Concreteness, Concretely: A Case Study for Validation in Natural Language

Michael Yeomans

September 30, 2019

* In preparation for journal submission - please do not distribute without permission *

Abstract. Concreteness is central to psychological theories of learning and thinking, and increasingly has practical applications to domains with prevalent natural language data, like advice and plan-making. However, the literature provides diffuse and competing definitions of concreteness in natural language. In this paper, we develop a concrete definition of concreteness, to understand how concreteness is expressed during social goal pursuits. We review several proposed algorithms for detecting concreteness, and systematically compare these algorithms across datasets with ground truth concreteness labels from several domains, including written advice, and plan-making in online courses (total N = 9,780). We generate simple guidelines for automated concreteness detection within and across domains, which are provided in a corresponding R package, *doc2concrete.*

Keywords. concreteness; planning prompts; advice; goal pursuit; conversation

Disclosure. 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/

1. Introduction

1.1. Concreteness, Abstractly

Concreteness is deeply rooted throughout psychological theory, dating back to Jean Piaget and William James (Trope & Liberman, 2010; Burgoon, Henderson & Markman, 2013). It is central to cognitive models of learning, in which concrete sensory experiences are synthesized into increasingly abstract concepts and representations (Kolb, 1976; Paivio, 1991; Bengio, 2009). These internal representations also vary in their concreteness, which can be detected from the natural language descriptions people generate to express their thoughts (Snefjella & Kuperman, 2015; Bhatia & Walesek, 2016).

Behavioral science has also begun to apply models of concreteness to understand more complex social goal pursuits. For example, planning prompts are meant to encourage people to generate representations of their future actions, so as to better pursue their goals (Gollwitzer & Sheeran, 2006; Rogers et al., 2016). Likewise, advice is a means by which people suggest future actions to recipients, who can then better pursue their own goals (Ashford & Cummings, 1983; Bonnacio & Dalal, 2006; Berger, 2014). In both of these domains, natural language can improve the translation of high-level goals into low-level behaviors. Accordingly, concreteness is often suggested as a way to affect planners' follow-through, or advice recipients' future performance.

Advice and plan-making are two domains that demonstrate the potential value of understanding how concreteness is expressed in natural language. However, the academic literature is diffuse, and has not settled on any consistent guidance for how concreteness should be measured from text. Although many methods have been proposed, they have often been not been validated against objective measures of concreteness, or against one another. More importantly, the validity of different linguistic concreteness measures has never been systematically evaluated across domains, or contexts.

1.2. Overview of Current Research.

In this paper, we develop a concrete definition of concreteness - that is, to understand how well concreteness can be detected in natural language. We empirically estimate the generalizability of concreteness, by drawing strength from multiple datasets, linguistic models and theoretical frameworks. Our exercise highlights the hallmarks of open science - data and methods pooled from many different researchers - in order to address questions that can only be answered at a larger scale. Our goal is to understand the role of concreteness in two goal pursuit domains - advice and planning.

We first review several psychological models of linguistic concreteness, and their algorithmic implementations. In Study 1, we compare these algorithms across datasets from a variety of advice-giving and description experiments (9 studies, 4,608 participants) that have concreteness labels. In Study 2, 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). These results generate simple guidelines for automated concreteness detection within and across domains, which we roll into a new R package *doc2concrete* to reproduce these analyses in new texts.

2. Relevant Literature

2.1. Domains of Interest.

The concept of concreteness has been invoked to make sense of the psychology of many domains with rich natural language. This includes 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), and social media (Snefjella & Kuperman, 2015; Bhatia & Walesek, 2016). A complete review of all of these domains is not possible here. Instead we focus on two common sources of natural language during long-term goal pursuit - either for someone else ("giving advice") or the speaker herself ("making plans").

2.1.1. Giving 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; Bonnacio & 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, surely, any effect of feedback will depend 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 et al, 1979; Baron, 1988; Goodman, Wood & Hendrickx, 2004; Kraft & Rogers, 2015; Reyt, Weisenfeld & Trope, 2016).

This literature has almost exclusively relied on manipulated specificity, or else human-annotated specificity, to determine the concreteness of a piece of advice. However, advice is common in many domains, and it is possible that concreteness may be a structural or stylistic aspect of all advice, or simply a property of the particular content of a domain. To test this empirically, we combine data from several different organizational and educational contexts, each of which has a ground truth measure of concreteness. We compare whether the linguistic features of advice that demarcate concreteness are consistent from one context to another.

2.1.2. Making Plans. A long literature has found positive effects of generating plans as a means to follow through on one's current intentions for future behavior. 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., 2016). However, 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).

But 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; Dai et al., 2018; Beshears et al., 2019). In open-ended domains, the many facets of concreteness as a measure are complicated by the many facets of the concept itself. Several dimensions can vary in their concreteness, which may or may not correlate with one another, and which may or may not carry the same linguistic features. To test this empirically, we collect two different measures of concreteness in the same plan-making dataset. We then train separate machine learning models to see how these two concreteness measures relate to the language of the plans, and to one another.

2.2. Linguistic 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 prominent examples of both below. For clarity throughout, all measures are oriented so that higher numbers indicate more concreteness, which means some models (e.g. the "abstractness index") are reversed from their original orientation.

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" vs. "friday"; "it" vs. "you"). However, homonyms ("leaves", "bank", "like", etc.) are muddled. More importantly, this approach cannot capture any aspects of concreteness that are compositional, or contextual, or subjective.

The more recent 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 & Walesek, 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 (Colthart, 1981), and a more recent dictionary that uses bootstrapped embeddings to algorithmically extrapolate the original MRC to 85,000 words (Paetzold & Specia, 2016). An example for each of these dictionaries is demonstrated in Table 1.

Dictionaries are defined and validated for single words, but for documents the individual word scores must be combined and weighted into a summary score. Previous research has primarily used unweighted averages across all matches for document-level scores (Snefjella & Kuperman, 2015; Bhatia & Walesek, 2016), which we adopt as a baseline. Although these previous applications did not include stop words ("it", "you", "where", "how") or numbers ("one", "ten"), we always include both, since they are likely relevant in our domains of interest.

2.2.2. Linguistic Category Model. The categorical measure most commonly associated with concreteness is the Linguistic Category Model (henceforth "LCM"; Semin & Fiedler, 1988), which 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 "abstract" is both a noun and an adjective; as is the word "concrete". However, it was developed from lab experiments that focused on texts from descriptions of people, where syntactical variation may be limited.

Originally, the LCM was developed to be annotated by hand, However, this is only practical on smaller sample sizes. However, automated grammar parsing has been improving substantially, for a variety of NLP tasks (Manning et al., 2014; Honnibal & Johnson, 2017). 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 simply be tallied according to the original LCM formula (Seih, Beier & Pennebaker, 2017), as an approximation of the more sophisticated human process.

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, we note that 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 automated grammar parsing to process the documents. Like the part-of-speech LCM, they offer this scoring rule from this model as a potential measure of concreteness in all domains. We test the generalizability of the pre-trained model, and also conduct exploratory analyses using the syntax features by re-training new models in our domains of interest.

2.2.3. Categorical LIWC Constructs. We test several categorical models developed from the 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. Specifically, previous work has suggested that combinations of these lists can approximate concreteness in natural language. We focus on three examples that have been used in multiple papers, which produce results representative of the larger approach - however, we acknowledge the

LIWC allows for authors to mix and match their own combinations of constructs, which have sometimes been used under the guise of concreteness (e.g. Pan et al., 2018).

One combines five categories - first person singular; present focus; discrepancies; long words (reversed); articles (reversed) - to form a measure of "verbal immediacy" (Mehl, Robbins & Hollerans, 2012). This was developed from prior conceptual work (Wiener & Mehrabian, 1968) on clinical responses to traumatic responses or events. Recent work has built on this to example retrospective self-distancing descriptions of experiences/stimuli, though it has also been used with language collected from email and experience sampling methods (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, and has been used in other domains as well (Larrimore et al., 2011; Toma & Hancock, 2012; Markowitz & Hancock, 2016). Finally, a set of five features - perception; space; time; affect; cognition (reversed) - has been used to estimate "reality monitoring", as a theory of concreteness in memory encoding and retrieval (Johnson & Raye, 1981; Johnson et al., 1993; Bond et al., 2017).

3. Study 1: Concreteness in Advice

In our first study, we tested linguistic concreteness across many different contexts, that all have some natural language with a ground truth "concreteness index". This sample of studies is not intended to be representative, and was mainly gathered through informal conversations (see Table 2). Our primary objective in this search was

to collect text where the goal was to give advice or feedback. However, we also include some datasets where the writer's goal is to describe a stimulus, as asked by the experimenter. Description language allows for tight control over what is being described, which makes it (by far) the most common form on text collected in psychology experiments. However, it is not always clear what goal writers are pursuing so the external and prescriptive validity can be limited.

3.1. Study 1 Datasets

3.1.1. Workplace Feedback. Employees (387 total) 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 anonymously. 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.

3.1.3. Teacher Feedback. Middle school students (283 total) 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 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, 2019). 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, 2019). 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, 2019). 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, 2019). 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 what they did, while the other half were told to explain "why" what they did. This task is commonly used as a mindset induction in construal level research, used over a variety of domains and measures (e.g. Freitas et al., 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 can detect the ground truth labels (the "concreteness index") within each dataset. 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 models received a similar transformation, calculated separately for each dataset.

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 .639 to .738. The two LCM measures are always positively correlated, but not strongly so, with an average correlation of .332. There are also a surprising number of negative correlations across all of the categorical models, with some datasets in particular stoking concern - 16 of the 28 pairwise correlations between models are negative in the Task Tips dataset, and 6 of the 26 in the Self-Distancing dataset were below -0.4. In general, this result suggests that concreteness often does not even correlate with concreteness.

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. However, word count is zero-bounded and right-skewed, and 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 $R^2 = .260$).

3.2.3. Ground Truth Concreteness. We first estimated the concreteness of a texts' content, controlling for log-transformed word count, using a hierarchical linear model (Bates et al., 2105). 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 ground truth measure of concreteness of the content of each document (all of our results are substantively similar if we use the unadjusted concreteness scores as our ground truth 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 variability across measures and datasets.

Consider the perspective of a researcher studying the Emotion Words paradigm - the literature on linguistic concreteness might seem like an embarrassment of riches, with most measures agreeing with the ground truth labels, and with each other. However, the perspective of a researcher studying the Workplace Feedback paradigm is considerably more bleak. Some measures correlate with concreteness positively, others negatively, and others not at all - and yet there is little guidance on how to resolve these apparent contradictions.

All of the dictionaries 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 mTurk dictionary ($r = .155$, 95% CI = [.121, .188]; $t(3287) = 9.0$, $p < .001$) than either of the MRC-based dictionaries (Bootstrap: $r = .116$, 95% CI = [.082, .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]; $t(1317) = 11$, $p < .001$; mTurk: $r = .131$, 95% CI = [.078, .184]; $t(1317) = 4.8$, $p < .001$; Bootstrap: $r = .163$, 95% CI = [.110, .215]; $t(1317) = 6.0$, $p < .001$).

The results are less positive for the categorical models. Some categorical measures perform well on the pooled description datasets (Immediacy: $r = .325$, 95% CI = [.276, .373]; $t(1317) = 12$, $p < .001$; Part of Speech LCM: $r = .264$, 95% CI = [.214, .314]; $t(1317) = 10$, $p < .001$; Syntax LCM: $r = .181$, 95% CI = [.128, .233]; $t(1317) = 6.7$, $p < .001$). However, when the advice datasets are pooled, those same categorical models either found no or opposite effects of concreteness, on average (Immediacy: $r =$

-.074, 95% CI = [-.107, -.039]; t(3287) = 4.2, p < .001; Part of Speech LCM: r = -.034, 95% CI = [-.068, .000]; t(3287) = 1.9, p = .052; Syntax LCM: r = .003, 95% CI = [-.031, .037]; t(3287) = 0.2, p = .849).

3.2.4. Category-Level LCM Concreteness. One concern we had during our review of the Linguistic Category Model was that scoring rules varied. However, there was consistency on the relative ordering of the categories - nouns (when they were included) were always the most abstract, adjectives (when they were included) were next, and so on. So rather than iterating through every possible scoring rule, we tested each category separately and empirically estimated a scoring rule, which is plotted in Figure 3. The results are grouped by domain, although each dataset is also plotted separately in Appendix A. The description tasks mostly validate the linguistic category model, as the correlations roughly line up in ascending order (although the Interpretive Action Verb category seems to be somewhat more abstract than the theory predicts). However, the advice datasets stand in stark contrast. It is hard to identify any previous LCM scoring rule that is consistent with these results.

3.2.5. Category-Level LIWC Concreteness. We perform a similar exercise with the combined categories of all the LIWC constructs. In all these models, each category should be weighted equally, with either a positive or negative sign. In Figure 4, we plot the correlation with concreteness content for the features inside each construct, separated by domain (each dataset is plotted individually in Appendix B). The most

notable result is the sheer variability across domains - 7 of the 13 features were estimated in opposite directions for advice and description tasks.

The categorical breakdowns reinforce the summary statistics in Figure 2, suggesting that categorical models fail to consistently capture the concreteness of advice. For example, the five immediacy features closely reproduce previous results in description tasks (Nook et al., 2017; 2019). However, the categorical breakdown also explains why it fails to detect concreteness index in all of the advice datasets. Across all three LIWC constructs in the advice datasets, four features were correlated in the predicted direction (Prepositions, Articles, Discrepancies; and First Person); three features were correlated in the opposite direction (Present Tense, Articles, and Affect) while the rest of the correlations are too small to interpret.

3.3. Study 1 Discussion

Concreteness is broadly ingrained across many psychological models of social learning. Recent work has suggested several approaches to detecting concreteness in language, each developed in separate contexts. We compare a wide range of concreteness 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 large-scale dictionary methods were somewhat reliable across domains, though the effect sizes were typically small.

We also found results that were consistent across datasets, but not across domains. While some of the categorical linguistic models (Immediacy, Part of Speech LCM) were able to detect concreteness in the description datasets, they mostly failed to detect concrete advice. In fact, the Linguistic Category Model was initially proposed for measuring trait descriptions (Semin & Fiedler, 1991), which are not included here. The closest is perhaps the Teacher Feedback dataset, in which teachers wrote feedback to parents about their children, rather than to the students themselves. It is interesting that across the individual advice datasets, this third-person feedback was the dataset on which the categorical models all performed best.

4. Study 2: Plan-Making in Online Education

In Study 2 we turn to a new social goal pursuit domain - plan-making. Like advice, plan-making is an expression of someone's (i.e. the self's) desired future behavior towards a goal, and plan-making has long been thought to be better when it is concrete and specific (Gollwitzer & Sheeran, 2006; Rogers et al., 2016). However, much of that work has manipulated plan specificity, in pursuit of one-time actions like a doctor's visit, or voting (Nickerson & Rogers, 2010; Milkman et al., 2011). But it is also possible that over longer goal pursuits, flexibility and persistence are more important, in which concreteness might be ineffective, or even counterproductive (Townsend & Liu, 2012; Dai et al., 2018; Beshears et al., 2019).

In Study 2 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., 2019). 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 compare the written plans against two ground-truth measures of concreteness: 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 for a separate project with this dataset, including exclusion criteria, and model specifications. The preregistration (which focused on treatment effects) did not include any text cleaning, which is necessary for any trace data. For this research, we created an automated 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 was specifically actionable for the writer herself, and whether it would be specifically actionable 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 annotated - however, we confirm the results are robust across the larger sample of 151 classes from the original study.

4.2. Study 2 Results

Our analyses follow the preregistration, by including course fixed effects and clustering standard errors at the course level. However, for robustness we also systematically vary some details of the model specifications (and generally find similar results). In some cases, we do not include course effects. In other cases, we also include control covariates - expected hours/week, intention to pass, previous MOOCs

completed, date of enrollment - that were collected before the planning prompts (and included in the pre-registration).

4.2.1. Word counts. Following Study 1, we also included log-transformed word counts in some of the regression specifications, for robustness. In general, the average specificity ratings were positively correlated with the log-transformed word count (β = .762, SE = .048, $z(5170) = 16$, $p < .001$). However, long-term plans had higher word counts, on average, than short-term plans (β = -.147, SE = 0.39, $z(5170) = 3.7$, $p <$ .001).

4.2.2. Plan Distance. In Figure 5 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, and the Brysbaert dictionary performed weakly better (β = .101, SE = .040, $z(5170) = 2.5$, $p = .013$). Interestingly, the Reality Monitoring construct showed a strong effect (β = .151, SE = .047, $z(5170) = 3.3$, $p =$ .001), driven primarily by the time subscale. However, the other categorical models showed no significant relationship with plan distance.

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 plan distance on annotated specificity (β = .008, SE = .035, $z(5170) = 0.2$, $p = .812$).

In Figure 5, we also plot the relationship between the linguistic models of concreteness and the specificity ratings. The two dictionaries were once again consistent, and the Brysbaert was directionally the closest to ground truth specificity ratings ($\beta = .147$, $SE = .010$, $z(5170) = 14$, $p < .001$). Syntax LCM and Abstractness found much smaller correlations with specificity, and both Part of Speech LCM and Immediacy were directionally the opposite of the ratings.

4.3. Study 2 Discussion

Study 2 extended our understanding of linguistic concreteness by comparing different models across multiple measures of concreteness within the same dataset. Like Study 1, the word-level concreteness models were more reliable indicators of both kinds of linguistic concreteness. However, results also showed that concreteness itself is multifaceted, and a manipulation of concreteness (via temporal distancing) had no effect on our annotated measure of concreteness (via specificity). This suggests that there may be underlying variation in the construct itself.

5. Domain Specificity of Concreteness.

An implicit assumption of all off-the-shelf concreteness models (and many off-the-shelf text analysis tools) is that they are generalizable. That is, they apply the same scoring rules to all text, and purport to measure the same construct, regardless of the speakers or their goals. That is, the boundary conditions of these measures are often left unclear. However, the results above suggest that there may be substantial domain-

level differences in concreteness. This could also explain why some of the off-the-shelf linguistic measures from previous research are not robust in our data.

We conducted 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, although their outcomes are different). We then compared the accuracy of those models in each of the other domains. This gives us an estimate of how similarly concreteness is expressed in each domain.

For each domain, we trained a supervised machine learning model to predict concreteness from the text. For features, we used bag-of-ngrams counts (all 1-, 2-, and 3-word sequences, including stop words, that occur in more than 1% of documents) and summary scores from the two dictionaries (the results are similar with the politeness and syntax features). The estimation algorithm we used was a 20-fold LASSO algorithm (Hastie, Tibshirani & Friedman, 2010). When the training domain matched the test domain, we used a nested cross-validation loop, treating each context - datasets in Study 1, or courses in Study 2 - as an outer fold (Varma & Simon, 2006). When the training domain and test domain differed, all of the training domain data was used to estimate the model.

In Table 3, 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. Indeed, the in-domain results underestimate the true stability because all

predictions are generated from in-domain but out-of-context models. However, the results cast doubt on the scope of any reliable domain-general model of concreteness. For example, while plan distance seems to be reliably detected from class to class, it does not seem to share similar features with any other concreteness domain in our data.

6. Concreteness Detection in Practice

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 discuss two approaches that can be used off-the-shelf, without any new annotations. Those approaches are: pre-trained models in domains with good training data; and as a last resort, a dictionary-based domain-general model. As a starting point for other researchers, we implement both of these off-the-shelf approaches in an open source R package, *doc2concrete.*

6.1. Annotated in-domain data. Hand-labeled data are usually more accurate than domain-general measures, because humans are often able to account for the domain in their interpretation of the text. Researchers may fairly be concerned that annotation does not come cheaply. However, even when researchers only have annotations for part of their dataset, they can train a model on the labeled data and apply predicted annotations in the unlabeled data. In Appendix C, we benchmark the effect of training set size on accuracy in our own advice and plan-making data.

Broadly, our results suggest that our simple models tend approached their in-domain accuracy plateau around 500 labels. This is a rough guide, that will surely vary based on domain, population, and task. However, it gives some context to existing validation exercises. For example, the Syntax LCM model was developed on a similarly-sized dataset (Johnson-Grey et al., 2019), and the LIWC's part of speech LCM was initially tested on a dataset roughly half that size (Seih et al., 2017). However, this suggests that validation sets that are orders of magnitude smaller (e.g. Pan et al., 2018) will be underpowered, and unlikely to accurately estimate in-domain validity.

Annotations are even more useful as a complement to automated methods, rather than just a substitute. Human annotations are typically given holistically, without explanations as to why a particular label was given. But when paired with natural language processing, we can carefully scrutinize how a construct is expressed in text. As an example, we use the politeness R package, which identifies a small set of stylistic and structural features of social language (Yeomans, Kantor & Tingley, 2018). In Figure 6, we conduct a visual analysis of the usage rates of various politeness features in the top- and bottom-third most concrete documents. On average, concrete advice is more negative than positive in emotional tone, and can be achieved both through more concrete language (e.g. bare command verbs) and less abstract language (emphatic praise, minimizing instructions with "just"). This example shows that a summary score is just the beginning of how researchers could treat text as data. Direct measures of language can help us intervene along the annotated dimension in people's conversational choices.

6.2. Pre-trained in-domain models. Our primary interest was in the practical consequences of concreteness in two applied domains - plan-making and advice. If other researchers share our interest in this domain, it may be unlikely that they will also collect training data on a similar scale to ours. To that end, we also provide two pre-trained models, which are intended to apply only to concreteness in the domains of advice or plan-making, respectively. Specifically, our 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.

6.3. Domain-general model. Although it is not ideal, researchers may have to rely on a domain-general model if they are in an unfamiliar domain, or conducting exploratory work. In this case, our results suggest that the mTurk dictionary provides the most robust measure of concreteness across the domains we tested here (Brysbeart et al., 2014). We offer an implementation of this dictionary, with some adjustments to the standard protocol.

Previous practice commonly excluded documents with insufficient word counts, and produced skewed distributions (i.e. short documents had much higher variance). Instead, our package suggests quasi-bayesian smoothing, which calculates a weighted combination of each document's raw score and the group average, with the weight proportional to document length. This smoothing somewhat improved the accuracy of the model for advice (raw: r = .154, 95% CI = [.121, .188]; smoothed: r = .156, 95% CI =

[.123, .189]) and for plan specificity (raw: β = .147, SE = .010; smoothed: β = .163, SE = .011). We suspect there may be other gains from fine-tuning the standard dictionary approach (for example, varying the weights on words) that should be explored in future research.

7. 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 context-specific, and multifaceted. This underscores the value of supervised machine learning as an empirical benchmark for theory-driven constructs in observational data.

7.1 Specific Lessons for Concreteness.

Our cross-domain approach provides useful context for some widely-used off-the-shelf measures. Our results provide tentative support for the dictionary methods as a weak-but-robust measure of concreteness across domains (Brysbaert, Warriner & Kuperman, 2016; Paetzold & Specia, 2017). However, our tests of the other off-the-shelf measures were less promising. There were some domain-specific successes - immediacy and LCM as a measure of description concreteness; and reality monitoring

as a measure of plan distance. Apart from those isolated cases, however, we failed to find any robust relationship with concreteness among the LIWC and LCM constructs.

Our results are not an attempt to replicate previous findings, but to test whether they generalize. In many cases, the evidence for that generalization was weak. Domain specificity is common, even in the most basic linguistic phenomena (e.g. Hamilton et al., 2016). We suspect that constructs may be especially domain-specific in social 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 is approached differently by the advisee and the advisor (Eggleston et al., 2015; Yeomans, 2018). Likewise, our linguistic model could at best estimate a generic model of how plans are expressed, because it did not include any individuating student characteristics (intentions, availability, experience, patience, etc.).

Our work also raises questions about the underlying psychological model of concreteness. That is, we cannot conclude whether the cause of domain specificity is related to concreteness as a behavioral measure, or to concreteness as a psychological construct. We are certainly not the first to suggest that concreteness may be multifaceted, or domain-specific (Trope & Liberman, 2010; Troche, Crutch & Reilly, 2017; Borghi et al., 2017; Pollock, 2018). However, concreteness may still characterize a domain-general cognitive architecture (e.g. Paivio, 1991), even if its expression in language is domain-specific. Unfortunately, almost all research on concreteness in natural language is conducted one domain at a time. One of the contributions of a large-

scale, multi-domain study like ours is that we can actually estimate the degree of generalizability on a level playing field.

7.2. Abstract Lessons for 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; Benoit, Munger & Spirling, 2019; Tackman et al., 2019). Of particular note are two recent studies that failed to replicate a long-hypothesized correlation between deception and concreteness (Kleinberg et al., 2019; Calderon et al., 2019). Both papers use off-the-shelf methods across large samples from different contexts, and conclude that linguistic concreteness is not systematically correlated with deception. Unfortunately, they do not discuss whether linguistic concreteness is systematically correlated with concreteness.

A similar tension arises whenever a linguistic model, trained to detect a construct in one domain, is transferred to another. Often, domain-general tools (like dictionaries) are cited as having "been validated", without describing the domain in which it was validated. Our results suggest that claims of unconditional validation are often unwarranted. The LIWC was primarily validated on experimenter-prompted descriptions - of memories, feelings, other people, and so on - rather than any kind of natural language generated to pursue social goals (Tausczik & Pennebaker, 2010). This contrast is not unique to concreteness. 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 (Sun et al., 2019; Kross et al., 2019). Description tasks typically constrain the topic (e.g. "what did you think about this product?"), which reduces the distribution of words, and goals and increases internal validity in that context. However this can come at the cost of external validity in open-ended natural language tasks.

Construct validity is a classic psychometric concern (Cronbach & Meehl, 1955) but it is a particularly vexing problem in natural language research. Unstructured text is extremely high-dimensional - the same document could be quantified in an essentially infinite number of ways, which makes family-wise correction techniques meaningless. Furthermore, the meaning of individual words is compositional, and domain-specific, so clever researchers can easily cherry-pick examples from broad word categories to arguably demonstrate any construct (or its opposite).

Authors of new language constructs are justifiably wary of discussing boundary conditions. While it is inevitable that a model will fail in some domains, in principle it could be applied to any text. But this puts a large burden on future researchers, who must vigorously report the null results of many tested constructs, that may not relate to the central claims of the paper writ large. In practice, this is likely to lead to a massive file drawer problem. A complex, but comprehensive, solution would be for more systematic reviews like this one, that combine datasets from many domains. But that will require authors to be more open to sharing their data and code with one another. A simpler solution would be to assume, by default, that language models are invalid in domains where they have not yet been validated.

7.3. 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). While human annotations will remain the gold standard for the foreseeable future, annotations are often costly, slow, and uninterpretable. 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). These two reasons - scalability and ecological validity - make a compelling case for all kinds of social science researchers to become more familiar with text as a source of data.

However, this paper demonstrates that this tremendous research opportunity also comes with unique challenges for careful researchers. 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. Automated model selection and validation becomes necessary for robustness when measuring constructs in high-dimensional data.

We argue that our field should consider broad adoption of machine learning techniques - regularization and transfer learning, in particular. Practically, this means

researchers can offload some of the effort and guesswork of construct validation. Instead of parsing the literal meaning of particular words, researchers can spend more time trying to understand the goals of the writers themselves. 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

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.

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

Beshears, J., H.N. Lee, K.L. Milkman, R. Mislavsky, & J. Wisdom (2019). Creating Exercise Habits Using Incentives: The Tradeoff between Flexibility and Routinization. *Working Paper.*

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).*

Bond, G. D., Holman, R. D., Eggert, J. A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., ... & Rustige, R. (2017). 'Lyin'Ted','Crooked Hillary', and 'Deceptive Donald': Language of Lies in the 2016 US Presidential Debates. *Applied Cognitive Psychology, 31(6),* 668-677.

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

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.

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*

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.

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

Dai, H., Dietvorst, B. J., Tuckfield, B., Milkman, K. L., & Schweitzer, M. E. (2018). Quitting When the Going Gets Tough: A Downside of High Performance Expectations. *Academy of Management Journal, 61(5),* 1667-1691.

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.

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.

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.

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.

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

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.

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

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hausser, J., & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research, 10(Jul), 1469-1484.

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.

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.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological review, 88(1),* 67.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological bulletin, 114(1), 3.

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.

Kizilcec, R., Reich, J., Yeomans, M., Lopez, G., Rosen, Y., Dann, C., Brunskill, E. & Tingley, D. The Limits of Scalable Interventions: A Case Study from Massive Open Online Courses. *Working paper.*

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. (2019) Advice from top performers feels (but is not) more helpful. *Working Paper.*

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance.* Prentice-Hall, Inc.

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.

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

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).

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.

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 observerrated and linguistic measures. *Emotion. In press.*

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.

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).

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.

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.

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.

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

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(1),* 82.

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.

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

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.

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.

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.

Yeomans, M., & Reich, J. (2017, March). Planning prompts increase and forecast course completion in massive open online courses. In P*roceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 464-473). ACM.*

Yeomans, M., Kantor, A., & Tingley, D. (2018). The politeness Package: Detecting Politeness in Natural Language. *R Journal, 10(2).*

Yoon, J., Blunden, H., Kristal, A. & Whillans, A. (2019). Seeking Constructive Feedback? Ask for Advice Instead. *Working Paper.*

Yoon, J., Whillans, A.V. & O'brien, E. (2019). Connecting the Dots: Superordinate Framing Enhances the Value of Unimportant Tasks. *Harvard Business School Working Paper, No. 20-011*.

Zhang, T., North, M. (2019). Wunderkind wisdom: Younger advisers discount their impact. *Working Paper.*

Table 1: 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 2: Summary of Datasets in Study 1

Dataset Name	Ground Truth	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 3: 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 Short Plans**	**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 Short Plans	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]**

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

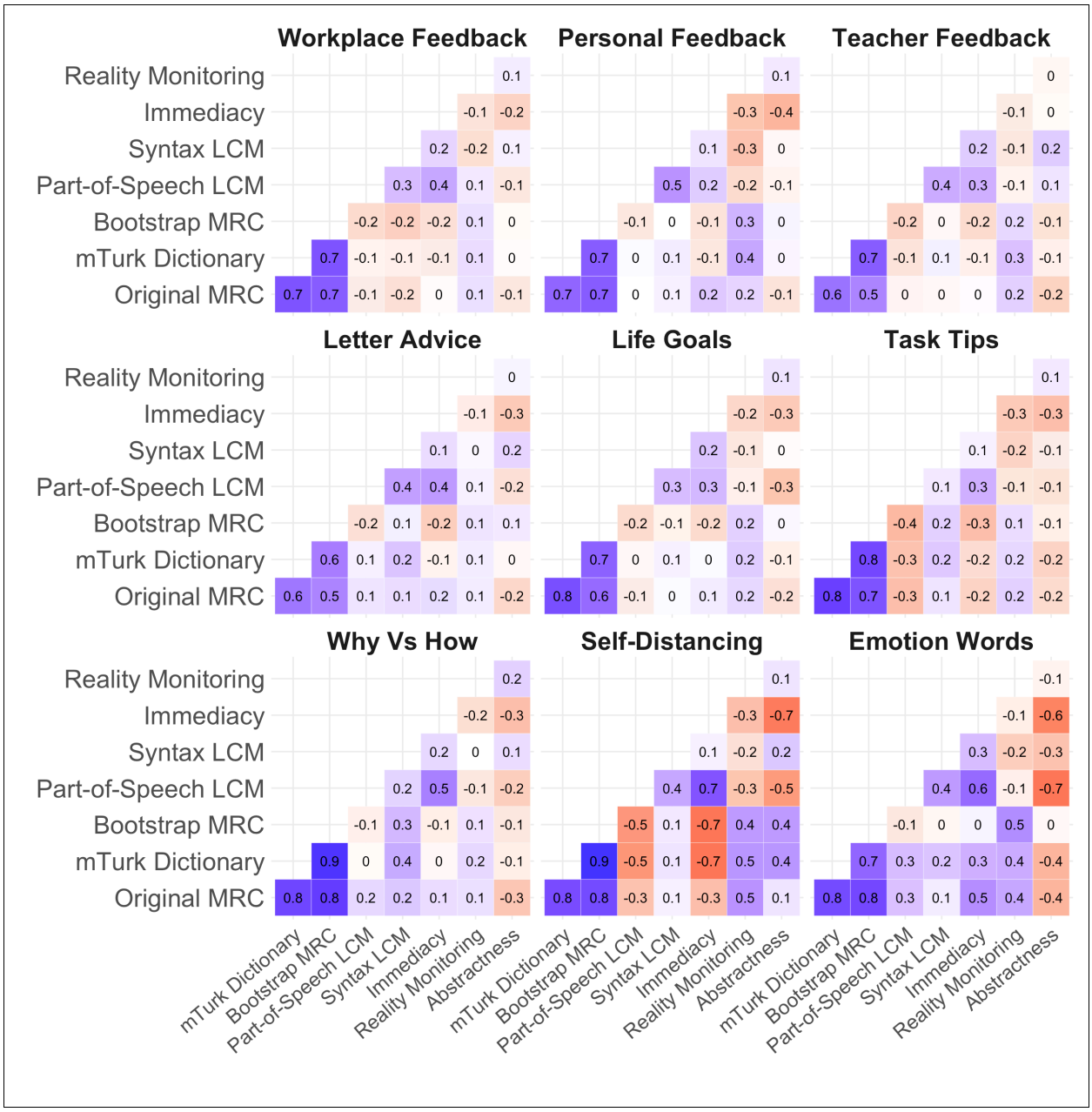


Fig 2. Correlation with concreteness content (and 95% CI) for linguistic measures of concreteness. The Y axis distinguishes different datasets, and each panel shows a different measure.

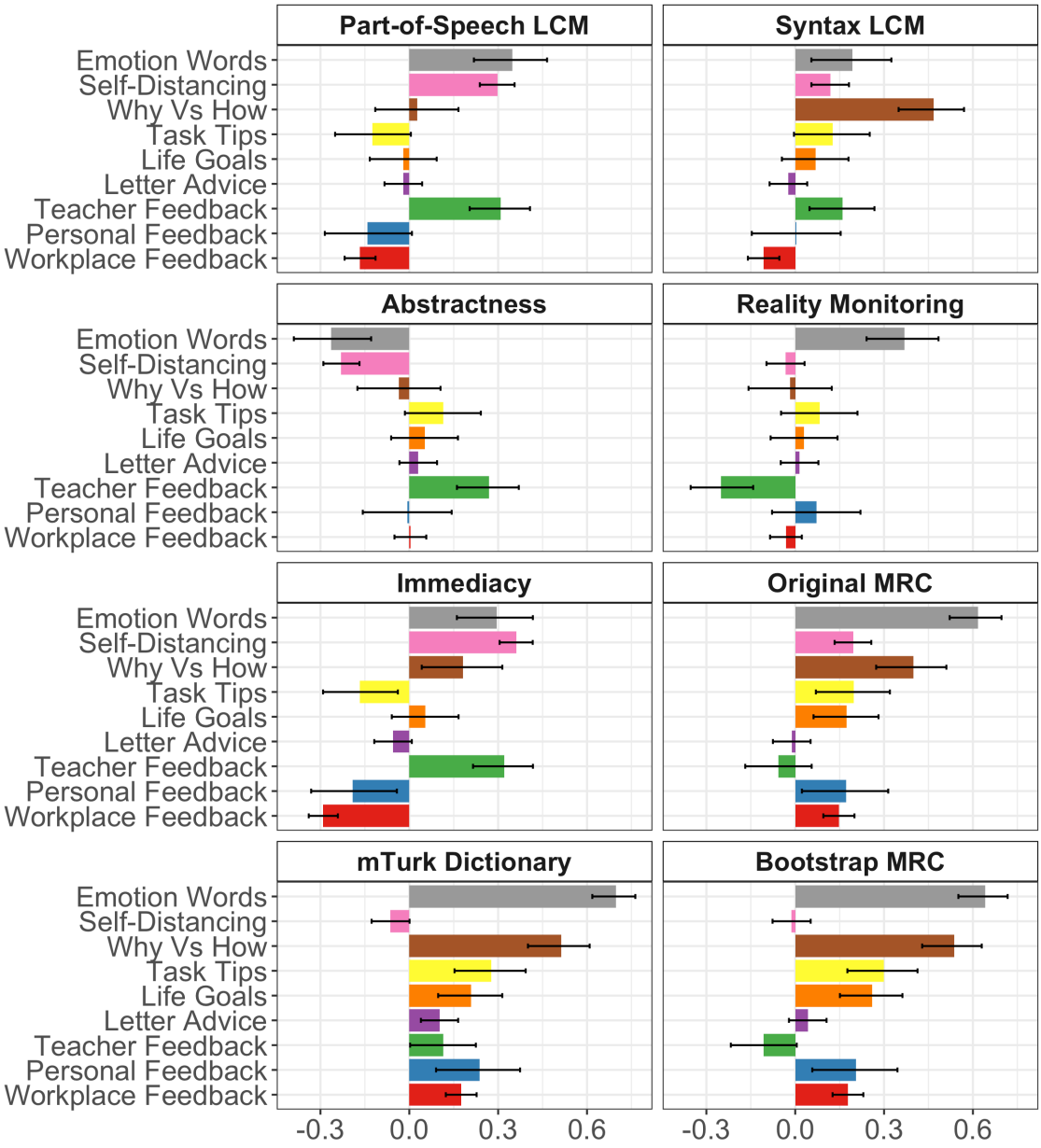


Fig 3. Correlation with concreteness content (and 95% CI) for part of speech categories from the Linguistic Category Model, plotted separately for each domain in Study 1.

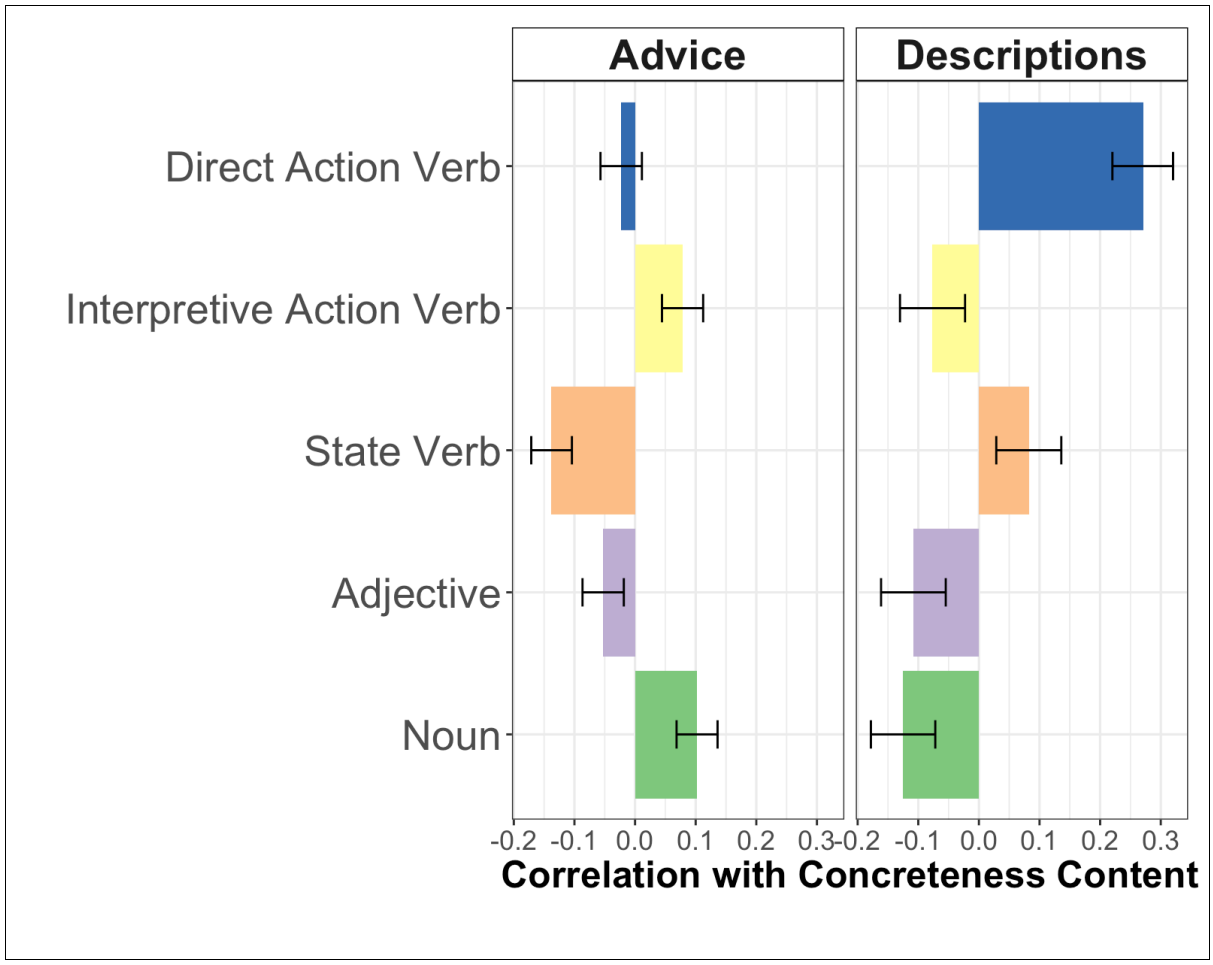


Fig 4. Correlation with concreteness content (and 95% CI) for categories in the LIWC measures, plotted separately for each domain in Study 1`. Red bars are features that are supposed to be negative, and light blue bars are supposed to be negative, according to the original LIWC construct.

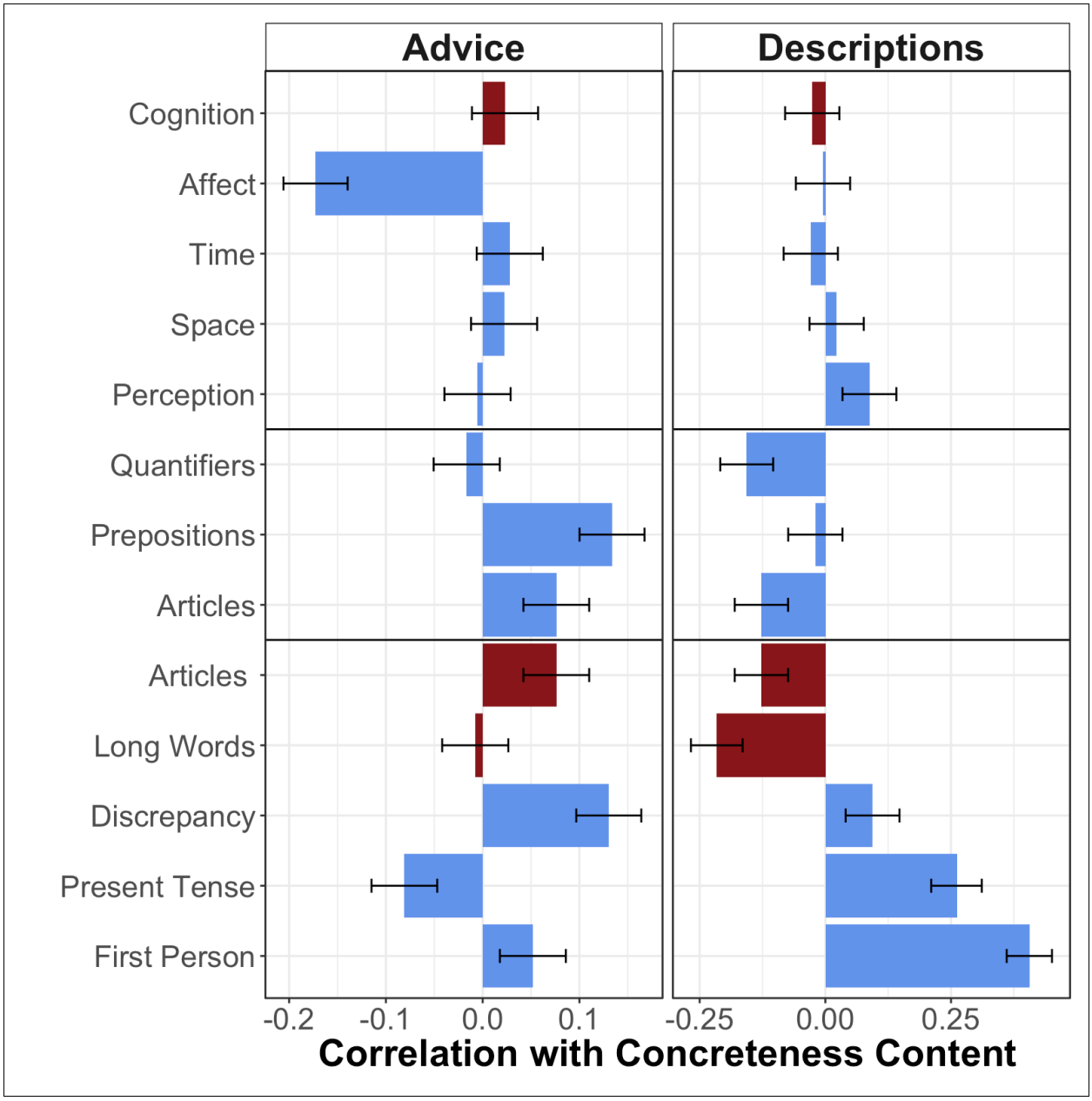


Fig 5. Relationship of concreteness to plan distance (short-term vs. long-term) and plan specificity (annotated) in Study 2. The Y axis distinguishes different linguistic models, and the X axis represents a standardized regression coefficient and 95% confidence interval. Colors identify regression specifications that include different sets of control variables.

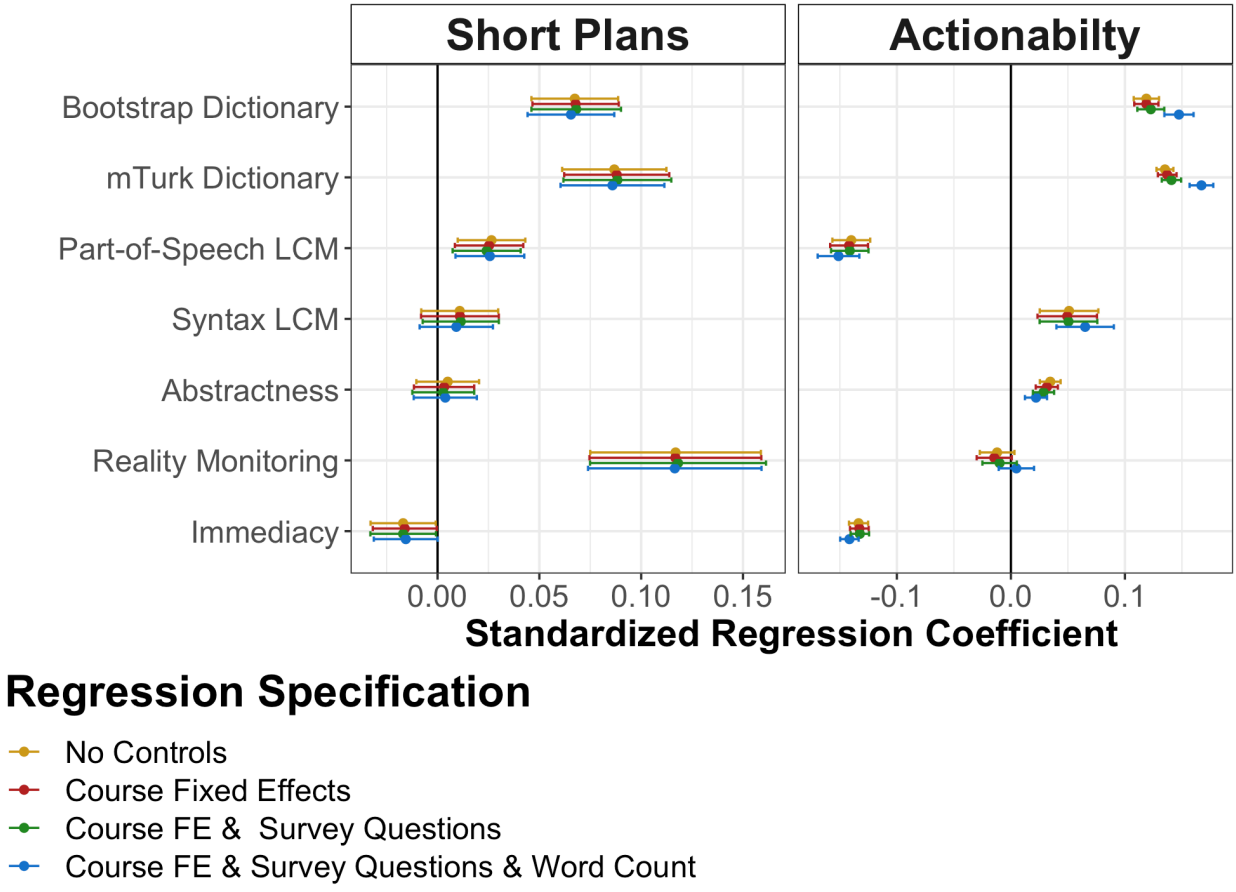
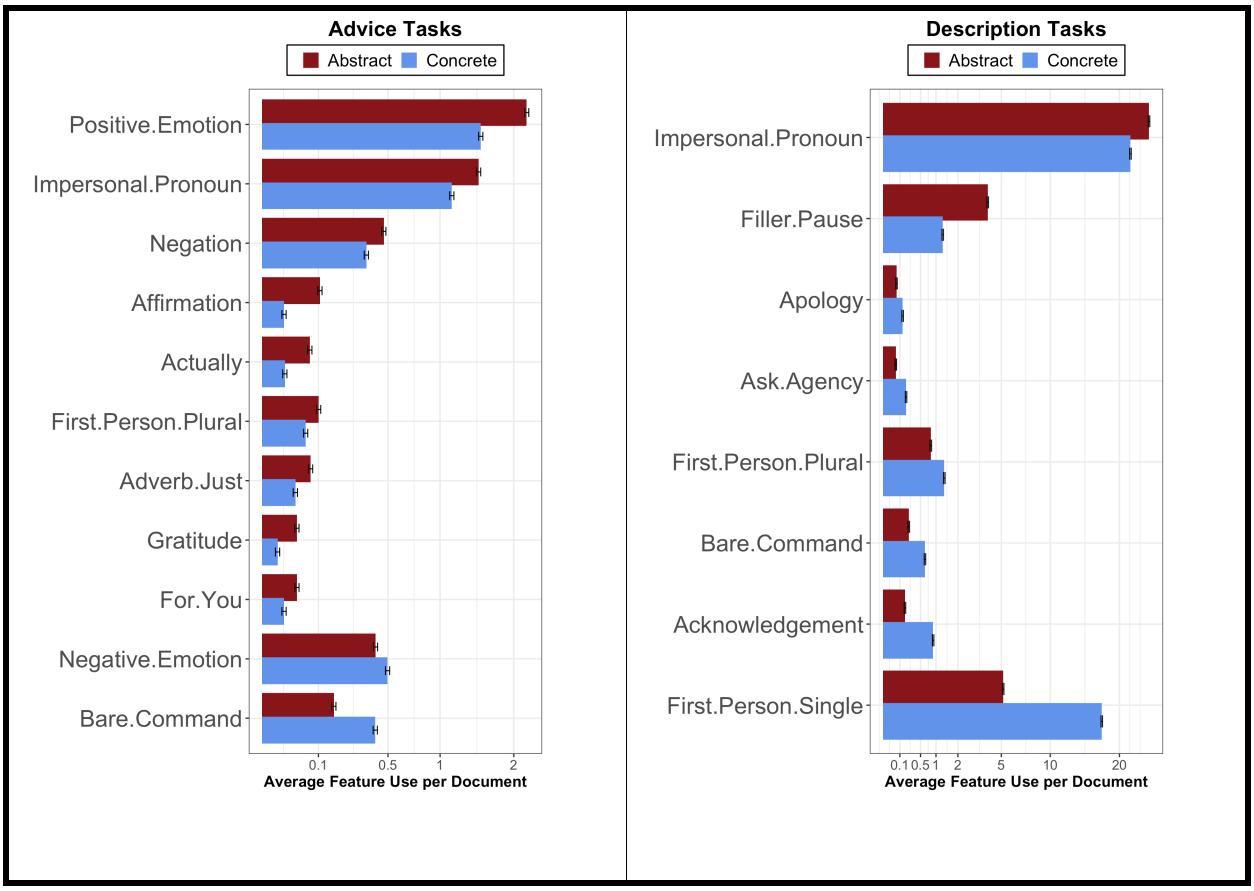
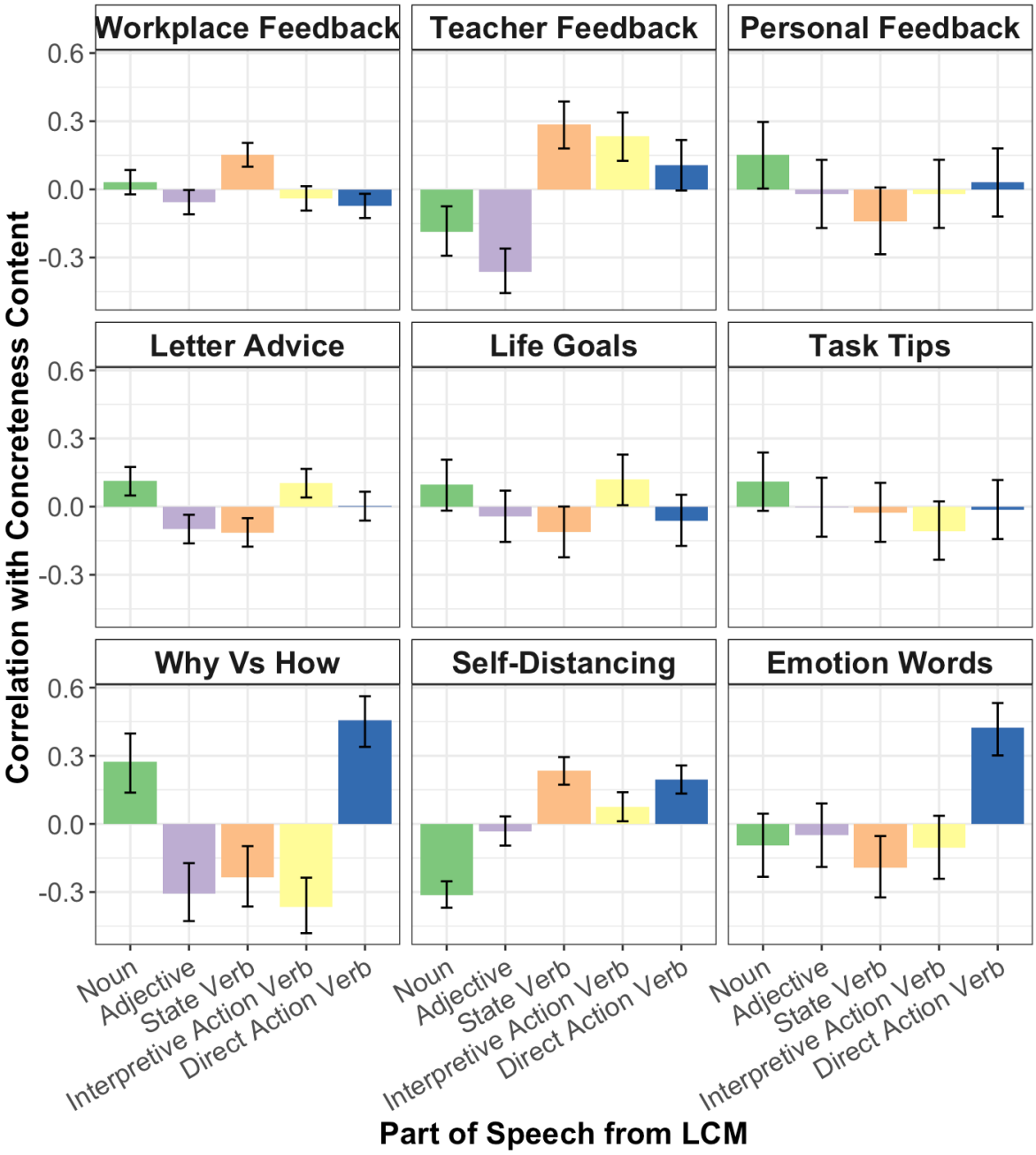


Fig 6. Differences in politeness features between top and bottom terciles of length-adjusted concreteness, plotted separately for each domain in Study 1. Bars show average feature counts (and SEs).



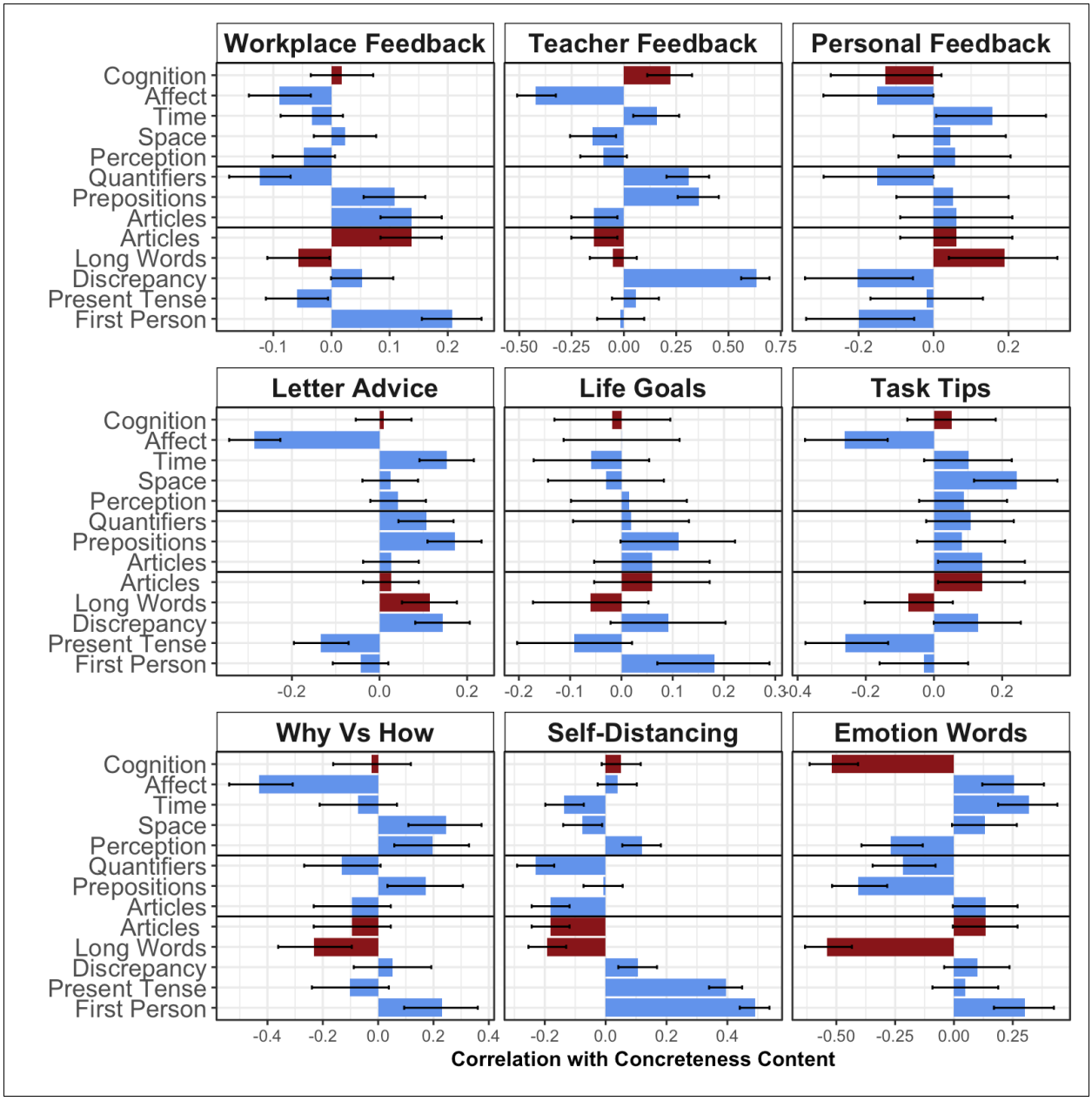
Appendix A: LCM Category Scores for Study 1

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



Appendix B: LIWC Category Scores for Study 1

Figure A2: Correlation with concreteness content (and 95% CI) for LIWC concreteness constructs. The Y axis distinguishes LIWC categories, and each panel is a different dataset.



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 again vary the feature sets across runs (bag of words and syntax LCM; both 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). 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, at least).

Figure A3: 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.

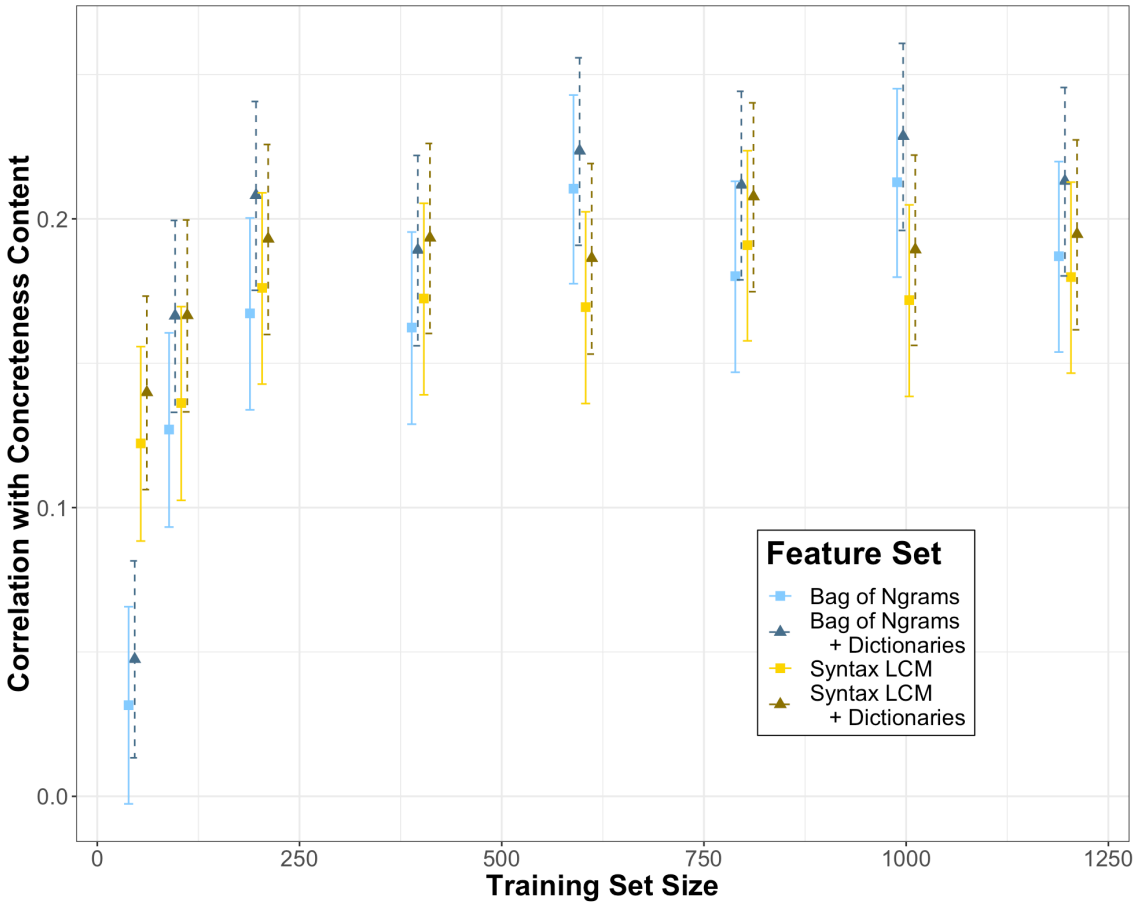


Figure A4: Effect of training set size on accuracy of supervised models. All points represent the correlation with concreteness content (and 95% CI), pooled across all annotated data in Study 2.