A Concrete Application of Open Science for Natural Language Processing

[author names blinded]

Keywords. concreteness; planning prompts; advice; goal pursuit; open science

Disclosure. The authors declare no conflicts of interest. For each study, we report how we determined our sample size, all data exclusions, and all measures. All data and analysis code from each study are available as Online Supplemental Material, stored on OSF at https://osf.io/dyzn6/

Abstract. Concreteness is central to theories of learning in psychology and organizational behavior. However, the literature provides many competing measures of concreteness in natural language. Indeed, researcher degrees of freedom are often large in text analysis. Here, we use concreteness as an example case for how language measures can be systematically evaluated across many studies. We compare many existing measures across datasets from several domains, including written advice, and plan-making (total N = 9,780). We find that many previous measures have surprisingly little measurement validity in our domains of interest. We also show that domain-specific machine learning models consistently outperform domain-general measures. Text analysis is increasingly common, and our work demonstrates how reproducibility and open data can improve measurement validity for high-dimensional data. We conclude with robust guidelines for measuring concreteness, along with a corresponding R package, *doc2concrete,* as an open-source toolkit for future research.

1. Introduction

1.1. Concreteness in Organizations.

Concreteness is a deeply rooted construct in our understanding of how people think. Concreteness is theorized to be a quality of a mental representation - as being specific and observable, rather than a broader schema or category (Brown, 1958; Burgoon, Henderson & Markman, 2013). In particular, concreteness is thought to vary across distance - things that are close (temporally, spatially, socially) are represented more concretely, while things that are further away are represented more abstractly (Trope & Liberman, 2003; 2010). Many models of learning are be defined as a process of synthesizing concrete sensory representations into abstract concepts and representations (Kolb, 1976; Paivio, 1991; Bengio, 2009). In language, concreteness is often defined as the degree to which the concept denoted by an utterance refers to a perceptible entity (Paivio, 1991). This implies that the concreteness of these representations is thought to be detectable from the natural language people generate to describe those representations (Snefjella & Kuperman, 2015).

Researchers in organizational behavior has begun to incorporate concreteness as a framework to understand how people pursue many kinds of personal and organizational goals (Wiesenfeld, Reyt, Brockner & Trope, 2017). For example, the linguistic expression of concreteness has been studied in a diverse set of goal pursuit domains, including deception detection (Kleinberg et al., 2019; Calderon et al., 2019), clinical interventions (Querstret & Cropley, 2013), personality assessment (Mairesse et al., 2007), word of mouth (Schellekens, Verlegh & Smidts, 2010), leadership

communication (Carton & Lucas, 2018), entrepreneurial pitches (Joshi et al., 2020) and social media (Snefjella & Kuperman, 2015; Bhatia & Walasek, 2016).

In this paper, we focus on two organizational domains in which natural language can support goal pursuit - either for someone else ("giving advice") or the speaker herself ("making plans"). This builds off prior work that has theorized an important role for concreteness in both domains. Specifically, research has suggested that advice is often too abstract, and that advisors can be more successful when they provide concrete, specific details to recipents (Ilgen, Fisher & Taylor, 1979; Baron, 1988; Hinds, Patterson & Pfeffer, 2001; Goodman, Wood & Hendrickx, 2004; Kraft & Rogers, 2015; Reyt, Weisenfeld & Trope, 2016). Likewise, a similar literature has been building to suggest that plan-making is most successful when it is concrete and specific (Gollwitzer & Sheeran, 2006; Milkman et al., 2011; Rogers et al., 2015). This theoretical grounding in both domains suggests that one way to improve these kinds of organizational communication is to encourage advisors and plan-makers alike to be more concrete. Building a better measure of concreteness could aid in the development and evaluation of interventions based along these lines. More practically, these two conversational goals are pervasive, and consequential. These domains naturally produce lots of text data, across a diverse set of field contexts within each domain.

1.2. Concreteness in Natural Language.

This rich conceptual framework for linguistic concreteness has naturally spurred an interest in measurement tools. And the previous literature has generated a substantial set of candidate measures, that all lay claim to essentially the same task -

algorithmically generating a single concreteness "score" for a piece of text (Paivio, Yuille & Madigan, 1968; Pennebaker & King, 1999; Hart, 2001; Larrimore et al., 2011; Brysbaert, Warriner & Kuperman, 2014; Paetzold & Specia, 2016; Seih, Beier & Pennebaker, 2017; Pan et al., 2018; Johnson-Grey et al., 2019). From one perspective, a researcher might be grateful for this diversity of potential tools at their disposal.

However, we argue that the multiplicity of plausible measures creates more problems than it solves. First, it increases the number of researcher degrees of freedom, which is a threat to credible inference (Simmons, Nelson & Simonsohn, 2011; Gelman & Loken, 2014). In a canonical example, Simmons et al. (2011) demonstrate via simulation that when researchers can choose from among two correlated dependent measures, their false positive rate approximately doubles. Second, even if a researcher wanted to restrict their analytical flexibility by pre-registering only one of these measures, the literature does not provide reliable guidance for which of these models accurately capture linguistic concreteness, and under what circumstances.

These issues are well-exemplified in two recent studies that failed to find a long-hypothesized correlation between deception and concreteness (Kleinberg et al., 2019; Calderon et al., 2019). Previous previous papers have suggested a deep conceptual link between the concreteness of a description and its veracity (Johnson, 1988; Masip et al., 2005). Accordingly, both papers test several measures, across large samples from different contexts, and conclude that linguistic concreteness is not systematically correlated with deception. But they do not examine whether the linguistic concreteness measures they use are valid measures of concreteness.

1.3. Measurement in Natural Language.

These problems are not unique to concreteness. While measurement validity is a classic psychometric concern (Cronbach & Meehl, 1955; John & Benet-Martinez, 2000; Flake, Pek & Hehman, 2017; Fried & Flake, 2018), it is a particularly vexing when a latent construct is measured from open-ended data, like text. This is because text is extremely high-dimensional - even after data have been collected, they can be quantified in an essentially infinite number of ways. And, like the ancient Greek paradox of the heap of sand, the distinctions between measures can be made arbitrarily small: if a single word is removed from a dictionary, is the new dictionary the same measure as the original, a new measure of the same construct, or a new construct entirely?

Prior research has suggested family-wise correction techniques as a remedy for multiple hypothesis testing. For example, researchers could compare the correlation of a measure to its construct, relative to a set of other comparable measures and constructs (Campbell & Fiske, 1959). Researchers could also alter their threshold for statistical significance based on the number of other measures under consideration (Holm, 1979; Hochberg, 1988). Alternatively, researchers could report the results of analyses using every possible specification of a measure (Steegen et al., 2016).

Family-wise adjustments are impractical when the number of potential measures approaches infinity. Take, for example, the Linguistic Inquiry Word Count, the most common text analysis software in psychology (Tausczik & Pennebaker, 2010). This software produces ~90 separate language metrics for each document. Furthermore, users are encouraged to combine different scales for their application, and to reverse-score items where needed. Even limiting ourselves only to three-item measures, the

consideration set is over 5.6 million. A Bonferroni correction would imply a threshold for significance of less than 10^{-8} - the required sample sizes would make credible text analysis all but impossible.

1.3. Overview of Current Research.

In this paper, we describe a set of protocols for systematically constructing and evaluating measures in natural language. We use linguistic concreteness as an example, that highlights concerns common to all kinds of text analysis. This is important because the natural language processing toolkit is improving rapidly (Grimmer & Stewart, 2013; Hirschberg & Manning, 2015; Jurafsky & Martin, 2019), and these tools are becoming more popular in organizational research (Kabanoff, 1997; Pollach, 2012; Short, McKenny & Reid, 2018).

In Section 2, we review the existing literature, which offers many competing measures of this single construct. Next, we evaluate measurement validity by conducting empirical tests of these models in two domains of substantive interest. In Section 3, we compare these algorithms across datasets from a variety of experiments that involved writing tasks, like giving advice (9 studies, 4,608 participants). In Section 4, we then conduct similar analyses with manipulated and annotated concreteness labels from a field experiment testing planning prompts in online education (7 classes, 5,172 students). In Section 5, we use basic machine learning tools to directly estimate new domain-specific models of concreteness. Overall, our results suggest that many existing models of linguistic concreteness have little or no measurement validity in these domains, although machine learning can produce valid in-domain language measures.

In Section 6, we discuss how our systematic review shows that principles of open science - data and methods pooled from different researchers, and transparent, reproducible code - allow for a more cumulative contribution to the literature. In that spirit, we provide a new R package *doc2concrete* that contains reproducible, and contextually valid models of concreteness in natural language, as an open-source a toolkit for future research. Our investigation highlights the need for improved standards of measurement validity in organizational research, especially in the case of text analysis, and suggest meta-science as one productive way forward.

2. Linguistic Measures of Concreteness

2.1. Human vs. Algorithmic Text Analysis

Traditionally, constructs from language data are measured using human annotators. Consider, for example, a researcher who has a sample of natural language texts, and has a hypothesis about how the concreteness of these texts varies with respect to some other variable (e.g. by gender, or by role). They would train a group of annotators - perhaps research assistants, or crowdsourced online workers. Each annotator would read some texts, and independently assess their concreteness using one or a set of scales, or some other predefined rubric (Semin & Fiedler, 1988; Vallacher & Wegner, 1989). The inter-rater reliability of the annotations would be assessed based on the correlation of their ratings on the same texts (Shrout & Fleiss, 1979). The independent annotations would be averaged together to form a final score for each document, and then those scores are entered into a regression.

Although we do not focus on human annotations here, we acknowledge that they have clear benefits, compared to algorithmic measures. The primary advantage human annotators have is measurement validity - whether the generated score is correlated with the construct it claims to be measuring (Cronbach & Meehl, 1955; John & Benet-Martinez, 2000; Flake, Pek & Hehman, 2017; Fried & Flake, 2018). Humans excel at reading comprehension, spelling and grammar correction, and can adjust their interpretations to the domain. Although natural language processing has made substantial advances, many complexities of language are still glazed over. For example, none of the measures reviewed here take into account word order. These technological limits impose a hard ceiling on the validity of algorithmic models for complex constructs.

Algorithmic measures have their own advantages. However, these advantages require reproducibility - that is, an analysis must easily be reproduced on the same data by an outside researcher (Peng, 2011; Bollen et al., 2015; Bergh et al., 2017). Reproducibility is a foundational principle of open science, but we argue it is especially important for natural language measures, for three reasons. First, if an algorithm is reproducible, it is often perfectly reliable. An algorithm can give the same score to the same text every time, whereas the same text can receive different scores when given to different humans (or the same human at different times). Second, if an algorithm is reproducible, then it is transparent. An open-source algorithm can reveal exactly how a measure is calculated, whereas humans usually give holistic scores that leave room for misinterpretation. Finally, if an algorithm is reproducible, then it is scalable. If an algorithm is written well, the marginal cost of applying an algorithm to new texts is almost zero, whereas employing annotators at scale can be costly.

2.2 A Review of Algorithmic Models of Concreteness

Previous research has primarily measured the concreteness of a document in one of two ways. *Word-level* measures have assigned individual scores to a long list of common words, using human judges. *Categorical* measures create groupings of common word types, and the total counts for each group are scored. We review three word-level dictionaries, and six categorical measures - two of which are based on the Linguistic Category Model (Semin & Fiedler, 1988); three of which are derived from the Linguistic Inquiry Word Count (Tausczik & Pennebaker, 2010); and the last of which is included in the DICTION software package (Hart, 2001). For reference, we list all of these measures in Table 1, along with qualitative summaries of our results below.

2.2.1. Word-level Concreteness. Word-level measures use a long table of words that have been annotated for concreteness, one at a time, out of context (Paivio, Yuille & Madigan, 1968; Brysbaert, Warriner & Kuperman, 2014). This has some clear advantages - the results are easy to reproduce, and capture some general intuitions (e.g. "whenever" and "it" are more abstract than "friday" and "you"). However, homonyms (words with two meanings, such as "bank", or "like") are muddled. More importantly, this approach cannot capture any aspects of concreteness that are compositional, or contextual, or subjective.

One of these dictionaries (Brysbaert, Warriner & Kuperman, 2014) has already been successfully applied out-of-domain to recover concreteness-adjacent constructs

(temporal/social/geographic distance) in large-scale social media data (Snefjella & Kuperman, 2015; Bhatia & Walasek, 2016). Pragmatically, it covers most words in common usage (~40,000 entries, rated by 5+ Mechanical Turk workers). But we will also benchmark against the older and sparsely documented MRC Psycholinguistic database (annotated by trained researchers), which has ~9,000 entries (Coltheart, 1981). We also test a more recent dictionary, that was created with a word embedding technique to extrapolate the original MRC list to 85,000 words (Paetzold & Specia, 2016). An example for each of these dictionaries is demonstrated in Table 1.

In previous work, these dictionaries were defined for single words, and the measures were validated by correlating the scores of individual words to previous word-level scores. However, for most applied research, these individual word scores must be combined and weighted into a document-level summary score. Previous research has primarily generated this score using unweighted averages of all the words in a document (Snefjella & Kuperman, 2015; Bhatia & Walasek, 2016), which we adopt as a baseline. Although their preprocessing is not entirely clear, we chose to include stop words ("it", "you", "where", "how") and numbers ("one", "ten"), since they are likely relevant in our domains of interest.

2.2.2. Linguistic Category Model. The Linguistic Category Model (henceforth "LCM"; Semin & Fiedler, 1988) is the categorical measure most commonly associated with concreteness. The LCM identifies language categories based on parts of speech - nouns, adjectives, state verbs, interpretive action verbs, and descriptive action verbs. Each category frequency is multiplied by a score to determine the documents'

concreteness. On its face, there are obvious elements of concreteness that the LCM cannot capture - for example, the word "concrete" is both a noun and an adjective; while the word "abstract" can be a noun, an adjective or a verb. However, it was initially developed from controlled lab experiments that focused on texts from descriptions of people, which constrained the ways in which words could be used in-context.

Originally, the LCM was developed to be annotated by hand, which limited these analyses to smaller sample sizes (including measurement validation). However, algorithmic grammar parsing has been improving substantially, for a variety of NLP tasks (Manning et al., 2014; Honnibal & Johnson, 2015). Furthermore, the verb categories can be parsed using word lists from the Harvard General Inquirer (Dunphy, Stone & Smith, 1965). One recent paper proposed that a document's part-of-speech tags can be tallied according to the original LCM formula (Seih, Beier & Pennebaker, 2017). They validated this approximation by showing this measure is affected by a distance manipulation (third-person vs. first-person perspective) using a dataset of 130 reflective writing samples from college students.

Seih and colleagues (2017) recommend a pre-trained scoring rule, which we follow: Direct Action Verbs = 1; Interpretive Action Verbs = 2; State Verbs = 3; Adjectives = 4; Nouns = 5. While all LCM papers follow a somewhat similar rule, the scores themselves vary from paper to paper. Nouns are a recent addition (Semin et al., 2002); sometimes the verb subtypes are collapsed (Reyt, Wiesenfeld & Trope, 2016), or expanded (de Poot & Semin, 1995; Reyt & Weisenfeld, 2015); and adjectives have also been divided into subcategories (Louwerse et al., 2010). However, the five categories usually fall in the same order across implementations.

Another recent model, the "Syntax LCM", implements the spirit of the LCM using a different approach (Johnson-Grey et al., 2019). First, they annotated a small set of documents - sentence-length descriptions of daily student life - using the original LCM procedure (i.e. by hand). Then they trained a machine learning model to predict the annotations using a broader set of 24 syntactic features, again relying on algorithmic grammar parsing to process the documents. In the original, their measure was validated on a sample of 500 sentences from descriptions of daily college student life, that included a manipulation of the distance of the audience (close vs. far).

2.2.3. LIWC Categories. We test several categorical models developed from the Linguistic Inquiry Word Count ("LIWC"), proprietary software that uses word lists to define content-focused categories (e.g. food, family, work, anger; Tausczik & Pennebaker, 2010). The LIWC is the most commonly-used category-based text analysis tool in psychology, and follows a similar approach to many kinds of constructs. Previous work typically combines sets of these lists to approximate a construct in natural language. Although this approach is common, we focus on three examples that have already been applied to measure concrete language collected from field settings.

One measure, "verbal immediacy", combines five categories - first person singular; present focus; discrepancies; (reversed)long words; and (reversed) articles (Pennebaker & King, 1999). This was developed from prior conceptual work on clinical responses to traumatic responses or events (Wiener & Mehrabian, 1968), More recently it has been applied to descriptions of more mundane experiences/stimuli, including language collected from email and experience sampling methods (Nook, Schleider, &

Somerville, 2017; Nook et al., 2019). The original measure was constructed from an exploratory factor analysis of 838 stream of consciousness texts from college students, however it has been replicated in many other papers (Cohn, Mehl & Pennebaker, 2004; Mehl, Robbins & Holleran, 2012), including in some direct pre-registered replication studies (e.g. Nook, Schleider, & Somerville, 2017).

Another set of three features - articles; prepositions; quantifiers - was originally applied as an "abstractness index" in a dataset of peer-to-peer lending decisions (Larrimore et al., 2011). It has been applied in other domains (Markowitz & Hancock, 2016; Toma & Hancock, 2012; Parhankangas & Renko, 2017). To our knowledge, no published work has validated it against an annotated measure (or manipulation) of concreteness.

The final LIWC scale we consider was developed to estimate "concreteness" in CEO earnings calls (Pan et al., 2018), a set of six features - verbs; numbers; past focus; (reversed) adjectives; (reversed) quantifiers; and (reversed) future focus. This was created ad-hoc for the paper in question, although others have used their formulation directly (Jacobsen & Stea, 2019). In the original, the authors validated this measure on a non-randomly selected sample of 60 texts, and then applied it to a larger sample.

2.2.4. DICTION. DICTION is a proprietary content analysis tool that counts the rate at which words from a set of dictionaries are used in a document. It was originally developed in political science (Hart, 2001), although more recently, management scholars have argued for the value of DICTION (Short & Palmer, 2008). Like the LIWC, DICTION encourages users to mix and match from among their forty content categories

- for our purposes, though, we use a single dictionary, labelled "concreteness". We find some evidence that organizational scholars have used the concreteness dictionary - for example, among entrepreneurial pitches (Allison, McKenny & Short, 2013), or in public statements from professional organizations (Rogers, Dillard & Yuthas, 2005). However, the original documentation does not describe how the categories were validated.

3. Study 1: Concreteness in Advice

One of the most important mechanisms for social learning is giving advice. People routinely seek and benefit from other people's opinions when making their own choices (Goldsmith & Fitch, 1997; Bonaccio & Dalal, 2006; Berger, 2014). Likewise, people often seek advice on their performance, including feedback on past performance (Ashford & Cummings, 1983). However, the net effects of feedback are less clear (Kluger & DeNisi, 1996), and the effect of feedback depends on the content of that feedback. Advice is often theorized to be more effective when it includes specific, actionable suggestions that can be followed, rather than abstract evaluations (Ilgen, Fisher & Taylor, 1979; Baron, 1988; Hinds, Patterson & Pfeffer, 2001; Goodman, Wood & Hendrickx, 2004; Kraft & Rogers, 2015; Reyt, Weisenfeld & Trope, 2016). However, this literature has almost exclusively relied on manipulated specificity, or else human-annotated specificity, to determine the concreteness of a piece of advice.

To study concreteness in this domain, we collected a group of datasets from other researchers. Our primary objective in this search was to collect text where the goal was to give advice or feedback. Furthermore, we wanted to sample from advice in a variety of contexts, to see whether concreteness has structural or stylistic similarities

across many kinds of advice, or else if it is a simple property of the particular content of a domain. For breadth, we also include some datasets from more traditional language tasks in the lab, where the writer is simply prompted to describe a stimulus.

Every observation in each dataset consists of a single text document, and a valid measure of concreteness (e.g. human annotations) that we can use as a "concreteness index" to benchmark the language models. The sample of studies is not intended to be representative - instead they were gathered from published or working papers from a range of other authors via informal conversations (see Table 3). Due to the diverse progeny of these datasets, the protocols of each study differ slightly within the theoretical umbrella of concreteness. Arguable, this natural variation actually supports the aim of our investigation, because we are trying to evaluate the reproducibility and validity of concreteness models across contexts and research teams.

3.2. Study 1 Datasets

3.2.1. Workplace Feedback. Employees at a food processing company were included in an annual developmental review process (Blunden, Green & Gino, 2018). Each person was asked to write feedback for 5-10 of their peers, which would then be shared with that person. The feedback was annotated for specificity one at a time by two RAs on 1-7 scale (ICC = .82), and that average RA rating was used as the concreteness index.

3.1.2. Personal Feedback. Participants on mTurk were asked to think of a person in their life to whom they could give feedback on a recent task. Then, they were

asked to write what feedback they would provide (Blunden, Green & Gino, 2018). The written feedback was shown to 5-6 raters (also mTurkers) who evaluated the specificity of the feedback, and we take the average of these raters as the concreteness index (ICC = .86).

3.1.3. Teacher Feedback. Middle school students were enrolled in an education intervention designed to facilitate communication with the parents of their students (Kraft & Rogers, 2015). Up to four times over a single summer school term, teachers wrote single-sentence feedback to their students' parents, which was then embedded in a form letter and sent out in some conditions. Each student was assigned to receive either Improvement or Positive feedback all summer, and afterwards a research assistant blind to condition confirmed that the Improvement feedback was much more actionable (89% vs. 8%). We used the condition labels as the concreteness index. We also collapse all four pieces of feedback for each student-class pair (some students took multiple classes) and drop students who did not receive all four pieces of feedback, in line with the original intervention.

3.1.4. Task Tips. Participants were recruited to an on-campus behavioral lab to participate in a study on task performance (Levari, Wilson & Gilbert, 2020). They first played a skill game (e.g. boggle, darts) and then wrote advice about how to do well to the next participant. Each piece of advice was hand-coded by a pair of RAs (r = 0.69-0.73) for several features - here, the only relevant feature is how "actionable" the advice is, which we use as the concreteness index.

3.1.5. Letter Advice. Participants on mTurk were given a cover letter for a job application with errors in it, and were told to provide their input - either "advice" or "feedback" - to the writer (Yoon, Blunden, Kristal & Whillans, 2020). These written responses were then shown to six raters who used a three-item likert scale to evaluate several dimensions, including "actionability" and "specificity". The average ratings of these two scales were highly correlated (r=.92) so we standardized them into a single concreteness index.

3.1.6. Life Goals. Participants on mTurk were told to give general advice on how to live a happy life to someone either younger or older than they were (Zhang & North, 2020). Each document was then shown to 7-10 raters (also mTurkers) who annotated several dimensions. The most relevant for our purposes are "abstract" and "specific" - the averages of these two ratings were quite negatively correlated (r=-0.63) so we standardize and average them for the concreteness index.

3.1.7. Why vs How. Participants from mTurk were told to describe the beginning, middle and end of their work day (Yoon, Whillans & O'Brien, 2020). Participants wrote in three separate text boxes, that we combined into a single document for each person. Here, the concreteness index is randomly assigned: half of participants were told to explain "how" they did things that day, while the other half were told to explain "why" they did things that day. This task is commonly used as a mindset induction in construal level research, used over a variety of domains and measures (e.g. Freitas, Gollwitzer &

Trope, 2004; Fujita et al., 2006), though the language produced is not often analyzed as a manipulation check.

3.1.8. Self-Distancing. Participants from mTurk were told to describe their reactions to a series of emotionally negative cue words (Nook, Schleider & Somerville, 2017). The concreteness index was randomly assigned and blocked within-subjects, with two blocks of 20 words each. In one condition, participants were told to imagine the cue word at a distance - either in another place, at another time, or to another person - and in the other condition they imagined it close (along the assigned dimension). We combine all the descriptions within each of the two conditions (i.e. two documents per person).

3.1.9. Emotion Words. Participants from mTurk were presented with 20 emotion words, one at a time, and told to told to write a definition of the word (Nook et al., 2019). We combine all twenty texts to producing one document per person. Each person's set of descriptions was annotated by two research assistants. They answered three scale items asking about the abstractness/generality of the definition (correlation across raters = .89, and Cronbach's alpha across scales = .93). The concreteness index was created as a standardized average of all of these ratings.

3.2. Study 1 Results

Our primary research question was to know how well these models of linguistic concreteness correlate with the "concreteness index" within each dataset, and with one

another. To create a consistent comparison across methods, we always model each concreteness index as a linear outcome, transformed to have a mean of zero and variance of one. Likewise, all the predictions from the linguistic measures received a similar transformation, calculated separately for each dataset. For clarity throughout, all measures are oriented so that higher numbers indicate more concreteness, which means some models (e.g. the LCM) are reversed from their original orientation.

3.2.1. Correlation between models. One possibility is that these models all correlated with one another, in which case they would not need to be differentiated. In Figure 1, we show the correlation between the different off-the-shelf models within each dataset. The dictionaries hold together quite well, with average pairwise correlations ranging from .663 to .738. The two LCM measures are always positively correlated, but not strongly so, with an average correlation of .352. The figure also shows a surprising number of negative correlations. These results present an initial quandary - sometimes, linguistic measures of the same construct do not correlate with each other.

3.2.2. Word Count Baseline. The most consistent measure of concreteness in Study 1 was the total number of words in the document. The raw correlations were significantly positive in six of the nine datasets, and all of the advice datasets (pooled r=.536, 95% CI=[.511, .560]) ranging from Life Goals (r=.233, 95% CI=[.123, .337]) to Workplace Feedback (r=.763, 95% CI=[.740, .785]). However word count was not a significant predictor of concreteness in the description tasks (r=.009, 95% CI= [-.045,

.063]). While advice may be abstract due to a lack of specific detail, this result has limited prescriptive value - that is, people may not know what to say.

We wanted to control for word count, to more clearly identify concreteness in the content of what someone is saying. As word count is zero-bounded and right-skewed, a logarithmic transformation of word count produces a more normal distribution. While both measures were significantly correlated with concreteness, the overall model fit is much higher with the log-transformed word count (R^2 = .060) than the linear term (R^2 = .002). This result holds when we include dataset fixed effects, as well (linear: R^2 = .007; log-transformed R^2 = .260). We confirm all our results below are substantively similar without this control, as well.

3.2.3. Correlation with Concreteness. We first estimated the concreteness of a texts' content, controlling for log-transformed word count, using a hierarchical linear model (Bates et al., 2007). This model predicted concreteness, using a random intercept at the dataset level, and a random slope for an effect of log-transformed word count that varies across datasets. The residual of this model was then treated as our index of concreteness content in each document (all of our results are substantively similar if we use the unadjusted concreteness scores as our measure).

In Figure 2, we plot the correlation between concrete content and each of the language measures, separately for each study. The results suggest that some of these measures do capture meaningful concreteness in the content of what someone writes. However, the most prominent finding is the sheer variability across measures and

datasets. Some measures correlate with concreteness positively, others negatively, and others not at all, and these relationships change from context to context.

There are some consistent results. All of the word-level measures were able to detect concreteness above chance in most of the advice datasets. However, performance on the pooled advice data seemed to be higher for the Brysbaert dictionary ($r = .155$, 95% CI = [.122, .188]; $t(3287) = 9.0$, $p < .001$) than either of the MRC-based dictionaries (Bootstrap: $r = .117$, 95% CI = [.083, .150]; $t(3287) = 6.7$, $p < .001$; Original: $r = .076$, 95% CI = [.042, .110]; $t(3287) = 4.4$, $p < .001$). But this relative order was reversed in the pooled description datasets, with the original MRC coming out on top ($r = .286$, 95% CI = [.236, .335]; $t(1317) = 11$, $p < .001$; Bootstrap: $r = .163$, 95% CI = [.110, .215]; $t(1317) = 6.0$, $p < .001$; Brysbaert: $r = .131$, 95% CI = [.078, .184]; $t(1317) = 4.8$, $p < .001$).

The results were less positive for the categorical models. Some categorical measures performed well on the pooled description datasets (Immediacy: $r = .363$, 95% CI = [.315, .409]; $t(1317) = 12$, $p < .001$; Part of Speech LCM: $r = .264$, 95% CI = [.214, .314]; $t(1317) = 10$, $p < .001$; Syntax LCM: $r = .181$, 95% CI = [.128, .233]; $t(1317) = 6.7$, $p < .001$; DICTION: $r = .066$, 95% CI = [.012, .119]; $t(1317) = 2.4$, $p = .017$). However, others were negatively correlated in the description data (Larrimore-LIWC: $r = -.114$, 95% CI = [-.167, -.060]; $t(1317) = 4.2$, $p < .001$; Pan-LIWC: $r = -.049$, 95% CI = [-.103, .005]; $t(1317) = 1.8$, $p = .075$). In the advice datasets. most categorical models either found no correlation or a negative one, with concreteness, on average (Immediacy: $r = -.115$, 95% CI = [-.149, -.082]; $t(3287) = 6.7$, $p < .001$; Part of Speech LCM: $r = -.034$, 95% CI = [-.068, .000]; $t(3287) = 1.9$, $p = .052$; Syntax LCM: $r = .002$, 95% CI = [-.033,

.036]; t(3287) = 0.1, p = .934; Pan-LIWC: r = -.065, 95% CI = [-.099, -.031]; t(3287) = 3.8, p < .001). Two other models found positive correlations that were smaller than all the word-level models (Larrimore-LIWC: r = .047, 95% CI = [.013, .082]; t(3287) = 2.7, p = .006; DICTION: r = .056, 95% CI = [.022, .090]; t(3287) = 3.2, p = .001).

One concern we had during our review of the Linguistic Category Model was that scoring rules varied. Rather than iterating through every possible scoring rule, we estimated a score for every category separately, within each dataset. We summarize these results graphically in Appendix A. The description tasks mostly validate the linguistic category model, as the correlations roughly line up in ascending order. However, the advice datasets stand in stark contrast. It is hard to identify any previous LCM scoring rule (for example, removing the noun category) that is consistent with these results. We conduct a similar exercise in Appendix B with the features from the LIWC categorical models. Consistent with the top-line results, the category-by-category analyses do not reveal any consistent pattern, with the exception of the immediacy categories in the description datasets.

3.3. Study 1 Discussion

Concreteness is broadly ingrained across many psychological models of social learning, and several approaches to measuring concreteness in language have been proposed. We compared all of these measures in datasets from a wide range of contexts. And we did find that some results generalized well. First, word count typically predicted concreteness in open-ended language, sometimes quite strongly. Additionally, the content also reliably contained indicators of the speaker's concreteness. The word-

level methods were somewhat reliable across domains, though the effect sizes were typically small (and the effect sizes were smaller still for DICTION).

We also found results that were consistent across datasets, but not across domains. While some of the categorical measures (Immediacy, Part of Speech LCM) were able to detect concreteness across description datasets, they mostly failed to detect concrete advice. Domain specificity is common, even in the most basic linguistic phenomena (e.g. Mehl, Robbins & Holleran, 2012; Hamilton et al., 2016). For example, while positive and negative words signal felt emotions in descriptions (like product reviews; Pang & Lee, 2008), they fail to reveal felt emotions in everyday speech (Beasley & Mason, 2015; Sun et al., 2019; Kross et al., 2019; Jaidka et al., 2020).

Description tasks typically constrain the topic (e.g. "what did you think about this product?"), which reduces the distribution of words and goals. This increases internal validity and experimental control, which makes it a natural fit for lab experiments. However this can come at the cost of external validity in open-ended natural language. The Linguistic Category Model was initially proposed for measuring trait descriptions (Semin & Fiedler, 1988). Perhaps that is why, in these data, the LCM performs best on the Teacher Feedback dataset, in which teachers wrote feedback to parents about their children, rather than to the students themselves.

4. Study 2: Plan-Making in Online Education

In Study 2 we extend our results to a new goal pursuit domain - plan-making. A long literature has found positive effects of generating plans as a means to follow

through on one's current intentions for future behavior. And the mechanisms may not be so different from advice - plan-making can be thought of advice for one's "future self". Early research on planning has primarily drawn from lab experiments (Gollwitzer & Sheeran, 2006), although the effects have been extended more recently into field experiments (Rogers et al., 2015).

The bulk of the evidence on planning interventions has primarily focused on the pursuit of one-time actions like voting, or a doctor's visit (Nickerson & Rogers, 2010; Milkman et al., 2011). Often, in these cases, it is recommended that a plan is more likely to be executed when it includes concrete, specific details to follow. However, many intentions require complex and long-term goals, that cannot be summarized in a single plan, and where concreteness may not even be ideal (Townsend & Liu, 2012; Beshears et al., 2020). For these complex goals, the concreteness of a plan might vary along many dimensions. That is, a plan's concreteness could potentially be driven by the specificity of one (or both) of the actions in the plan, and the temporal scope on which those actions occur. Here, we make this subtle distinction an empirical question, by collecting two different measures of concreteness in the same plan-making dataset.

Here, we use data collected during an intervention conducted in every online course released by HarvardX, MITx and StanfordX from September 2016 - December 2017 (from Kizilcec et al., 2020). Each of those courses had a pre-course survey that included a block for randomly-assigned interventions, of which one was an open-ended planning prompt (see Appendix C for exact stimuli). We then compare the linguistic measures of concreteness of the written plans against two concreteness indices - random assignment to short- or long-term plans, and human ratings of specificity.

4.1. Study 2 Methods

We delegate most of our analysis choices to the pre-registered analysis plan generated from the original research (Kizilcec et al., 2020), including exclusion criteria, and model specifications. However, the original research (which focused on intent to treat analyses) did not include any accommodation for cleaning text. For this research, we created an algorithmic filter to remove people whose true plan-making would not be captured by our NLP (e.g. if they wrote in another language, or if they provided an insincere response like pasting copied text or typing nonsense). We also asked our annotators to filter cases where the response was clearly insincere. Observations were filtered at similar rates across conditions (X^2 (1) = 0.2, p = .674), and all non-filtered text was analyzed raw, with no corrections (e.g. for spelling).

4.1.1. Annotated Concreteness. We trained two research assistants to annotate the specificity of the plans - i.e. if a plan could be executed without more detail, and its execution could be objectively verified (see Appendix D for exact instructions). After practicing together on three small pilot classes, they then produced independent ratings for a selection of seven larger classes (N = 5,172 students after exclusions) that covered a range of common subjects (e.g. computer science; law; biology; literature). Each annotator provided two ratings: whether a plan could be concrete for the writer herself, and whether it could be concrete for another student. We average all four ratings to produce an annotated concreteness index.

4.1.2. Manipulated Concreteness. The experiment also included two types of planning prompts, randomly-assigned, which provides a second potential concreteness index. Students were asked to make a plan for either the first week of the course ("short plans"), or for the entire course ("long plans"). Similar kinds of temporal distance manipulations have often been used in construal level research (Trope & Liberman, 2003). So we also tested whether the concreteness models were able to detect the difference between short plans or long plans. For ease of comparison, we report results from the seven classes where the data was also annotated - however, we confirm the results are robust across the larger sample of 151 classes from the original study.

4.2. Study 2 Results

The preregistration in the original research included course fixed effects and clustering standard errors at the course level, as well as a set of control covariates from the - expected hours/week, intention to pass, previous MOOCs completed, date of enrollment - that were collected before the planning prompts. This is the model we report in the text below. For robustness we also systematically varied some details of the model specifications, as shown in the Figure 3, and find similar results.

4.2.1. Word counts. Following Study 1, we also included log-transformed word counts in some of the regression specifications, for robustness. The average specificity ratings were positively correlated with the log-transformed word count ($\beta = .575$, $SE = .039$, $z(5158) = 15$, $p < .001$). However, long-term plans had higher word counts, on average, than short-term plans ($\beta = -.116$, $SE = 0.34$, $z(5158) = 3.4$, $p < .001$).

4.2.2. Plan Distance. In Figure 3 we show estimates for the effect of the manipulation of plan distance on linguistic concreteness. Several concreteness measures detected more concreteness in the short plans condition. In particular, the dictionary methods performed well - while the Brysbaert dictionary demonstrated the largest raw coefficient ($\beta = .098$, $SE = .040$, $z(5158) = 2.4$, $p = .015$) the other dictionaries were also somewhat valid measures (Bootstrapped MRC: $\beta = .075$, $SE = .034$, $z(5158) = 2.2$, $p = .029$; Original MRC: $\beta = .053$, $SE = .028$, $z(5158) = 1.9$, $p = .059$). However, none of the categorical models showed a positive significant relationship with plan distance (all $p > 0.12$).

4.2.3. Specificity Ratings. The two human raters were closely correlated with one another ($r = .642$, 95% CI = [.626, .658]). Interestingly, we also observed no effect of the assigned plan distance on annotated specificity ($\beta = .009$, $SE = .035$, $z(5158) = 0.3$, $p = .796$). Figure 3 also shows the relationship between the linguistic measures of concreteness and the specificity ratings. The dictionaries were once again consistent, and the Brysbaert was directionally the closest to annotated specificity ratings ($\beta = .151$, $SE = .009$, $z(5158) = 16$, $p < .001$), while the others were close behind (Bootstrapped MRC: $\beta = .136$, $SE = .011$, $z(5158) = 13$, $p < .001$; Original MRC: $\beta = .039$, $SE = .011$, $z(5158) = 3.7$, $p < .001$). Three of the categorical measures also found a weaker correlation with specificity (Pan-LIWC: $\beta = .091$, $SE = .016$, $z(5158) = 5.8$, $p < .001$; Larrimore-LIWC: $\beta = .053$, $SE = .012$, $z(5158) = 4.5$, $p < .001$; Syntax LCM: $\beta = .055$, $SE = .027$, $z(5158) = 2.0$, $p = .044$). However, the other measures were either

indistinguishable zero (DICTION), or significant in the opposite direction (Immediacy and Part-of-Speech LCM).

4.3. Study 2 Discussion

Like Study 1, the word-level concreteness measures were more reliable indicators of both kinds of linguistic concreteness, while the categorical measures found much smaller effects, or no effect at all. These results also showed that concreteness itself is multifaceted, even within the same dataset. A manipulation of concreteness (via temporal distancing) had no effect on our annotated measure of concreteness (via specificity). One possible interpretation of Study 1 is that while the linguistic expression of concreteness is domain-specific, it may still reflect a domain-general cognitive architecture (Paivio, 1991; Trope & Liberman, 2003, 2010). The results of Study 2 suggest something deeper - that the underlying construct of concreteness may be multifaceted, or domain-specific (Fiedler, et al., 2003; Troche, Crutch & Reilly, 2017; Borghi et al., 2017; Pollock, 2018). Regardless, both potential mechanisms suggests that the generalizability of a language measure developed on a single domain should not be taken for granted.

5. Estimating Domain Specific Models of Concreteness.

The results above suggest that there may be substantial domain-level differences in concreteness. Our open science approach allows for an empirical test of the domain specificity of concreteness. That is, we systematically estimated four new scoring rules, for one domain at a time - advice concreteness, description concreteness, plan

distance, and plan specificity (the two plan domains use the same text, albeit with different outcomes). We then compared the accuracy of those models in each of the other domains. This procedure gives us an estimate of the similarity in how concreteness is expressed in each domain.

5.1. Methods

For each domain, we trained a supervised machine learning model to predict concreteness from the text. We created a large list of features, using a relatively simple bag-of-ngrams approach. That is, we tallied all 1-, 2-, and 3-word sequences, including stop words, that occur in more than 0.5% of documents, using the quanteda R package (Benoit et al., 2018). We also included summary scores from the Brysbaert and Paetzold dictionaries. We then used a simple estimation algorithm, the LASSO, to build a model that predicted a document's concreteness score based on using the feature counts (Friedman, Hastie & Tibshirani, 2010).

To evaluate the out-of-sample accuracy of the models, we used a nested cross-validation loop (Varma & Simon, 2006). When we tested a model across domains, we used all of the training domain data in each model, and it was tested all at once on the test domain data. When we we tested a model within a single domain, we generated predictions one context (dataset in Study 1, or courses in Study 2) at a time, using all of the other contexts from that domain as a training set.

5.2. Results

5.2.1. In-Domain Accuracy. In Table 4, we compare the performance of the different models, using correlation with the concreteness index in each domain. Each of these models was somewhat successful in its own domain, affirming concreteness as a stable linguistic construct. These in-domain tests also reliably outperform the cross-domain tests. This suggests the potential for any domain-general concreteness measure is limited. For example, while plan distance seems to be stable across classes, it does not have any validity for measuring concreteness in the other domains.

It is also interesting to compare these in-domain results with the domain-general measures reviewed above. In three of the four domains, the machine learning model was clearly more accurate. This was true for the advice data (ML: r = .228, 95% CI = [.195, .260]; Brysbaert: r = .155, 95% CI = [.122, .188]); for assigned plan distance (ML: r = .228, 95% CI = [.195, .260]; Brysbaert: r = .047, 95% CI = [.020, .075]) and for annotated plan specificity (ML: r = .733, 95% CI = [.720, .745]; Brysbaert: r = .438, 95% CI = [.416, .460]). This was not true for the descriptions (ML: r = .092, 95% CI = [.038, .145]; Immediacy: r = .363, 95% CI = [.315, .409]), although this may be because that model had the least available training data (only 1319 observations from 3 contexts). Additionally, all of these these results underestimate the power of in-context machine learning because they were trained on data that was in-domain but out-of-context.

5.2.3. Training Set Size Simulations. Although hand-labeled data are usually more accurate than domain-general measures, researchers may fairly be concerned that annotation does not come cheaply. However our results suggest even models trained on hand-labeled data can reliably outperform domain-general measures.

This suggests that one cost-effective approach is to collect annotations for a portion of a dataset, and then train a model to apply predicted annotations in the unlabeled data.

In Appendix E, we benchmark the effect of training set size on accuracy using the advice data from Study 1 and the annotator data from Study 2. Specifically, we conducted a simulation in which we train many supervised models, just like the ones above, however we systematically train each model using only a randomly sampled subset of our full data. Broadly, our results suggest that the gains from additional training data for our simple models tend to taper off after approximately 500 labels. This is a rough guide for researchers considering an annotation exercise themselves, although surely the results will vary based on the domain, population, task, and model. This exercise also provide some perspective for the estimated validity of the models we reviewed in Section 2. Some models (e.g. Pan-LIWC, or Part-of-Speech LCM) were initially validated on samples that were likely too small to evaluate validity.

5.3. Discussion. These results suggest an upper limit on the domain-general validity of any language model. We suspect that constructs may be especially domain-specific in goal pursuit domains, where the meaning depends on external factors, and the recipient herself. For example, while some datasets in Study 1 focused on generic advice (e.g. Task Tips) or a single recipient (e.g. Letter Advice), many advice contexts involve personalized advice, which naturally changes the advisors' approach (Eggleston et al., 2015; Yeomans, 2019). Likewise, the machine learning model could only estimate a generic model of concrete plans, because it did not take into account any individuating

course characteristics (length, subject, structure, student pool). And plans for other kinds of long-term goal pursuits may require yet another new measure.

Domain specificity is not controversial in principle - situational and contextual moderators are a foundational concern in social psychology (Ross & Nisbett, 2011). However, this distinction is often glossed over when researchers borrow a language measure from another domain. An implicit assumption of off-the-shelf language measures - including all of those reviewed in Section 2 - is that they are domain-general. They apply the same scoring rules to all text, regardless of the speakers or their goals. This means they cannot capture domain-specific features by definition.

6. General Discussion

Our work provides a unique and systematic review of concreteness in natural language. Our most consistent result was that a machine learning model trained on within-domain data, even with unsophisticated language processing to extract features, consistently produced more reliable estimates of concreteness than any domain-general model available. Our work suggests above all that concreteness is domain-specific, and multifaceted. This underscores the value of supervised machine learning as an empirical benchmark for theory-driven measures in observational data.

6.1 Concreteness in Natural Language.

Our cross-domain approach provides useful context for some widely-used off-the-shelf measures. Our results provide tentative support for the word-level methods as a weak-but-robust measure of concreteness across domains (Brysbaert, Warriner &

Kuperman, 2014; Paetzold & Specia, 2016). However, our tests of the other off-the-shelf measures were less promising. There were some domain-specific successes - for example, immediacy and the LCM measures performed well in the description tasks. Apart from those isolated cases, however, we failed to find any robust relationship with concreteness among the other LIWC constructs, or the DICTION scale.

Based on our results, we suggest three potential approaches to concreteness detection in new data. Ideally, researchers should annotate new data in their context of interest. However, this may be impractical, so we also propose alternatives that can be done without any new annotations. For researchers interested in a domain for which there is a good training data, should used an existing domain-specific measure. Absent a good domain-specific measure, researchers should use a word-level measure.

We implement these alternative approaches in an open source R package, *doc2concrete.* This package includes two pre-trained measures, which are intended to apply only to concreteness in the domains of advice or plan-making, respectively. Specifically, the package includes the best-performing supervised models - the LASSO model with bag-of-ngrams and dictionary features - to calculate concreteness in a new set of texts. For domains where good training data is not yet available, our results suggest that the largest word-level measures provide a good starting point (Brysbeart, Warner & Kuperman, 2014; Paetzold & Specia, 2016). The package includes implementations for both of these measures, with the Brysbaert as a default.

6.2. Natural Language in Open Science

Our work follows the spirit of recent systematic reviews showing that linguistic measures of psychological constructs provide varying results in observational data (Carey et al., 2015; Sun et al., 2019; Kross et al., 2019; Benoit, Munger & Spirling, 2019; Tackman et al., 2019; Jaidka et al., 2020). The multitude of plausible language measures for any single construct presents a challenging question for applied researchers. Upon what criteria should a researcher select language measure to test their research question? Here, we discuss two criteria in detail - measurement validity and reproducibility. Although the validity of algorithmic measures may only approximate human annotations, this may still be worthwhile for algorithms that are reproducible.

6.2.1. Measurement Validity. Our results suggest that measurement validity should not be taken for granted in language measures. Our review finds that many existing measures do not have validity in our results - and the papers in which they were initially proposed were probably underpowered to demonstrate validity. This is a general problem in all kinds of applied research (Cronbach & Meehl, 1955; John & Benet-Martinez, 2000; Flake, Pek & Hehman, 2017; Fried & Flake, 2018), and is rightly a focus of the open science movement. This is especially important when there are many researcher degrees of freedom even after the data are collected - if hypothesis testing is not constrained by external criteria, then it is likely that a disproportionate number of results will be false positives.

In particular, our field can benefit from increased use of machine learning techniques - regularization, cross-validation, and transfer learning, in particular (Yarkoni & Westfall, 2017; Mullainathan & Spiess, 2017). Practically, this means researchers can

focus on defining the empirical criteria by which a measure should be judged a success, and allow algorithms to fine-tune the scoring rule. When paired with proper validation techniques, this means the high dimensionality of the data is actually a benefit. That is, the algorithm can consider many different scoring rules during validation, and provide empirical estimates for the out-of-sample validity of the best available scoring rule.

Our results also suggest an additional concern with measurement validity in language: generalizability. We found that even the best domain-general measures could not approximate the accuracy of a simple in-domain measure. Text data is constantly generated during interactions in all kinds of domains, and while in principle any reproducible algorithm could produce a score for any piece of text, in practice that score may not mean the same thing in one domain as it does in another. Although these boundary conditions are not controversial in principle, initial authors may not be eager to state them explicitly (Simons, Shoda & Lindsay, 2017). Furthermore, authors may be reluctant to report negative results (Rosenthal, 1979), and we suspect this dynamic is exacerbated in natural language, for two reasons. First, because a negative result may be hard to interpret without also collecting valid human annotations - is it a failure of the theory, or the measure? Second, because when the researcher degrees of freedom are high, authors are likely to find other positive results that may be more captivating. While the traditional selective reporting problem suggests a resource-intensive process of discarding entire datasets, text analysis allows researchers to still make use of the dataset by discarding the language measure instead.

One solution we encourage is for the research community to embrace more systematic reviews like this one, that combine datasets from many domains. That way,

positive and negative results can be report in contrast with one another, so that results can be more cumulative and boundary conditions can be clearer. But that is only possible if authors embrace the emerging norms of open science - such as sharing their data with one another, and producing transparent, reproducible analysis code.

6.2.2. Reproducibility. Reproducibility is often defined as the ease with which an analysis can easily be recreated on the same data by an outside researcher (Peng, 2011; Bollen et al., 2015). Reproducibility is especially important for language measures, because that ensures they can then be reliably scaled up across many datasets, including those that are too large or too confidential to be assigned to human annotators. Furthermore, text analysis often involves a broad set of preprocessing decisions (Denny & Spirling. 2018). For example: how to correct spelling and grammar; whether stop words should be included, or for that matter numbers, or proper names; and how that affects phrase construction. These small decisions can create room for error, or flexibility in implementation. Ideally, a language measure will be transparent about all of these decisions, and provide an exact implementation, as we have done with the *doc2concrete* package.

The algorithmic measures we review here provided a range of reproducibility. The Syntax LCM was the most reproducible - all of the analysis code is open source on OSF, and available in R (although not as an official CRAN package). The word-level measures were also quite reproducible. Tables of their word scores are all posted publicly, and the broad calculations for generating document-level scores are reported in main text of their original papers. However, analysis code was not posted, which

elides some of the smaller preprocessing decisions. The part-of-speech LCM also has gaps in its reproducibility - while each paper reports their category-level scoring rule, it is not always clear how words were assigned to categories, among other decisions.

The LIWC and DICTION measures are perhaps the least reproducible of the set. This is primarily because their software is closed-sourced and paywalled. Researchers who pay their license fee are able to exactly match the analyses of the original. However, those analyses are kept opaque - both the preprocessing pipeline, and the words included in each category. This prevents integration with other text analysis tools in open source languages like R or Python. This isolates users from the larger community of text analysis research, and impedes use of these tools in platforms, apps, or interventions. Finally, the high price may lead researchers to erroneously assume that these proprietary measures are higher-quality than open source tools (Rao & Monroe, 1989). Although their proprietary software ensures algorithmic reliability, this limits their scalability in practice.

6.2.3. Other Considerations. While the focus here is on validity and reproducibility, we acknowledge there are many other qualities of a language measure that applied researchers should consider. For example, interpretability - the ability to generate a meaningful explanation of why a score is given (Doshi-Velez & Kim, 2017; Rudin, 2019). Algorithms can be interpretable by revealing their exact scoring rubric, although many of the more complex models in NLP rely on opaque black box algorithms. Likewise, language models can encode discriminatory biases from their training data and unwittingly encourage unfair treatment of marginalized and

underrepresented groups (Caliskan, Bryson & Narayanan, 2017; Kleinberg et al., 2018). It is also worth noting that human judgment can itself be uninterpretable, and unfair.

6.3. Conclusions.

Overall, the use of text as data has become increasingly common in the social sciences (Grimmer & Stewart, 2013; Hirschberg & Manning, 2015; Jurafsky & Martin, 2019). The rapid rise of recorded language data, and the corresponding progress of text analysis tools, have both made it easier to study more (and larger) kinds of social interactions efficiently. Furthermore, humans are constantly using natural language to interact one another, which means that research will usually be more ecologically valid when it observes linguistic behavior directly, rather than by proxy (e.g. self-report, observer impressions, lay theoretical vignettes). The scalability and ecological validity of language datas suggest that it will take an even more prominent place in the future of organizational research (Kabanoff, 1997; Pollach, 2012; Short, McKenny & Reid, 2018).

However, this tremendous research opportunity also comes with unique challenges. Language technologies have dramatically increased what we *can* measure, but these must be adopted in parallel with the tools that help us know what we *should* measure. Conversation is far too complex us to expect independent researchers to make all of these modeling choices correctly. Our field will flourish if researchers prioritize reproducible measures, and embrace domain specificity as the rule, rather than the exception, when measuring constructs in high-dimensional data like language. And the conventions of open science make it much easier to combine strengths of many tools, datasets, and frameworks, within a community of inquiry, and have that conversation together.

References

Allison, T. H., McKenny, A. F., & Short, J. C. (2013). The effect of entrepreneurial rhetoric on microlending investment: An examination of the warm-glow effect. *Journal of Business Venturing*, *28*(6), 690-707.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational behavior and human performance, 32(3),* 370-398.

Baron, R. A. (1988). Negative effects of destructive criticism: Impact on conflict, self-efficacy, and task performance. *Journal of Applied Psychology*, *73*(2), 199.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version*, *2*(1), 74.

Beasley, A., & Mason, W. (2015, June). Emotional states vs. emotional words in social media. In *Proceedings of the ACM Web Science Conference* (pp. 1-10).

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends in Machine Learning, 2(1),* 1-127.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müler, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of consumer psychology*, *24*(4), 586-607.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization*, *15*(3), 423-436.

Beshears, J., H.N. Lee, K.L. Milkman, R. Mislavsky, & J. Wisdom (2020). Creating Exercise Habits Using Incentives: The Tradeoff between Flexibility and Routinization. *Management Science, in press.*

Bhatia, S., & Walasek, L. (2016). Event construal and temporal distance in natural language*. Cognition, 152,* 1-8.

Blunden, H., Green, P., & Gino, F. (2018). The Impersonal Touch: Improving Feedback-Giving with Interpersonal Distance. *Academy of Management Proceedings, 2018(1).*

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*, *3*(4).

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, *101*(2), 127-151.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263

Brown, R. (1958). How shall a thing be called?. *Psychological review*, *65*(1), 14.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46(3),* 904-911.

Burgoon, E. M., Henderson, M. D., & Markman, A. B. (2013). There are many ways to see the forest for the trees: A tour guide for abstraction. *Perspectives on Psychological Science*, *8*(5), 501-520.

Calderon, S., Mac Giolla, E., Luke, T. J., Warmelink, L., Ask, K., Granhag, P. A., & Vrij, A. (2019). Linguistic Concreteness of True and False Intentions: A Mega-analysis. *OSF Preprint* https://doi.org/10.31234/osf.io/h7g8b

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Carey, A. L., Brucks, M. S., Küfner, A. C., Holtzman, N. S., Große, F. D., Back, M. D., ... & Mehl, M. R. (2015). Narcissism and the use of personal pronouns revisited. *Journal of personality and social psychology, 109(3),* 1-15.

Carton, A. M., & Lucas, B. J. (2018). How can leaders overcome the blurry vision bias? Identifying an antidote to the paradox of vision communication. *Academy of Management Journal*, *61*(6), 2106-2129.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, *15*(10), 687-693.

Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497-505.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

de Poot, C. J., & Semin, G. R. (1995). Pick your verbs with care when you formulate a question!. *Journal of Language and Social Psychology, 14(4),* 351-368.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168-189.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dunphy, D. C., Stone, P. J., & Smith, M. S. (1965). The general inquirer: Further developments in a computer system for content analysis of verbal data in the social sciences. *Behavioral Science*, *10*(4), 468.

Eggleston, C. M., Wilson, T. D., Lee, M., & Gilbert, D. T. (2015). Predicting what we will like: Asking a stranger can be as good as asking a friend. *Organizational Behavior and Human Decision Processes*, *128*, 1-10.

Fiedler, K., Bluemke, M., Friese, M., & Hofmann, W. (2003). On the different uses of linguistic abstractness: From LIB to LEB and beyond. *European Journal of Social Psychology*, *33*(4), 441-453.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*(3).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Freitas, A. L., Gollwitzer, P., & Trope, Y. (2004). The influence of abstract and concrete mindsets on anticipating and guiding others' self-regulatory efforts. *Journal of Experimental Social Psychology, 40(6),* 739-752.

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology, 90(3),* 351.

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

Goldsmith, D. J., & Fitch, K. (1997). The normative context of advice as social support. *Human communication research*, *23*(4), 454-476.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in experimental social psychology, 38, 69-119.

Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology, 89(2),* 248.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 595). NIH Public Access.

Hart, R. P. (2001). Redeveloping DICTION: theoretical considerations. *Progress in communication sciences*, 43-60.

Hinds, P. J., Patterson, M., & Pfeffer, J. (2001). Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. *Journal of applied psychology*, *86(6)*, 1232.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349(6245),* 261-266.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Honnibal, M., & Johnson, M. (2015, September). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373-1378).

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of applied psychology*, *64*(4), 349.

Jacobsen, D. H., & Stea, D. (2019, July). The Use of Metaphorical Communication and Language Concreteness in An Equity Crowdfunding Setting. In *Academy of Management Proceedings* (Vol. 2019, No. 1, p. 15404). Briarcliff Manor, NY 10510: Academy of Management.

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison

of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165-10171.

John, O.P. & Benet-Martinez, V. (2000). “Measurement: Reliability, construct validation, and scale construction.” in *Handbook of Research Methods in Social and Personality Psychology,* edited by H.T. Reis and C.M. Judd. New York: Cambridge University Press, 339-369.

Johnson, M. K. (1988). Reality monitoring: An experimental phenomenological approach. *Journal of Experimental Psychology: General*, *117*(4), 390.

Johnson-Grey, K. M., Boghrati, R., Wakslak, C. J., & Dehghani, M. (2019). Measuring Abstract Mind-Sets Through Syntax: Automating the Linguistic Category Model. *Social Psychological and Personality Science*, 1948550619848004.

Joshi, P. D., Wakslak, C. J., Appel, G., & Huang, L. (2020). Gender differences in communicative abstraction. *Journal of personality and social psychology, in press.*

Jurafsky, D. & Martin, J. H. (2019). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River: Pearson/Prentice Hall.

Kabanoff, B. (1997). Introduction: computers can read as well as count: computer-aided text analysis in organizational research. *Journal of Organizational behavior*, 507-511.

Kizilcec, R., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Williams, J., Turkay, S. & Tingley, D. (2020). Scaling Up Behavioral Science Interventions in Online Education. *Proceedings of the National Academy of Sciences, in press.*

Kleinberg, B., van der Vegt, I., Arntz, A., & Verschuere, B. (2019, March 1). Detecting deceptive communication through linguistic concreteness. *Working paper.* https://doi.org/10.31234/osf.io/p3qjh

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, *10*.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

Kolb, D. A. (1976). Management and the learning process. *California management review, 18(3),* 21-31.

Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review*, *47*, 49-63.

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., ... & Jonides, J. (2019). Does Counting Emotion Words on Online Social Networks Provide a Window Into People's Subjective Experience of Emotion? A Case Study on Facebook. *Emotion*, *19*(1), 97-107.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research, 39(1),* 19-37.

Levari, D.E., Wilson, T.D, & Gilbert, D.T. (2020) Advice from top performers feels (but is not) more helpful. *Working Paper.*

Louwerse, M., Lin, D., Drescher, A., & Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. In *Proceedings of the Annual Meeting of the Cognitive Science Society 32(32).*

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research, 30,* 457-500.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, *35*(4), 435-445.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law*, *11*(1), 99-122.

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2012). How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences, 3(2),* 30-50.

Milkman, K. L., Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2011). Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences, 108(26),* 10415-10420.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Nickerson, D. W., & Rogers, T. (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, *21*(2), 194-199.

Nook, E. C., Schleider, J. L., & Somerville, L. H. (2017). A linguistic signature of psychological distancing in emotion regulation. *Journal of Experimental Psychology: General, 146(3),* 337.

Nook, E. C., Stavish, C. M., Sasse, S. F., Lambert, H. K., Mair, P., McLaughlin, K. A., & Somerville, L. H. (2019). Charting the development of emotion comprehension and abstraction from childhood to adulthood using observer-rated and linguistic measures. *Emotion. In press.*

Paetzold, G., & Specia, L. (2016). Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 435-440).

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45(3),* 255.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology, 76(1),* 1.

Pan, L., McNamara, G., Lee, J. J., Haleblian, J., & Devers, C. E. (2018). Give it to us straight (most of the time): Top managers' use of concrete language and its effect on investor reactions. *Strategic Management Journal*, *39*(8), 2204-2225.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

Parhankangas, A., & Renko, M. (2017). Linguistic style and crowdfunding success among social and commercial entrepreneurs. *Journal of Business Venturing*, *32*(2), 215-236.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226-1227.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, *15*(2), 263-287.

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior research methods*, *50*(3), 1198-1216.

Querstret, D., & Cropley, M. (2013). Assessing treatments used to reduce rumination and/or worry: A systematic review. *Clinical psychology review*, *33*(8), 996-1009.

Rao, A. R., & Monroe, K. B. (1989). The effect of price, brand name, and store name on buyers' perceptions of product quality: An integrative review. *Journal of marketing Research*, *26*(3), 351-357.

Reyt, J. N., & Wiesenfeld, B. M. (2015). Seeing the forest for the trees: Exploratory learning, mobile technology, and knowledge workers' role integration behaviors. *Academy of Management Journal*, *58*(3), 739-762.

Reyt, J. N., Wiesenfeld, B. M., & Trope, Y. (2016). Big picture is better: The social implications of construal level for advice taking. *Organizational Behavior and Human Decision Processes, 135,* 22-31.

Rogers, R. K., Dillard, J., & Yuthas, K. (2005). The accounting profession: Substantive change and/or image management. *Journal of Business Ethics*, *58*(1-3), 159-176.

Rogers, T., Milkman, K. L., John, L. K., & Norton, M. I. (2015). Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behavioral Science & Policy*, *1*(2), 33-41.

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

Ross, L., & Nisbett, R. E. (2011). The person and the situation: Perspectives of social psychology. Pinter & Martin Publishers.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Schellekens, G. A., Verlegh, P. W., & Smidts, A. (2010). Language abstraction in word of mouth. *Journal of Consumer Research*, *37*(2), 207-223.

Seih, Y. T., Beier, S., & Pennebaker, J. W. (2017). Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology, 36(3),* 343-355.

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54(4),* 558.

Semin, G. R., Görts, C. A., Nandram, S., & Semin-Goossens, A. (2002). Cultural perspectives on the linguistic representation of emotion and emotion events. *Cognition & Emotion, 16(1),* 11-28.

Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods*, *11*(4), 727-752.

Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*, 415-435.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

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

Snefjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological science, 26(9),* 1449-1460.

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

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2019). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of personality and social psychology, in press.*

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., ... & Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology, 116(5),* 817.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology, 29(1),* 24-54.

Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, *62*(1), 78-97.

Townsend, C., & Liu, W. (2012). Is planning good for you? The differential impact of planning on self-regulation. *Journal of Consumer Research, 39(4),* 688-703.

Troche, J., Crutch, S. J., & Reilly, J. (2017). Defining a conceptual topography of word concreteness: clustering properties of emotion, sensation, and magnitude among 750 english words. *Frontiers in psychology*, *8*, 1787.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological review*, *110*(3), 403.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117(2),* 440.

Vallacher, R. R., & Wegner, D. M. (1989). Levels of personal agency: Individual variation in action identification. *Journal of Personality and Social psychology*, *57*(4), 660.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 91.

Wiener, M., & Mehrabian, A. (1968). *Language within language: Immediacy, a channel in verbal communication.* Ardent Media.

Wiesenfeld, B. M., Reyt, J. N., Brockner, J., & Trope, Y. (2017). Construal level theory in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 367-400.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Yeomans, M. (2019). Some hedonic consequences of perspective-taking in recommending. *Journal of Consumer Psychology*, *29*(1), 22-38.

Yoon, J., Blunden, H., Kristal, A. & Whillans, A. (2020). Framing Feedback Giving as Advice Giving Yields More Critical and Actionable Input. *Harvard Business School Working Paper, No. 20-021*.

Yoon, J., Whillans, A.V. & O'Brien, E. (2020). Connecting the Dots: Superordinate Framing Enhances the Value of Unimportant Tasks. *Harvard Business School Working Paper, No. 20-011*.

Zhang, T., North, M. (2018). Wunderkind Wisdom: Younger Advisers Discount Their Impact in Reverse Advising Contexts. In *Academy of Management Proceedings* (Vol. 2018, No. 1, p. 15416). Briarcliff Manor, NY 10510: Academy of Management.

Table 1: Qualitative Summary of Results from Linguistic Concreteness Measures.

Name of Measure	**Measurement Validity**				**Reproducibility**
	Descriptions	**Advice**	**Plan Distance**	**Plan Specificity**	
Brysbaert	Low	Low	Low	Low	Medium
Original MRC	Medium	Low	Low	Very Low	Medium
Bootstrap MRC	Low	Low	Low	Low	Medium
Immediacy	**Medium**	Zero	Very Low	Zero	Low
Larrimore-LIWC	Zero	Very Low	Very Low	Very Low	Low
Pan-LIWC	Zero	Zero	Very Low	Very Low	Low
Part-of-Speech LCM	Medium	Zero	Very Low	Zero	Low
Syntax LCM	Low	Zero	Zero	Very Low	**High**
DICTION	Very Low	Very Low	Zero	Zero	Low
n-Grams NLP Model	Low	**Medium**	**Medium**	**High**	**High**

Table 2: Example of word-level concreteness scores.

word	mTurk Ratings	Original MRC	Bootstrapped MRC
This	2.14	240	212.36
example	3.03	--	335.35
sentence	3.57	--	397.16
has	2.18	267	272.31
both	2.97	322	256.11
concrete	4.59	562	506.81
and	1.52	220	277.14
abstract	1.45	--	373.73
words.	3.56	--	389.48

Table 3: Summary of Datasets in Study 1

Dataset Name	Concrete Index	Goal	Sample Size	Word Count mean (sd)
Workplace Feedback	Annotated	advice	1334	20 (20)
Teacher Feedback	Randomized	advice	304	36 (19)
Personal Feedback	Annotated	advice	171	36 (21)
Letter Advice	Annotated	advice	951	32 (22)
Life Goals	Annotated	advice	301	36 (25)
Task Tips	Annotated	advice	228	38 (25)
Why Vs How	Randomized	description	195	61 (47)
Self-Distancing	Randomized	description	928	315 (120)
Emotion Words	Annotated	description	196	710 (440)

Table 4: Correlation with concreteness content (and 95% CI) for supervised machine learning models. Each cell represents an estimate of out-of-sample accuracy for a model trained on one dataset, and tested on another. On the diagonal cells where the training and test datasets are the same, we cross-validated by holding out different studies/courses one at a time.

	Test Dataset			
Training Dataset	**Study 1 Advice**	**Study 1 Descriptions**	**Study 2 Distance**	**Study 2 Specificity**
Study 1 Advice	**0.228 [0.195, 0.26]**	-0.113 [-0.166, -0.059]	0.004 [-0.024, 0.031]	0.258 [0.232, 0.283]
Study 1 Descriptions	0.119 [0.085, 0.152]	**0.092 [0.038, 0.145]**	0.012 [-0.015, 0.039]	0.417 [0.394, 0.439]
Study 2 Distance	0.022 [-0.012, 0.056]	-0.012 [-0.066, 0.042]	**0.339 [0.315, 0.363]**	0.026 [-0.001, 0.053]
Study 2 Specificity	0.191 [0.158, 0.224]	-0.032 [-0.086, 0.022]	0.038 [0.011, 0.065]	**0.733 [0.72, 0.745]**

Figure 1: Pearson correlations between linguistic measures of concreteness in Study 1, calculated separately for each dataset.

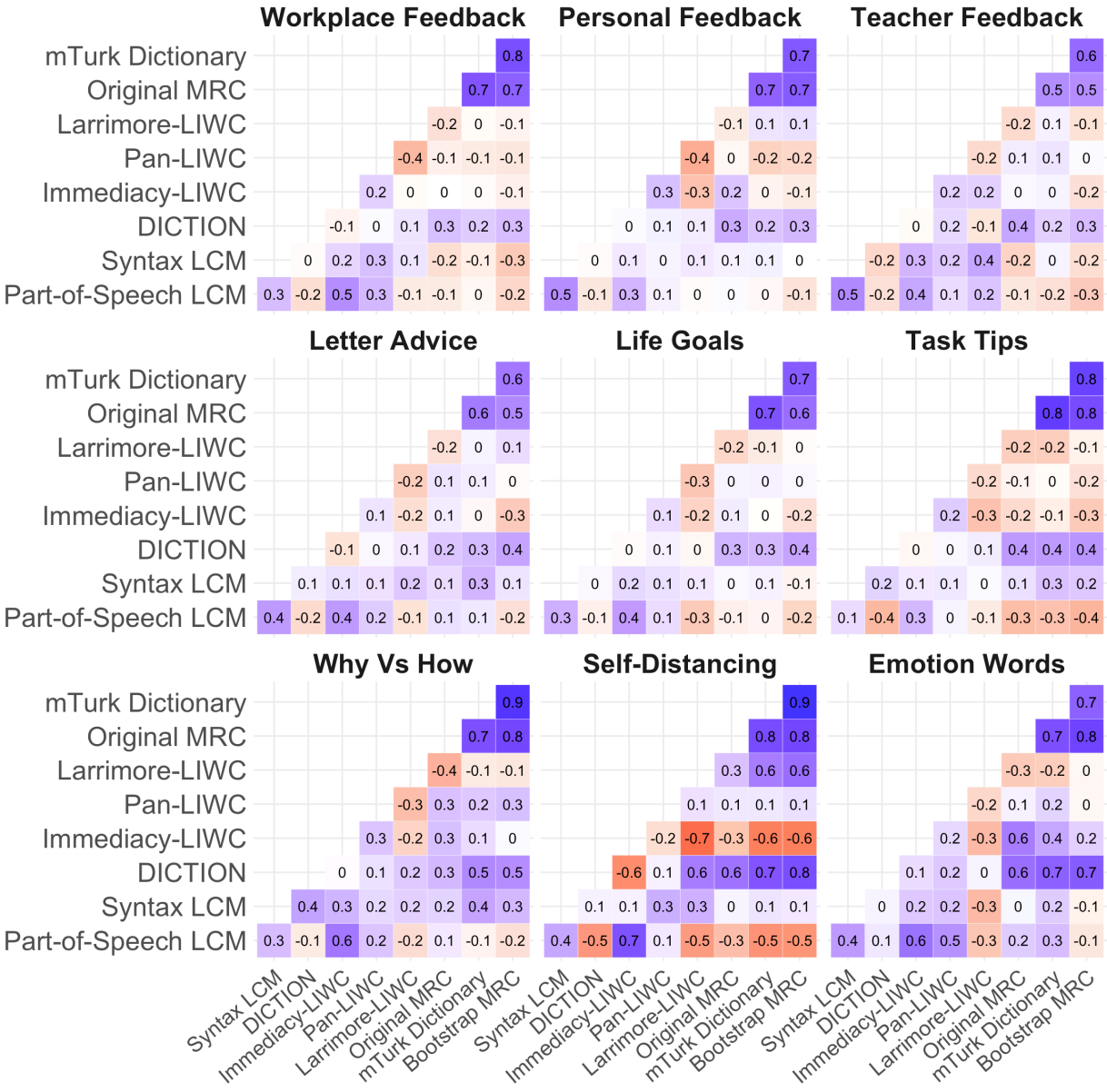


Figure 2. Correlation with concreteness content (and 95% CI) for linguistic measures of concreteness. The Y axis distinguishes different datasets, and each panel shows a different measure.

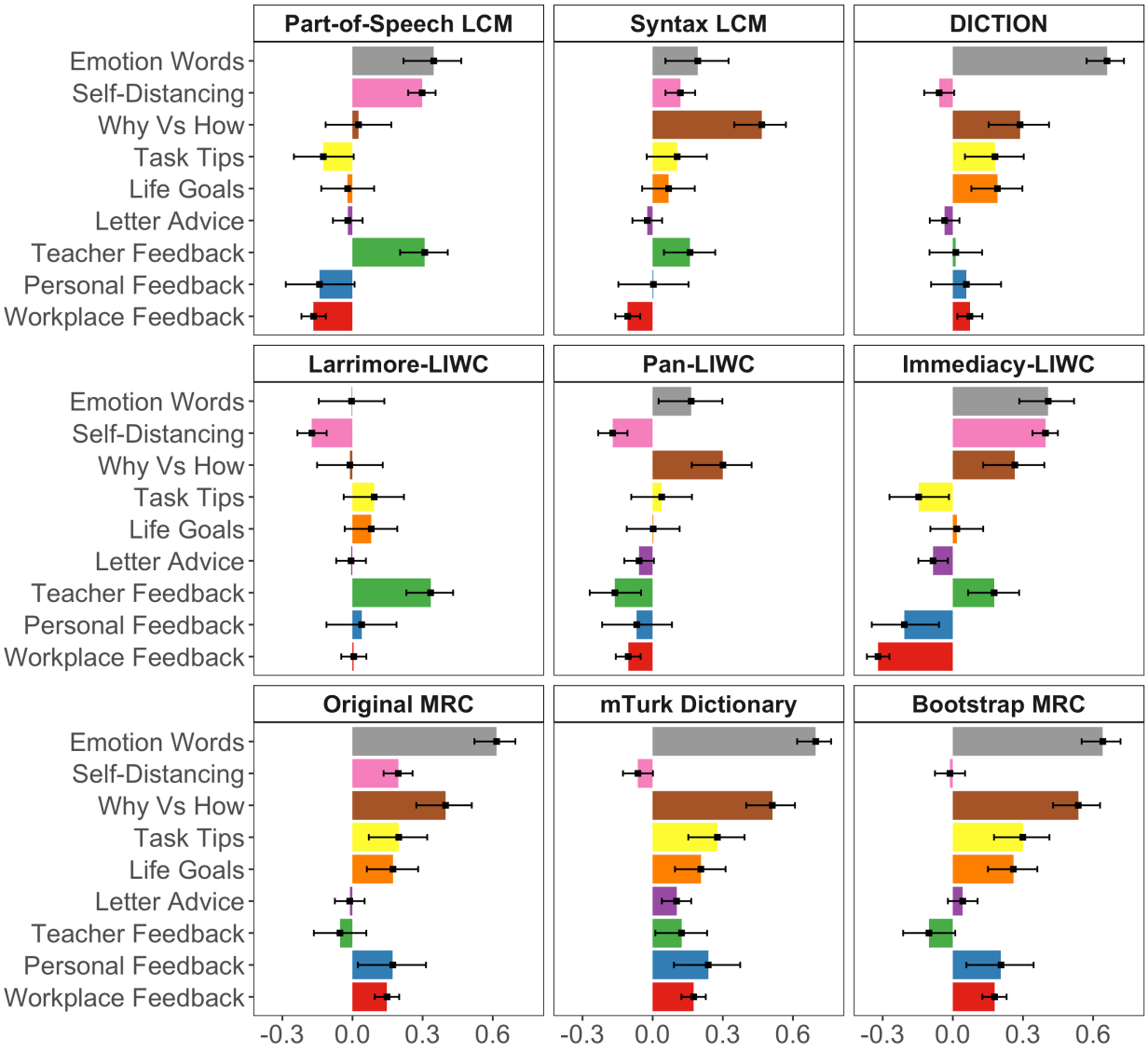
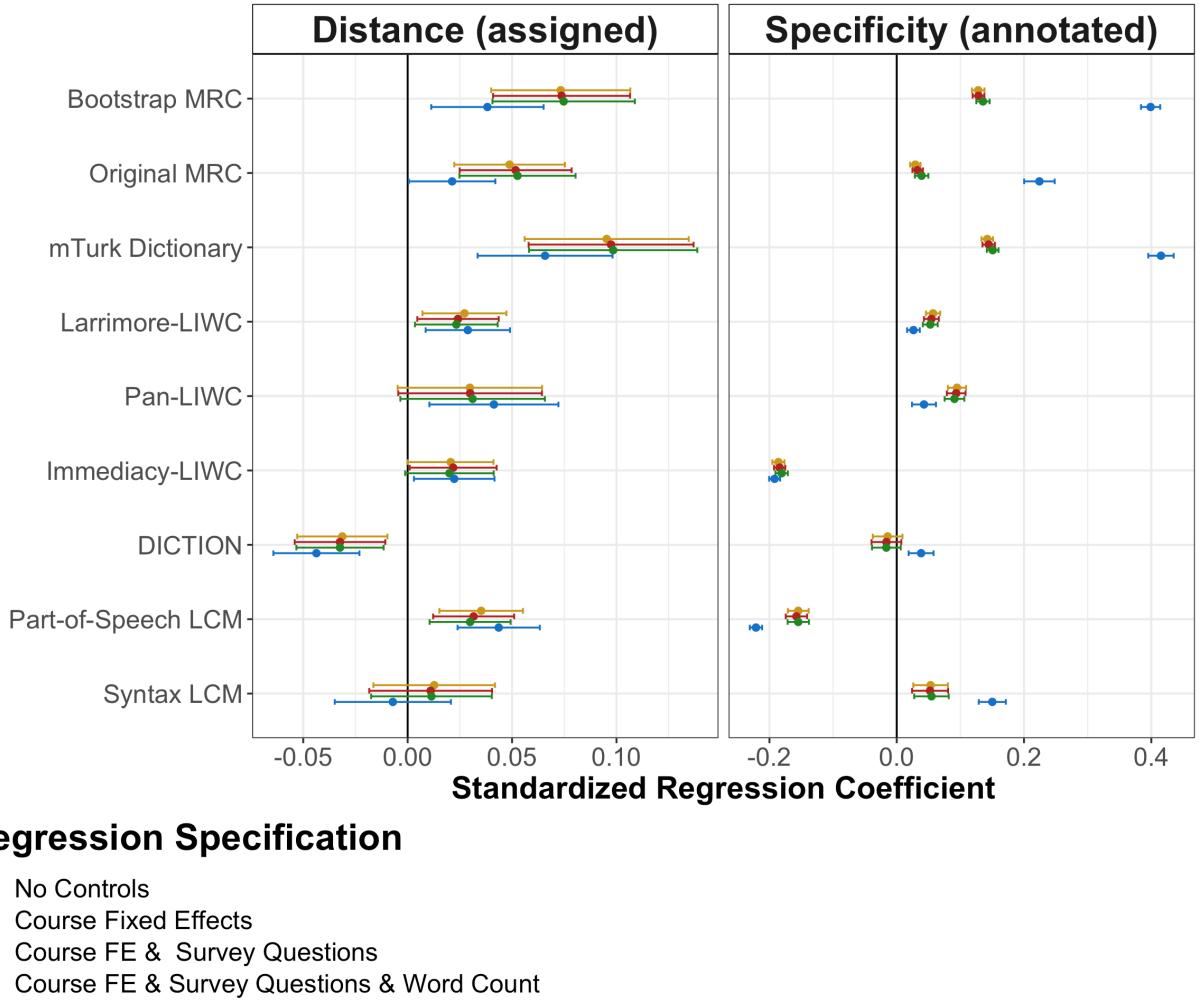


Figure 3. Relationship of concreteness to assigned plan distance (short-term vs. long-term) and plan specificity, as annotated by research assistance in Study 2. The Y axis distinguishes different linguistic measures, and the X axis represents a standardized regression coefficient and 95% confidence interval. Colors identify regression specifications that include different sets of control variables.



Appendix A: LCM Category Scores for Study 1

Figure A1: Correlation with concreteness content (and 95% CI) for part of speech categories from the Linguistic Category Model. Each panel compiles all the data in a different domain.

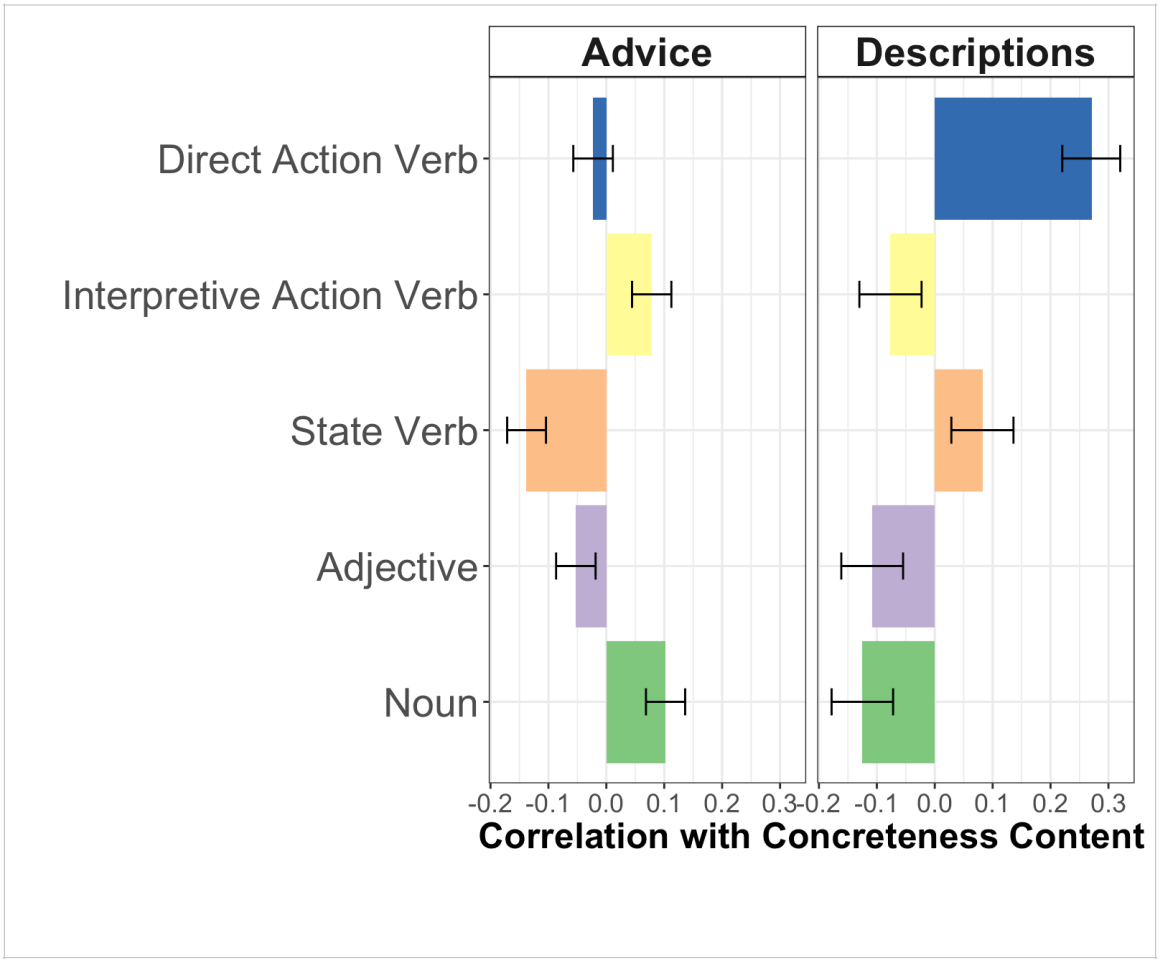
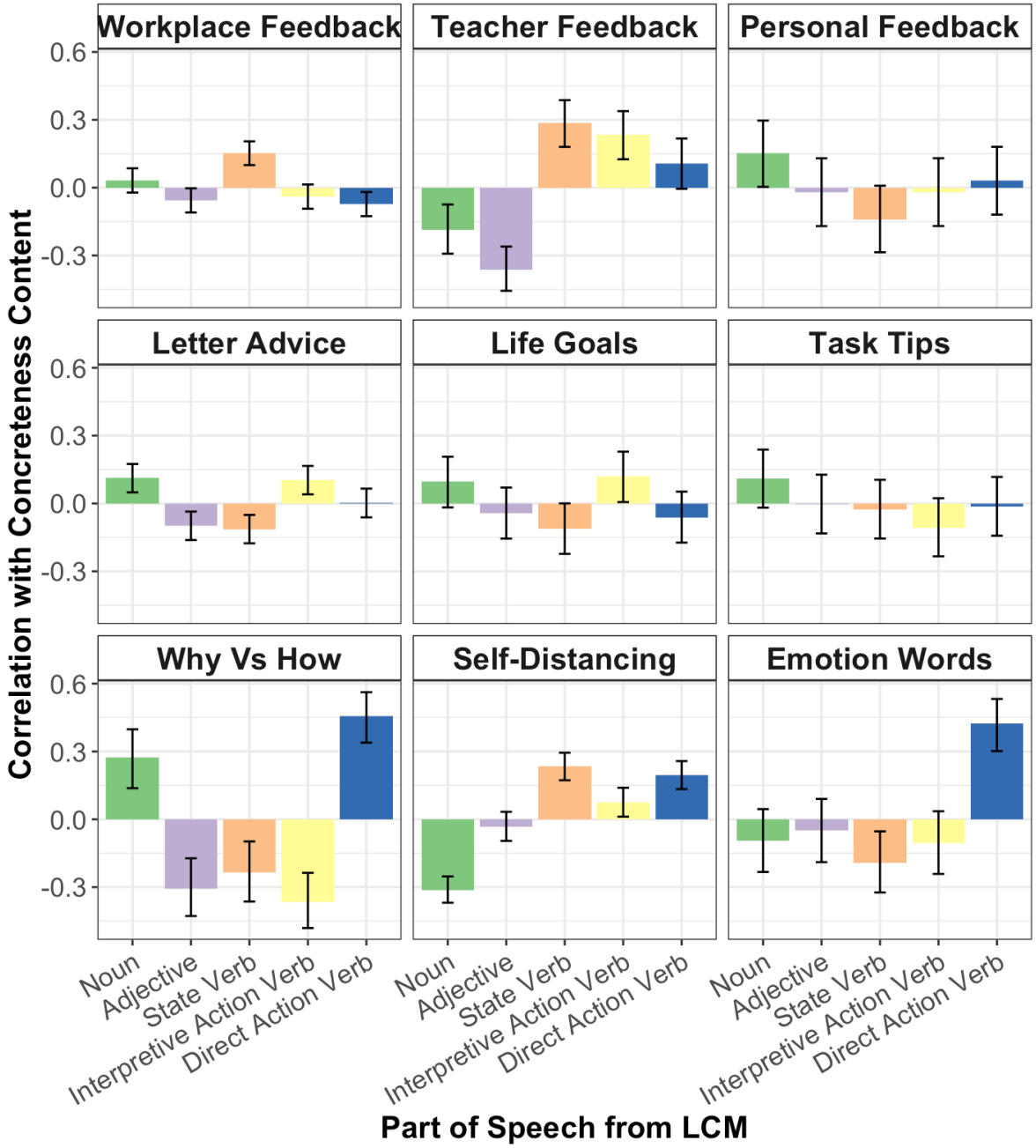


Figure A2: Correlation with concreteness content (and 95% CI) for part of speech categories from the Linguistic Category Model. Each panel compiles all of the data from a different dataset.



Appendix B: LIWC Category Scores for Study 1

Figure A3: Correlation with concreteness content (and 95% CI) for categories in the LIWC measures. The Y axis distinguishes LIWC categories, and the results are plotted separately for each domain in Study 1. Red bars are features that are supposed to be negative, and light blue bars are supposed to be negative, according to the original LIWC construct.

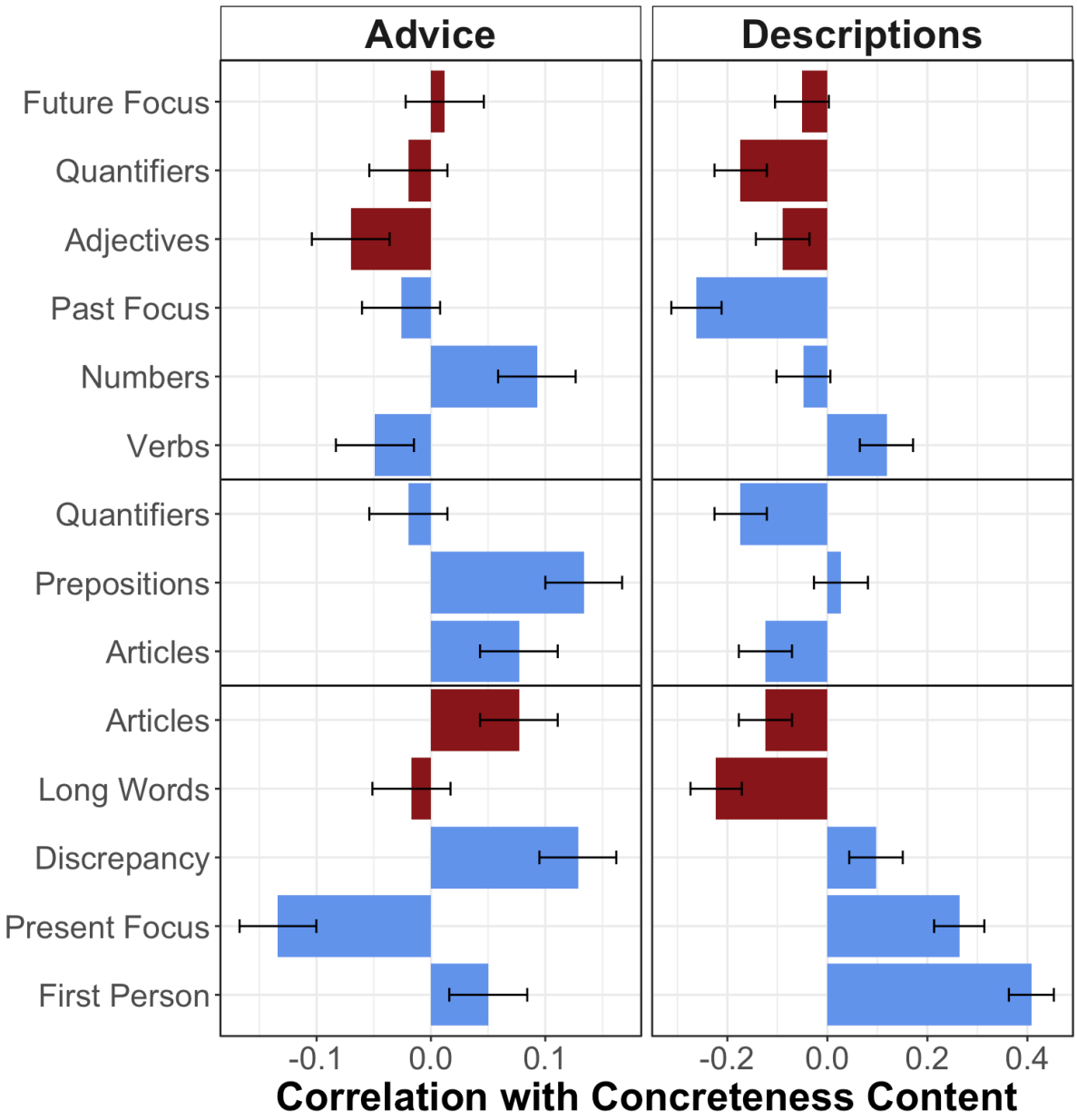
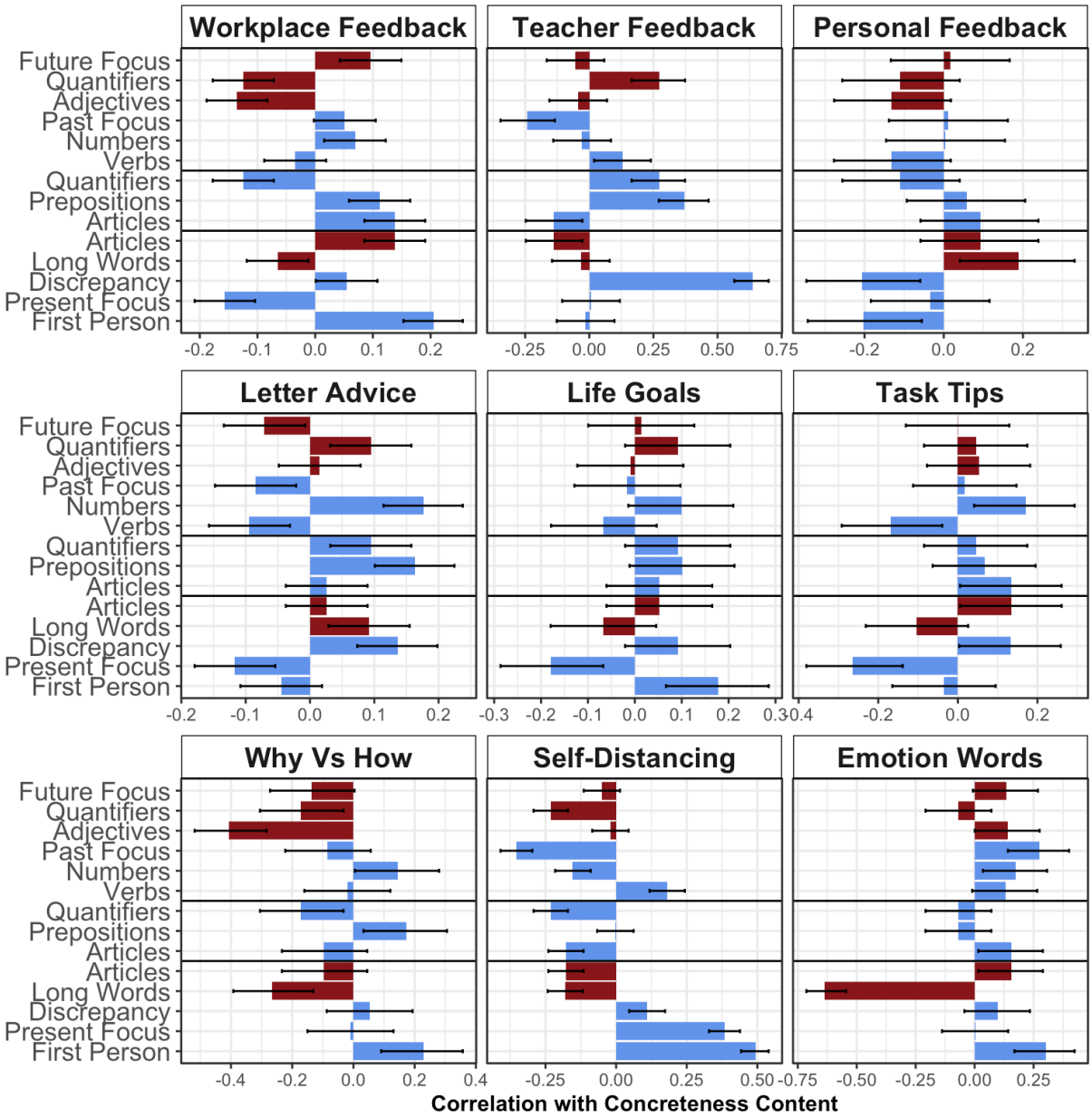


Figure A4: Correlation with concreteness content (and 95% CI) for LIWC concreteness measures. The Y axis distinguishes LIWC categories, and the results are plotted separately for each dataset in Study 1. Red bars are features that are supposed to be negative, and light blue bars are supposed to be negative, according to the original LIWC construct.



Appendix C: Full text of Study 2 Planning Prompts

Note: Both plan-making interventions were similar, and all text that differs between short and long conditions is *[italicized in brackets].*

Please write down a clear, concrete plan to follow through on your goals in *[the first week of]* the course. Plan-making can be a helpful tool in MOOCs! Successful students in previous courses have made detailed plans for how they will engage *[in the first week of / throughout]* the course.

In the text box below, write out your plans to complete tasks for the course *[this upcoming week]*. Please be as specific as you can!

[open text box]

You might find it helpful to consider these questions when you make your plans:

-When and where do you plan to engage with the course content?
-How much time will you spend studying in the *[first week / course]*?
-What will you do to ensure you complete the required course work?
-How will you overcome potential obstacles in the *[first week / course]*?

Here are some examples to inspire your plan-making (replace them with your own):

"I will watch videos Wednesday night*[s]* after work, and complete the readings on Saturday morning*[s]*."
"If I haven't done *[the/a]* week's work by Sunday, then I will prioritize the videos to stay on schedule."
"I will add these times to my calendar so that I don't forget."
"If I have trouble understanding the material, I will visit the class discussion forum."

---------- NEXT PAGE ----------

It's great that you have written down your plans. They will be a useful tool for overcoming difficulties and achieving your goals.

Take another look at your plans below. How will you make sure to remember them? For example, take a moment now to: write them down on paper, email them to yourself or a friend, add to a calendar with a reminder, or tell someone about them!

[text of plans piped in from previous page]

Appendix D: Full text of Study 2 Annotation Instructions

Your task is to provide human annotations for a set of plans that people have written for online classes. Participants were real students in real online classes, who were responding to this prompt. Note that it was randomized, so that participants were nudged to write plans for either the first week of class, or else the entire course. However, you will be blinded to their true condition, and in any case it is not strictly relevant to the dimensions you will be evaluating.

[text of prompt]

You will evaluate each plan on three dimensions.

Sincerity [0/1] - did this person actually attempt to write out their plans? Or did they simply dump enough text into the box to advance in the survey? Do not evaluate whether they are good plans – just ask whether they are plans at all.

Concreteness [1-7] - Is this plan concrete? Did this person's plans describe specific steps, like a recipe? Does it describe tangible concepts (i.e. things you can see, hear, smell, taste or feel), rather than intangible, abstract concepts (i.e. thoughts, goals, feelings, ideas)? Are the plans focused on the "hows" of class completion, rather than "whys" of class completion? Is it obvious how this person will fulfill their plans? Do you think it will be obvious to evaluate whether or not that person has fulfilled their plans afterwards?

Concreteness is split into two scales – self and other. They describe the same concept, but from the perspective of either the writer herself, or another student in the class (who us not the writer). The “self” rating should identify whether the plan seems actionable for the writer to carry out, while the “other” rating should identify whether the plan seems actionable for someone else who was given this plan.

Appendix E: Supervised Models and Training Set Size

Hand-labeled data are the most accurate way to detect concreteness in language, and enrich the results of automated methods. However, they are costly to collect, in several ways: time spent developing an annotation scheme and teaching annotators, paying for their time to read, in a way that preserves the original writers' privacy. Furthermore, researchers may sometimes want to estimate concreteness in a dataset that is much larger than they could feasibly annotate.

In theses cases, we suggests that researchers consider hand-labeling a portion of their data, and estimating a model to label the rest. However, it is not trivial to estimate how much hand-labeling needs to be done to produce a supervised model that is at least as accurate as an untrained off-the-shelf model. The right answer depends on many factors that will vary from context to context. However, we can use the data we have to at least benchmark this calculation in the domains of advice and plan-making. and labels during training improves accuracy on the rest of the set.

We conducted the same nested cross-validation procedure that was used in Section 5. And we test two feature sets across runs - bag of ngrams with and without the dictionaries added. But rather than use all available held-out data to train in each fold, we iteratively sampled a subset for training (50,100, 200, 400, 600, 800, 1000, or 1200). The result from each combination was produced from an averaged over five separate runs, to smooth out cross-validation error. In both studies, accuracy improved with training set size. However, our results suggest that even a training set of 200 is enough to outperform many domain-general models, and the gains from additional data tend to taper off after 500 or labels for our relatively simple algorithm.

Figure A5: Effect of training set size on accuracy of supervised models. All points represent the correlation with concreteness content (and 95% CI), pooled across all advice datasets from Study 1.

Figure A6: Effect of training set size on accuracy of supervised models. All points represent the correlation with concreteness content (and 95% CI), pooled across all annotated data in Study 2.