Expressly Different:
Discursive Diversity and Team Performance

Katharina Lix[a], Amir Goldberg[a], Sameer B. Srivastava[b], and Melissa A. Valentine[a]
[a]Stanford University; [b]University of California, Berkeley

Abstract

How does diversity among members of a team affect their performance? Prior research has found diversity to be a double-edged sword that sometimes boosts and in other cases dampens performance. Yet empirical evidence on the link between diversity and performance remains mixed, and theoretical progress in understanding the contingencies has begun to stall. To help reinvigorate research in this field, we propose a novel conceptualization of team cognitive diversity and introduce a language-based technique to measure it. We focus on a particular aspect of cognitive diversity—discursive diversity—that reflects *realized*, rather than potential, divergence among group members; *expressed*, rather than latent, differences in the way members construe and ultimately communicate about a given set of topics; and *temporal variation* in construals and expressions over a team's life cycle rather than the assumption of stability. We use the tools of natural language processing to develop a time-varying measure of discursive diversity. Using data from 117 remote teams of freelance software developers who collaborate via an online communication tool, we find that discursive diversity is generally associated with better team performance. However, levels of discursive diversity fluctuate significantly over teams' life cycles. In more fine-grained analyses, we find that discursive diversity's effects on performance are contingent on time: it is positive when a team's next milestone is distant but turns negative as the next milestone approaches. We discuss implications of this work for research on diversity and cultural heterogeneity and the potential for computational methods to inform the design and implementation of diversity and inclusion initiatives in organizations.

March, 2019

INTRODUCTION

Knowledge production frequently occurs in the context of teams that bring together individuals with different backgrounds, expertise, and perspectives. How do various forms of diversity among these individuals relate to their performance as a team? One perspective emphasizes diversity's role in enabling teams to generate novel ideas and recombine existing ideas in new ways by traversing a wider search space and bridging divergent perspectives (Pelled, Eisenhardt, & Xin, 1999; Amabile et al., 1996; Williams & O'Reilly, 1998). An alternative view focuses on the challenges that diversity poses for the coordination of activity and cohesion and alignment among group members (Weber & Camerer, 2003; Knight et al., 1999). Although the team diversity literature is expansive, empirical evidence on its net consequences for group performance remains mixed, leading researchers to search for an ever-growing but still unsettled list of moderating factors (see Joshi & Roh, 2009, for a review).

We believe that progress in understanding the relationship between diversity and team effectiveness has begun to stall out because of three core limitations in extant research. First, whereas the theorized benefits and costs of diversity relate primarily to how group members interpret, engage with, and respond to each other's ideas and beliefs, it is typically not possible to observe these aspects of cognition directly. Instead, prior work has relied on proxies of cognition such as the team's demographic composition, members' personality traits, and their self-reported attitudes, values, and beliefs (Knight et al., 1999; Kilduff, Angelmar, & Mehra, 2000; Cox & Blake, 1991). Yet, in many cases, these ascriptive and attitudinal characteristics of group members may correspond poorly to how they actually think when interacting with one another.

In addition, studies of team diversity have tended to assume that, if various forms of difference exist within a team, these divergent viewpoints will necessarily be expressed. Yet in

many team contexts, certain individuals may feel inhibited from voicing their views or may hold back from doing so at critical junctures in a team's life cycle (Detert & Edmondson, 2011; Murnighan & Conlon, 1991). Finally, prior work has typically measured diversity based on surveys, which are implemented once or at best episodically. In part for this reason, team diversity has generally been conceptualized as a static construct. Yet diversity that manifests in group interactions is likely to vary from day to day and across different phases of the team lifecycle (McGrath, 1991). Without being able to observe these fine-grained temporal fluctuations, it is impossible to know whether they might have meaningful performance implications.

In this article, we seek to address these gaps and help reinvigorate diversity research by focusing attention on a specific form of diversity: that manifested in group members' interactional language use. The language a person uses when communicating with peers reflects her underlying categories of thought. Language also represents a set of shared conventions that connect symbols to meanings, thereby enabling coordination among individuals. Guided by these insights from sociolinguistics, as well as recent work that uses language as a window into cultural dynamics within (Goldberg et al., 2016; Doyle et al., 2017; Srivastava et al., 2018) and across (Corritore et al., forthcoming) organizations, we introduce a novel construct to diversity research: *discursive diversity*, or the differences in meaning conveyed in group member interactions at a given point in time.

We note, however, that group members can diverge in meaning in at least two analytically distinct ways—when individuals discuss distinct topics and when they discuss the same topics in different ways. For example, a conversation in a product development team may involve a marketing person talking about pricing and a manufacturing person talking about costs. We refer to this difference as *topical diversity*. By discursive diversity, we instead focus on the second form

of divergence. In the product development example, this might entail a discussion of the same topic—say pricing—that the marketing person discusses in reference to the external marketplace and the manufacturing person talks about relative to internal gross margins. Importantly, discursive diversity can wax and wane over time as a function of the topics that are being discussed, the individuals who choose to voice their views, and how they choose to express themselves.

In the context of teams that are focused on knowledge production rather than mere execution of routine tasks, we theorize that—all else equal—discursive diversity will enable broader knowledge search and more recombination and thereby enhance team performance, even when accounting for the degree of topical diversity in group interactions. Yet we also propose that the effects of discursive diversity on team performance will be contingent upon time: it will be beneficial when milestone deadlines are distant and new ideas can still be implemented, but it will instead be detrimental when milestones are imminent such that new ideas serve as a distraction from executing the best prevailing idea.

We test these ideas using a unique and rich data set that includes 117 software development teams whose 421 members were remotely located but assembled together via an online platform and who communicated with each other exclusively through the Slack collaboration tool. We use the content of over 800,000 Slack messages sent by these developers to map words into a multidimensional space of meaning and then derive time-varying measures of discursive diversity. We also use standard text analysis techniques to measure topical diversity and other linguistic features that serve as control variables in our models. Moreover, we use data from the online platform to derive measures of the most consequential performance outcome for these teams: the proportion of milestones they achieve on time. Using coarsened exact matching to identify observationally similar teams that vary on discursive diversity, we find strong empirical support

for our two main hypotheses. We discuss implications of this novel approach to measuring a core dimension of cognitive diversity for research and practice.

THEORY

Demographic Diversity and Team Performance

Research on the consequences of group diversity for performance, which is too vast to comprehensively review here, spans social psychology (e.g., Dahlin, Weingart, & Hinds, 2005), sociology (e.g., McPherson & Rotolo, 1996), and organizational behavior (e.g., Williams & O'Reilly, 1998). Although the various literatures on team diversity use different terms and emphasize distinct constructs, they are broadly united in conceiving of it in two main forms: demographic or ascriptive diversity, which focuses on variation in easily observed characteristics such as gender, age, race, and occupational background; and deep-level or cognitive diversity, which emphasizes differences in underlying ways of thinking, perceptions, and beliefs (Harrison et al., 1998).

While a majority of diversity research has focused on demographic diversity given that it is easier to measure than cognitive diversity (e.g., Lawrence, 1997), meta-analyses suggest mixed support for the proposition that variation in a group's demographic composition enhances team performance (e.g., Joshi & Roh, 2009; Horwitz & Horwitz, 2007). Whereas some researchers have reported positive effects of demographic diversity on creativity (Amabile et al., 1996) and problem-solving effectiveness (Hambrick, Cho, & Chen., 1996), others have found that it can lead to increased group conflict (Jehn, Northcraft, & Neale, 1999) and hinder communication flows (Zenger & Lawrence, 1989). In many cases, the documented effects of diversity on performance do not hold when potentially confounding contextual factors such as task and industry

characteristics are taken into account (Joshi & Roh, 2009). Recent years have seen a growing but still inconclusive search for moderators of the relationship between demographic diversity and team performance (Horwitz & Horwitz, 2007).

Implicit in the demographic approach to diversity research is that membership in certain categories relates to team members' cognition in ways that drive team performance outcomes. For example, some researchers have argued that a shared organizational background contributes to the development of common schemata, which are thought to shape interaction patterns in teams (Michel & Hambrick, 1992). Yet, for the most part, the assumption that demographic heterogeneity begets cognitive diversity has been largely unexamined in this stream of work, and it remains unclear how specific forms of demographic diversity might manifest in individual and group cognition.

Cognitive Diversity and Team Performance

A second strand of diversity research has focused directly on the relationship between cognitive diversity and team performance (e.g., Phillips & Loyd, 2006). For example, Harrison et al. (1998) showed that deep-level characteristics can be, but often are not, associated with differences in easily observable surface-level characteristics such as gender and professional background. A variety of studies have argued that divergent ways of thinking, attitudes, and beliefs will lead to positive team outcomes (Pelled, Eisenhardt, & Xin, 1999; Kilduff, Angelmar, & Mehra, 2000; Cox & Blake, 1991). The empirical evidence in support of this contention remains, however, mixed.

On one hand, cognitive diversity has been found to improve team problem-solving effectiveness (Kilduff, Angelmar, & Mehra, 2000; Cox & Blake, 1991), creativity and innovation (Amabile et al., 1996), as well as group learning (Fiol, 1994). The mechanism thought to underpin

this relationship centers on the search space for potential solutions: the more team members' perspectives are variegated, the broader is the solution space they are assumed to search (Pelled, Eisenhardt, & Xin, 1999; Amabile et al., 1996). Teams that traverse a broader search space have a greater likelihood of discovering novel ideas and of surfacing new ways to recombine known ideas (Fleming, 2001; Uzzi et al., 2013; de Vaan et al., 2015). For example, in a study of top management teams, Kilduff, Angelmar and Mehra (2000) found that teams whose members held more diverse beliefs about cause-effect relationships, team processes, and external events tended to outperform those with less diverse perspectives. They did not, however, find a significant relationship between team demographic and cognitive diversity.

At the same time, cognitive diversity can impede teams' ability to coordinate effectively. If team members differ in their perceptions of what must be done and when, or if they hold incompatible interpretations of key internal and external events, chaos can ensue. When team members' normative expectations and beliefs are incompatible, their ability to coordinate and accomplish tasks tends to suffer (Weber & Camerer, 2003; Knight et al., 1999; Dahlin, Weingart, & Hinds, 2005).

As with demographic diversity, mixed evidence on the relationship between cognitive diversity and team performance has spawned a large number of studies on the role of potential moderators. Psychological safety (Martins et al., 2013), relationship conflict (Jehn, Northcraft & Neale, 1999), trust (Olson, Parayitam, & Bao, 2007) and transformational leadership (Kearney & Gebert, 2009) are just a few examples of moderators that cognitive diversity researchers have studied.

Taken together, these literatures suggest that demographic diversity is often used as a noisy proxy for underlying differences in team members' cognition. Cognitive diversity, in turn, seems

to have important but complex relationships with team performance. On one hand, diversity in team members' underlying perceptions and beliefs can boost performance by increasing the likelihood of finding creative solutions. At the same time, it can also harm performance by impeding coordination. Indeed, a growing ensemble of diversity scholars has noted that it represents a double-edged sword for team performance.

To move beyond this self-evident conclusion that ultimately leaves more questions open than it answers, we propose that it is necessary to confront and address three core limitations of past research. First, rather than assuming that demographic differences correspond to cognitive differences or that self-reports of values, beliefs, and ways of thinking are necessarily reflective of how group members think and behave in practice, it is important to theorize about and measure the actual cognitive differences that arise in group dynamics. In other words, we propose to shift the analytical focus from *potential* cognitive diversity to its *realized* manifestations.

Prior work has tended to assume that cognitive differences that exist in group members' minds will necessarily be conveyed when they interact with one another. However, in many situations—for example, when a domineering new leader takes over a team—certain individuals may feel inhibited from voicing their views (Detert & Edmondson, 2011). Similarly, at certain times—such as before an impending deadline—group members may hold back dissenting opinions and focus instead on execution (Murnighan & Conlon, 1991). Thus, we suggest the need to shift from studying *latent* forms of cognitive diversity to those that are *expressed* by group members.

Finally, we believe there is a need to complement existing survey-based measures of team diversity, which provide mostly *static* (or at best episodic) snapshots of diversity over a team's life cycle (see, for example, Kilduff, Angelmar, & Mehra, 2000) and can be susceptible to various forms of self-report bias (e.g., Donaldson & Grant-Vallone, 2002; Greenwald & Banaji, 2005),

with more granular, *dynamic* measures. In particular, we propose that the language team members use when communicating with each other can reveal meaningful dimensions of expressed cognitive diversity and how it varies over time. Research in sociolinguistics shows the many ways that language connects symbols to meanings, which in turn enables interpersonal coordination (Lewis, 1969). Indeed, language operates as a high-level control system for the mind that enables people to construct mental representations of themselves and others (Lupyan, 2016). In other words, language provides a portal into individual and group cognition.

Building on these insights, Goldberg, Srivastava, and their colleagues (Goldberg et al., 2016; Srivastava et al., 2018) develop an interactional language use model of cultural alignment based on the linguistic styles people use when communicating to their colleagues via email and demonstrate that this language-based measure of cultural fit is predictive of consequential career outcomes such as promotion, involuntary exit, and favorable performance ratings. In a similar vein, the language employees use when describing their organizations on platforms such as Glassdoor.com can be used to derive time-varying measures of cultural heterogeneity (Corritore et al., forthcoming). Whereas these prior studies have focused on language as a window into cultural alignment at the individual level and heterogeneity at the organizational level, we instead propose to use language as means to assessing cognitive diversity at the intermediate level of groups and teams.

Discursive Diversity and Team Performance

Cognitive diversity can be thought of as an umbrella construct that encompasses many facets of how people in a group think, process information, interpret situations, and express their views. We focus on one core dimension: variation in the lenses through which individuals privately construe and ultimately communicate their understanding of topics that are being discussed by

their group at a given point in time. We use the term *discursive diversity* to describe this aspect of group-level cognition. Although the construct encompasses both construals and expressions, the former are typically unobserved by group members. We therefore focus on discursive diversity as reflected in group members' expressions. We also note two features of the construct (and our corresponding measure) that distinguish it from past diversity research: it focuses on behaviors rather than attitudes or beliefs, and it embraces the possibility of temporal variation.

Our first set of arguments focus on the aggregate level of discursive diversity present in a team and performance. Prior work has already established the tradeoffs of cognitive diversity: it can impair coordination, sow misunderstanding and mistrust, and slow execution (Weber and Camerer, 2003; Knight et al., 1999; Dahlin, Weingart, & Hinds, 2005), but it can also improve team problem-solving effectiveness, creativity, and group learning (Kilduff, Angelmar, & Mehra, 2000; Amabile et al., 1996; Page, 2007).

In assessing how these general tradeoffs of cognitive diversity manifest in the specific form of discursive diversity, we invoke toolkit theory from cultural sociology (Swidler, 1986). This perspective views culture as a loosely-held repertoire of symbols, worldviews, and styles that people use to construct strategies of action. The choices people make about which of these cultural "tools" to deploy vary situationally (Swidler, 2001; Hallett & Ventresca, 2006; Harding, 2007).

In organizational settings, the breadth of the toolkit available to people can have important performance implications. For example, Corritore, Goldberg and Srivastava (forthcoming) report that firms with high intrapersonal cultural heterogeneity—that is, those whose members had, on average, access to a wide range of cultural resources—produced more patents and higher quality patents than did their counterparts that were low on this dimension. Moreover, the former set of firms were more highly valued by the market than the latter.

We extend these arguments from the level of the organization as a whole to the team unit of analysis. In contexts where the primary objective is knowledge production, rather than routine task execution, we similarly propose that teams higher in discursive diversity can be thought of as invoking a broader cultural toolkit. As a result, such teams will tend to view the same functional topics through multiple, non-overlapping lenses and will be able to traverse a broader knowledge space. Doing so will enable them to better surface novel ideas and to recombine known ideas. Thus, we expect:

Hypothesis 1: In the context of knowledge production, teams exhibiting higher mean levels of discursive diversity will outperform teams that are less discursively diverse.

Intertemporal Tradeoffs of Discursive Diversity

Whereas our first hypothesis sidesteps the question of temporal variation in discursive diversity, our next one directly engages it. We begin with the premise that the activities, attention, and assumptions of a group vary across its life cycle. Whereas early research on teams assumed that they progressed gradually through discrete and predictable stages, Gersick's (1988, 1989, 1991) theory of punctuated equilibrium instead suggests that teams undergo fundamental cognitive shifts part of the way through their life cycle and often near the midpoint. For example, team members pay greater attention to time, try to conclude their prior deliberations, come to some agreement on their course of action, and pay greater attention to external stakeholders and their demands (Gersick, 1989).

Building on these insights, Ford and Sullivan (2004) develop a theoretical account of how teams are likely to process and value novel contributions across their life cycle. They propose that

novel contributions are likely to benefit teams at early stages, when people seek to learn about the problems they are tasked with solving, search for insights, and explore potential solutions. In contrast, novelty is likely to harm team performance in later stages when team members' attention shifts to meeting an impending deadline and responding to the demands of external stakeholders. In these later stages, novelty can distract and frustrate group members to the point that they become less effective.

Our argument follows directly from these perspectives, but we depart in one key way. Whereas these prior studies emphasized the midpoint in a team's life cycle as the key cognitive turning point, we acknowledge that knowledge-producing teams often remain intact for extended periods of time and typically produce multiple work products that are due at varying points in time. With the arrival of every new milestone, they repeat the cycle of exploration followed by execution. When milestones are far away in time, the team is exploring, and novelty is valued, we propose that discursive diversity will support team success by supporting the generation of novelty. Yet when milestones are imminent, the team is executing, and novelty is therefore a distraction, the positive effects of discursive diversity on novelty generation will instead become an impediment to team success. Thus, we anticipate:

Hypothesis 2: In the context of knowledge production, the effects of discursive diversity on team performance will be contingent on time: it will be beneficial when milestones are temporally distant and harmful when they are proximate.

METHOD

Research Setting and Data

Our research setting is an online platform that brings together individual software developers to serve on temporary teams that collaborate to produce on-demand software for both individual and business clients. The freelance software developers on this platform are distributed around the globe and work on a variety of projects ranging from mobile to web application development. The projects are uniformly knowledge-intensive and require high levels of creativity, technical problem-solving, and effective coordination. The platform acts as an intermediary between clients and freelance developers by taking in clients' project requests, refining technical specifications with clients, and composing suitable teams to work on the projects. The platform negotiates prices with clients and administers payments to freelancers.

Our dataset comprises 117 teams, representing 421 unique individuals (36% female). A typical team has 8 members and consists of one project manager, at least one backend or "fullstack" engineer, at least one front-end engineer, and a user interface expert. Depending on the type of project, teams may also contain writers, natural language processing engineers, and other types of specialized professionals. Among teams in our data, projects lasted 159 days on average (median: 150 days) and were structured in milestone phases that last between one and four weeks (mean: 25 days; median: 16 days). To join the platform, professionals have to pass a variety of technical interviews designed to verify their expertise in their field. On average, the members of an individual team represent 4.6 countries (median: 4). 42% of individuals in our sample listed their country of origin as located in North America. Another 13% hailed from Asia, followed by 12% from Europe. The remaining 23% resided in Latin America, Africa, and other parts of the world.

Because they were geographically distributed and lacked any physical office space, team members collaborated exclusively via an online communication tool called Slack and, for posting and editing code, the software development platform GitHub. The platform strongly encouraged adherence to these modes of communication. On average, teams exchanged 1873 Slack messages in public channels throughout their lifespan (median: 1220).

We collected detailed data on teams' demographic characteristics, project complexity, intra-team communications and performance. We measured teams' diversity with respect to their functional roles[1], gender, and country of origin. In addition, we collected the entire set of teams' Slack archives, resulting in 800,000 messages total.[2] Each message was timestamped and attributable to its author. We also collected extensive performance data, including the timeliness of delivery of individual milestones. Together, these data sources resulted in a rich, detailed and continuous history of teams' internal dynamics, processes, and outcomes.

Dependent Variable

The timely delivery of milestones is the most critical performance measure for teams. Company executives explained that clients prioritize timeliness, and this is the key metric used to evaluate individual freelancers. The final project deliverable, as well as the deliverables for each milestone and dates for corresponding deadlines, are agreed upon between the project manager and the client before the project starts. Timely delivery signals both effective coordination among team members, as well as high output quality, since a client must approve or reject the team's deliverables at each milestone deadline. If the agreed-upon deliverables for a given milestone are

[1] Typical roles included project manager, fullstack engineer, backend engineer, frontend engineer, designer, and so on.

[2] The data were anonymized to remove any personally identifiable information. Our dataset contains only messages that were sent to public channels. That is, we did not have access to private direct messages exchanged among team members. However, platform executives confirmed that most planning and discussion happened in public channels for easier traceability.

deemed by the client to be of poor quality or incomplete, that milestone is marked within the company's system as delayed. Teams are allowed to proceed from one milestone phase to the next only after the client approves a given milestone's deliverables as satisfactory. Team members are paid a pre-agreed sum upon the approval of each milestone, as well as upon successful completion of the project. Members of teams who do not deliver on time may experience financial penalties or limited opportunities to join lucrative projects in the future.

Independent Variables

We measure *discursive diversity* as the differences in meaning team members convey in interaction with each other over the course of a project. By differences in meaning, we refer to the differences in mental associations, construals and perceptions that team members convey around the same substantive issue in conversation with one another. Since language offers a window into people's attitudes, perceptions and beliefs (Srivastava, Goldberg, Manian, & Potts, 2017; Srivastava & Goldberg, 2017), task-relevant conversations among team members are a suitable basis for measuring team discursive diversity. The intuition behind our measure is that teams whose members talk about a given set of topics and convey similar meanings around those topics are less interpretively diverse than teams whose members discuss those same topics using language that reflects the different meanings they assign to and lenses through which they view the topics.

To develop this measure, we fit a statistical model to the entire corpus of teams' Slack archives (vocabulary size 10.5k), which is designed to capture the latent semantic patterns embedded in team communication patterns. This type of statistical model, known as a "word embedding," captures semantic relationships between words, such as gender, verb tense, and sentiment, and superordinate-to-subordinate category relationships (Mikolov, Chen, Corrado, & Dean, 2013). Embedding models locate words in a real-valued, multidimensional vector space

such that the coordinates of words with similar meanings are located close to one another in the space (Mikolov et al., 2013). The resulting dimensions of this vector space can be understood as the common latent features underlying language use in a given text corpus (here, the entire set of Slack archives). We pre-processed the Slack data according to standard procedures in natural language processing and trained our embeddings model using a Python implementation of Word2Vec, a popular tool for training word embeddings models.[3]

Using the resulting embeddings model, any two chat messages can be evaluated for their relative dissimilarity by computing the distance (1-cosine similarity) between their respective mean embedding vectors. The greater the distance between two messages' mean embeddings vectors, the greater is their discrepancy in content, style, and subjective meaning. And these differences are, in turn, reflective of heterogeneity in the underlying attitudes and beliefs of message senders. The relative dissimilarity between two speakers, A and B, in a given time window *w* can be calculated as the distance (1-cosine similarity) between the mean vector representing A's messages sent during *w* and the mean vector representing B's messages sent during *w*. We computed weekly and daily *discursive diversity scores* by computing the mean cosine distance between all dyads within the team in a given week and day, respectively.[4]

[3] We used Python gensim's implementation of Word2Vec to train embeddings vectors (window size 10). Words were tokenized and lemmatized. Stop words and tokens that occurred fewer than 30 times in the corpus were excluded. Tokens representing dates, elongations, URLs and other rare tokens were replaced with category-representing tokens. Some of these tokenizing functions were inspired by the Stanford NLP group's code for parsing Twitter data for training word embeddings (Pennington et al., 2014).

[4] Our dataset includes teams' public Slack channels but not the direct messages exchanged among team members. Conversations in public channels are, in principle, open to all team members. Individual messages are not usually explicitly addressed to a specific other but are part of a collective discussion. We treat conversations as team-level interactions that occur between all team members. Therefore, our measure is based on cosine distances for each dyad of active speakers on the team in a given time window.

Control Variables

Team size and project complexity. Because teams' communication patterns are likely to vary as a function of their size and the complexity of the projects they work on, we match teams on these variables using coarsened exact matching (Iacus et al., 2012). We describe this approach in detail in the Analytical Strategy section below. Project complexity was measured based on project revenue ("price"). Higher-priced projects tended to be more complex technologically and in their need for coordination across team members.

Demographic diversity. We measure three types of team demographic diversity: Functional role[5], gender, and country of residence diversity. Each was quantified using a standardized Blau index, where values closer to 1 indicate greater heterogeneity and values closer to 0 indicate greater homogeneity with respect to the focal characteristic. Interestingly, out of all three demographic diversity measures, only role diversity was significantly correlated with our measure of discursive diversity ($r=0.36$, $p<0.05$). Table 2 offers univariate summary statistics and bivariate correlations.

Engagement levels on communication platform. We controlled for team members' engagement levels on Slack during a given time window by introducing the average percentage of team members who were active on Slack (i.e., who sent at least one message) during that time window, as well as the average word length per message during that time window, as control variables. For the first set of analyses (H1), the time window was set to one week. For the second set of analyses (H2), it was set to one day.

Topical diversity. Word embeddings models capture the latent semantic features underlying language use. These latent features can include the topical content of language (for

[5] Team members could each hold one of 21 possible roles. However, most teams had highly similar role structures, with certain core functions being represented on every team.

example, when conversations revolve around backend-coding vs. design choices) but also stylistic aspects of language such as sentiment or the degree of subjectivity of language. Since the goal of our research is to isolate differences in meaning that speakers convey around a similar set of topics, we must disentangle discursive divergence from differences in the topics teams are discussing. To achieve this, we introduce a control measure for topical diversity.

Following a machine learning approach proposed by Blei, Ng and Jordan (2003), we trained a Latent Dirichlet Allocation (LDA) topic model on the entire set of teams' Slack archives to identify the key topics that team members discuss. LDA topic models model documents as distributions over topics, where the topics themselves are distributions over words (Blei et al., 2003). LDA "learns" the latent topics in a corpus (in this case, the entire set of Slack archives) based on the word co-occurrence patterns within documents. Treating the collection of Slack messages that an individual team member sent on a project as one document, we trained an LDA model to identify the latent topics that speakers discussed.[6] A model with 12 topics returned what appeared to us as the most coherent and cohesive set of topics. Examples of the topics we labeled through this exercise include project management, backend engineering issues, application design choices, payments, and leisure activities.

We quantified the set of topics a team member discussed during a given week as the probability distribution of her aggregated messages sent in that week over the 12 topics identified by the LDA model. The topical distance between speakers A and B during time window *w* can be

[6] In general, LDA topic models do not work well on very short texts because a certain amount of data is needed to meaningfully discern differences in word co-occurrence patterns between documents (Yan, Guo, Lan, & Cheng, 2013). Therefore, we included only speakers who uttered at least 500 words on a given project in our training data for the LDA topic model. Nearly all team members (98%) uttered at least 500 words on a given project and were included in the topic model training set.

calculated as the Hellinger distance[7] between their respective messages' probability distributions over the latent topics for time window *w*. Weekly team topical diversity was measured as the average dyadic Hellinger distance between all of a team's dyads' topic distributions in week *w*.

Variance in speaking styles. As mentioned earlier, word embeddings capture numerous latent categories of meaning conveyed in language, including stylistic elements, such as sentiment, subjectivity, and concreteness.[8] Variance in individuals' stylistic choices in language can also convey information about diversity in their interpretations and construals of the same topics. For example, if team member A's messages about a customer tend to be positive in sentiment, while team member B's tend to be negative, they are likely to hold different interpretations and construals of the customer. Thus, we consider variance in certain stylistic choices as parts of discursive diversity that can be measured through well-established, lexicon-based approaches. We sought to control for the proportion of variance in team discursive diversity that is accounted for by variation in common and easily measurable stylistic features of language: sentiment, subjectivity, and concreteness. We computed teams' weekly variance in sentiment, subjectivity and concreteness by computing each team member's weekly mean score with respect to each of these features, and then computing the variance between members' mean scores for each week.[9]

[7] The Hellinger distance is a popular measure for quantifying the distance between two probability distributions and is frequently used in the context of LDA topic models (e.g., Blei & Lafferty, 2007).

[8] For example, a simple OLS regression of 1000 randomly selected words' sentiment scores onto their 200 latent dimensions from the word embeddings model showed that 103 of those dimensions had a statistically significant ($p<0.05$) relationship with sentiment.

[9] Sentiment and subjectivity scores per message were computed using built-in dictionaries from Python's TextBlob package by computing the mean sentiment and subjectivity scores of a message's component words. Message concreteness was evaluated using the crowd-sourced dictionary of English lemmas provided by Bysbaert, Warriner and Kuperman (2014) by computing the mean concreteness scores of its component words. In their prompt to crowdworkers who provided the concreteness ratings, the authors described concreteness as follows: "A concrete word comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do." See original paper for additional details.

Analytical Strategy

We apply coarsened exact matching, or CEM (Iacus et al., 2012), to identify pairs of teams that vary with respect to discursive diversity and performance, but that are otherwise comparable on all observable characteristics. Specifically, we matched teams on project revenue ("price") and team size since those variables are likely to influence team communication dynamics in fundamental ways. For example, larger teams' conversations may be more likely to unfold among subgroups, and more complex projects may involve longer and technology-focused back-and-forths between technical experts. Matching teams on these characteristics helped in isolating the effect of discursive diversity on performance.

To test hypothesis 1, which focuses on the relationship between mean discursive diversity and performance, we ran fractional logit models (Papke & Woolridge, 2008) to predict the fraction of milestones teams delivered on time with strata fixed effects and weights, where the strata and corresponding weights were given by CEM. Controls for weekly topical diversity and various stylistic diversity dimensions (sentiment, subjectivity, and concreteness) were added to the model in stepwise fashion.

Hypothesis 2 proposes that the relationship between team discursive diversity and the probability of meeting an upcoming milestone deadline varies as a function of the time remaining until the deadline. To test hypothesis 2, we examined 101 14-day milestone phases that occurred across 40 teams. We chose 14 days because it was the modal length of milestone phases in our data. For each 14-day milestone phase, we computed daily lagged measures for discursive diversity, the three stylistic variance measures, and engagement levels on Slack, starting with three days before the deadline and going back to the first day of the milestone phase. This resulted in 12 panel data structures, one for each day between the third-to-last to the first day of the milestone

phase. For example, the first panel contained measures for the third-to-last day before the deadline, the second-to-last day before the deadline, and the last day before the deadline. Similarly, the second panel contained all measures taken on the fourth-to-last day before the deadline, as well as the third-to-last, second-to-last, and last day before the deadline. We constructed 12 of these data structures in total, each representing a different number of lag days (from 3 days' lag to 14 days' lag) from the deadline. As a robustness check, we repeated these analyses for milestone phases of different lengths and found the same general patterns as the ones described below.

To test hypothesis 2, we ran logistic regressions predicting whether an impending milestone was met on time. Because we were interested in the coefficient (in the case of logistic regression, coefficients represent log odds) of discursive diversity as a function of the number of days remaining until the deadline, we ran 12 logistic regressions with lag-period fixed effects, each one representing a different number of days remaining until the deadline and drawing on a different one of the lagged data structures described above. As before, all models included controls for team demographic diversity, daily engagement levels on Slack, and variance in daily stylistic features. The latter were also lagged to capture their time-dynamic effects.[10]

Because analyses for hypothesis 2 included only 101 observations across 40 teams, we blocked teams on team size and price (similar to the CEM approach pursued in the analyses for hypothesis 1). Blocking involves binning teams into groups ("strata") that are similar on observational characteristics, but without specifying discursive diversity as a treatment variable. This allowed us to control for the influence of "binned" team size and price on performance while

[10] Note that because the milestone-level models are based on daily data, they do not include controls for topical diversity given that most teams did not exchange enough messages on a given day to compute meaningful topical diversity scores.

preserving observations. We included strata fixed effects and clustered standard errors at the stratum level.

RESULTS

Discursive Diversity: Validating the Word Embeddings Model

There are two common approaches to evaluating the validity of word embeddings models: most-similar queries and word analogy tasks. In most-similar queries, the model is asked to return the words that it learned to be most similar to the vector of a given target word. For example, in software development, the word "bug" usually refers to a programming issue, whereas it is more likely to refer to an insect in other contexts. Our model evaluated the most similar words to "bug" to be "issue", "crash," and "problem," demonstrating that the meaning of "bug" was accurately captured in context-relevant way. Similarly, the most similar words to "sweet" were "intense," "dope," and "yay," while the most similar words to "dude" were "man," "bro," and "yessir." These examples demonstrate an important advantage of custom-trained word embeddings over non-customized approaches to building language models: They capture not only topical relationships, but also relationships between somewhat idiosyncratic cultural schemata that make up the meaning universe of the focal context—in this case, freelance software development teams. We conducted a wide range of most-similar queries for target words from within and outside the software development context and found that the model appeared to capture their meanings in contextually appropriate ways.

Second, Mikolov et al. (2013) show how mathematical operations in the vector space produced by an embeddings model can be used to solve analogical reasoning problems and can serve as a further check of model validity. For example, Mikolov et al. (2013) use their model to

evaluate the question, "Germany is to Berlin as France is to _____?". The answer is given by vec("Berlin") - vec("Germany") = x - vec("France"), or x = vec("Berlin") - vec("Germany") + vec("France"). Provided that the model learned vector representations of words correctly, the answer (in this case, "Paris") that the model returns will be given by the word vector that is closest to the coordinates that are obtained from this equation.

Applying this approach to our word embeddings model provided further evidence that our model performed well at capturing not just common semantic relationships, but also certain meanings that are idiosyncratic to the context of freelance software development. We tested our model through a number of word analogy tasks, some examples of which are shown in Table 1.

Both types of validity checks, most-similar queries and word analogy tasks, suggested that our model accurately captured context-relevant semantic relationships between words. Training a valid word embeddings model that performs well at capturing context-relevant semantic relationships was key to our subsequent analyses, since we were interested in capturing differences in the way team members convey meaning in interaction with one another.

---------- *Insert Table 1 about here* ----------

Descriptive Statistics

Table 2 shows summary statistics and bivariate correlations between the key dependent variable, the fraction of milestones delivered on time, and the key independent variable, discursive diversity, as well as various control variables. Teams varied considerably with respect to discursive diversity (mean: 0.578, SD: 0.213). Teams were relatively diverse with respect to roles and members' countries of origin, and relatively less diverse with respect to gender. On

average, 28.2% of team members were active on Slack per week (i.e., sent at least one message).[11]

Teams' Slack conversations were slightly skewed toward a positive tone (mean: 0.057, SD: 0.03; sentiment scores ranged from -1 for maximally negative to 1 for maximally positive language). Language tended to be relatively more objective than subjective (mean: 0.207; SD: 0.067; subjectivity scores ranged from 0 for maximally objective to 1 for maximally subjective language) and was neither extremely abstract nor extremely concrete, with a mean score of 2.55 (SD: 0.05) (the concreteness scale by Brysbaert et al. (2014) ranges from 0 for highly abstract words to 5 for highly concrete words).

Team performance, measured as the fraction of milestones delivered on time, was significantly and positively correlated with discursive diversity ($r=0.22$, $p<0.05$) and project price ($r=0.24$, $p<0.05$), but not with demographic diversity or any of the other control variables. Discursive diversity was significantly and positively correlated with team size ($r=0.27$, $p<0.05$), price ($r=0.36$, $p<0.05$), and role diversity ($r=0.36$, $p<0.05$). This makes intuitive sense, as larger teams that encompass individuals with more diverse professional backgrounds or that work on more complex projects are more likely to preserve a wider array of interpretations of various topics than smaller, less role-diverse teams working on less complex projects. Moreover, discursive diversity was significantly positively correlated with topical diversity ($r=0.29$, $p<0.05$), concreteness of the language teams used on Slack ($r=0.27$, $p<0.05$), and negatively with

[11] A weekly mean participation rate of 28.2% does not imply that team members were not engaged in the project. Rather, the modular nature of software development work implies that, except for the planning and brainstorming phase at the very beginning of the project, only a subset of team members are required to actively coordinate with each other during each subsequent phase. This subset of team members changes over time, as tasks need to be tackled in a certain order. For example, many projects begin with a collective ideation and brainstorming phase. Next, a designer might develop a mockup of the app's appearance, after which a backend engineer would build a prototype of the app in code. Frontend engineers and user interface experts tend to join the workflow in later stages once a backend prototype has been built. However, the detailed sequence of the workflow varies from project to project, and most projects involve considerable amounts of iteration, feedback-seeking and feedback-giving.

its subjectivity ($r=-0.38$, $p<0.05$). As we would expect, topical diversity and certain stylistic features explain some of the variation in discursive diversity, but discursive diversity still captures variation in meaning that is not accounted for by these covariates.

---------- *Insert Table 2 about here* ----------

Finally, we examined the degree to which discursive diversity varies over time within teams. Figure 1 plots team discursive diversity as a function of team life stage, where life stages of 0 and 1 correspond to the start and end points of the project, respectively. Each of the grey lines in the figure represents one team from a sample of 20 randomly selected teams. The blue line represents the mean level of discursive diversity for all 117 teams at a given life stage. The plot illustrates that individual teams' discursive diversity varies considerably over time, and that simply considering the mean of cognitive diversity would obscure all of this potentially informative variation. The average team in our dataset saw fluctuations of about 0.205 standard deviations around its mean level of discursive diversity, while the most volatile team experienced fluctuations of 0.305 standard deviations. Figure 1 illustrates the core advantage of our approach to measuring cognitive diversity relative to traditional survey-based measures: Even if collected at a few points in time during a team's life cycle, self-reports of cognitive diversity would simply be unable to capture the fine-grained temporal variation that language-based measures, such as ours, can uncover.

---------- *Insert Figure 1 about here* ----------

Covariate Balance Checks

Table 2 shows T-tests for covariate means before versus after coarsened exact matching, where discursive diversity was used as a treatment dummy variable. "Treated" teams were those with weekly average discursive diversity above the median level, while "control" teams were those

with weekly average discursive diversity at or below the median level. Before matching, high- and low discursive diversity teams differed significantly with respect to price and marginally with respect to team size. Using CEM, we were able to match 112 of the 117 teams. The 112 matched teams were included in our analyses for hypothesis 1. Teams in the matched sample were equivalent with respect to observable features such as gender and country diversity and engagement levels on Slack, but they differed with respect to role diversity. To account for this, we included role diversity as a control variable in all models.

---------- *Insert Table 3 about here* ----------

Main Results

Our first set of analyses focused on the main effect of discursive diversity on team performance. Consistent with hypothesis 1, we found a positive main effect of discursive diversity on team performance, measured as the fraction of milestones teams delivered on time. In Table 4, Model 1 predicts team performance based on role diversity alone, the only observable characteristic on which teams in the matched sample differed. Model 1 shows a positive and significant relationship between role diversity and performance. In Model 2, discursive diversity has a positive and significant relationship with team performance, and the coefficient for role diversity is no longer significant. Subsequent models indicate that the positive relationship between discursive diversity and performance is robust to the inclusion of controls for topical diversity (Model 3), variance in sentiment (Model 4), variance in subjectivity (Model 5), variance in concreteness (Model 6), as well as to the inclusion of all controls simultaneously (Model 7). In this final model, discursive diversity emerges as the only significant predictor of team performance. Figure 2 provides a graphical representation of this relationship.

---------- *Insert Table 4 about here* ----------

---------- *Insert Figure 2 about here* ----------

Our second set of analyses was concerned with time-varying effects of discursive diversity on teams' likelihood of meeting an impending deadline. We hypothesized that discursive diversity is beneficial to performance when the deadline is still relatively far away, and that it becomes detrimental to performance as the deadline draws nearer. Figure 2 shows the coefficient (log odds) of discursive diversity in logistic regressions predicting whether the impending milestone was met on time for 14-day long milestone phases (101 in total) that occurred across 40 teams. In the figure, the milestone deadline is represented by the red dashed line, while the x-axis represents days until the milestone deadline.

As Figure 3 illustrates, discursive diversity is positively associated with the likelihood of meeting the deadline until 6 days before the deadline ($p<0.001$), and negatively associated with the probability of meeting the deadline five to three days before the due date ($p<0.001$). Thus, consistent with hypothesis 2, discursive diversity appears to be beneficial for performance up until about one work week before the deadline, at which point it starts to become detrimental.[12]

---------- *Insert Figure 3 about here* ----------

DISCUSSION

The goal of this paper has been to help reinvigorate research on team diversity and performance by developing a novel conceptualization of cognitive diversity and introducing a language-based technique to measure it. Given that cognitive diversity is a superordinate construct,

[12] These findings are based analyses of 14-day milestone phases, which were the most common in our data by far (N=101 14-day milestone phases). When we examined milestone phases of different lengths (ranging from 10 to 31 days), we found a similar basic pattern of discursive diversity having a positive effect on timeliness when milestones were distance and detrimental to performance approximately one week before the deadline. Although the signs of the coefficients were consistent with expectations, many of the coefficients were not significant due to a loss of statistical power given that other milestone lengths were not as prevalent in our data.

we focused on one core dimension—discursive diversity—that reflects realized, rather than potential, divergence among group members; expressed, rather than latent, differences in the way members construe and ultimately communicate about a given set of topics; and temporal variation in construals and expressions over a team's life cycle rather than the assumption of stability. Using a word embedding model that defines a space of meaning, we constructed a time-varying measure of discursive diversity based on the dissimilarity of meanings expressed by members of a team. Using data from 117 remote teams of freelance software developers who collaborate via an online communication tool, and after accounting for the diversity of topics being discussed and variation in speaking styles, we found that discursive diversity is generally associated with better team performance. Yet considering mean levels of discursive diversity obscures considerable fluctuation that occurs over teams' life cycles. In more fine-grained analyses, we found that discursive diversity's effects on performance are contingent on time: it is positive when a team's next milestone is distant but turns negative as the next milestone approaches.

Contributions

This work contributes to research on team diversity and performance and cultural sociology. At the same time, it has potentially important implications for the design of diversity and inclusion initiatives in organizations.

We introduce to diversity research a novel construct, discursive diversity, which integrates social psychological and sociological perspectives on how people construe situations and substantive topics of discussion and how they then choose to express these meanings in interactional language use. Whereas the literature on team effectiveness has increasingly recognized the need for deeper understanding of temporal dynamics in group processes and outcomes (e.g., Mortensen & Haas, 2018), the workhorse research method in this field—self-

reports—is ill-suited to uncovering fine-grained temporal variation. Our approach demonstrates the utility of language as a complementary means to understanding various facets of cognitive diversity, how they vary over time, and what factors give rise to them.

We believe that language-based measures of team diversity can open up a number of promising pathways for future research. For example, how can leaders create climates—for example, by reinforcing norms of psychological safety (Edmondson, 1999)—that enhance discursive diversity when it is beneficial? Conversely, how can leaders tamp down discursive diversity when it is potentially detrimental but without reducing the team's capacity to be discursively diverse again in the future? Similarly, in the context of groups that tend to become more polarized over time (Myers & Lamm, 1976; Sunstein 2002; Macy, Kitts, Flache, & Benard, 2003), can the injection of members who introduce greater discursive diversity help moderate the tendency toward polarization? Beyond discursive diversity, we see the potential to develop new language-based measures of other established team cognition constructs such as need for cognition (Kearney, Gebert, & Voelpel, 2009), interpretive ambiguity (Kilduff, Angelmar, & Mehra, 2000), and problem-solving approaches (Kirton, 1989; Kurtzberg, 2005).

Results from this study also provide reasons to question the assumption in many diversity studies that demographic diversity is a useful proxy for cognitive diversity. Unlike prior work that has relied primarily on self-reports of cognitive diversity, our language-based measure of discursive diversity is not significantly correlated with gender diversity or diversity based on country of origin. Of course, demographic diversity may still be correlated with other facets of cognitive diversity that we do not measure, and demographic diversity may yield benefits to the team or the organization as a whole that are not reflected in our outcome measures. Nevertheless, insofar as discursive diversity matters for team performance, an open question for future research

is identifying the specific forms of demographic diversity that are most likely to yield discursively diverse teams.

Next, our work deepens our understanding of the contingent effects of time in team processes and performance. Gersick's (1988, 1989, 1991) seminal work on the punctuated equilibrium model focused on the midpoint of the team life cycle as a critical transition point with important implications for performance. We add nuance to this account by examining the cycles of exploration and execution at the level of team milestones. Modern work teams increasingly carry out projects that are structured into multiple milestone phases designed to achieve intermediate project goals. Our work advances theory to account for this shift in work practices.

In the realm of cultural sociology, our work builds on recent attempts to bring Swidler's (1986) cultural toolkit theory to the organizational realm (Weber, 2005; Giorgi, Lockwood, & Glynn 2015; Corritore, Goldberg, & Srivastava, forthcoming). Although the theory has been highly influential in cultural sociology, it has also been criticized for being vague and difficult to operationalize. Just as Corritore et al. (forthcoming) applied the theory to understand cultural heterogeneity among organizations, we demonstrate how the theory can be applied to conceptualize the diversity of cultural resources deployed by individuals who operate in teams. Moreover, being able to observe variation in the toolkits people use across time opens the possibility of more rigorously testing Swidler's proposition that innovation in toolkits is most likely to occur during unsettled times.

Finally, this work joins a growing body of research that draws a connection between cultural sociology and economic sociology (Khaire & Wadhwani, 2010; de Vaan, Stark, & Vedres, 2015; Askin & Mauskapf, 2017). We demonstrate how the use of cultural resources

within a team can have consequences for group productivity and, ultimately, financial performance.

Beyond these contributions to scholarly research, this work would appear to have potentially important implications for the design of diversity and inclusion practices in organizations. For example, our measure of discursive diversity could be used to predict when teams are at risk of missing milestones and introduce nudges that serve to dial discursive diversity up or down as needed. Such diagnostic tools and nudges could prove especially useful in the context of globally distributed teams, in which members often lack a shared context and shared identity and thus have demographic faultlines that are especially stark (Hinds & Mortensen, 2005). In a similar fashion, real-time dashboards of discursive diversity could serve as a useful barometer for how efforts to create inclusive work environments are translating into the breadth of ideas being expressed in work groups. Finally, individuals' past communication histories could be analyzed to help construct teams that are likely to have a high capacity for discursive diversity.

Limitations

While we believe that this work opens promising avenues for research and practice, we are also cognizant of its limitations. Even though we used coarsened exact matching to account for observable differences in the characteristics of teams that varied in discursive diversity, there is still the potential for unobserved heterogeneity among teams that is correlated with team performance. Experimental designs—for example, ones in which people's past communication behavior is used to randomly assign them into teams that are likely to vary in discursive diversity—are needed to pin down the causal link between cognitive diversity and performance.

Although our data include a relatively large number of participants from diverse geographies, it is important to keep in mind that our empirical setting involves high-end software development teams and individuals who have self-selected into the platform. Questions therefore remain about how our results would generalize to other contexts such as collocated teams and teams in which the focus is on routine task execution rather than knowledge production.

Finally, in our setting, team members communicated primarily through messages sent to the whole group. Although direct messages represented a relatively small fraction of group interactions, we did not have access to this form of communication. Thus, it remains unclear whether the discursive diversity team members expressed in one-one-one messages might have differed from that conveyed in group communications. In this regard, matching on team size was useful given that the proportion of direct messages exchanged likely varied as a function of size.

Conclusion

In sum, this study demonstrates the value of interactional language use and computational methods for the study of team diversity and performance. Relative to prevailing approaches to studying group heterogeneity, language offers an expressly different way to understand cognitive diversity, including subtle differences in the meaning people ascribe to substantive topics they discuss, and unlocks the role of time in explaining how dissimilarity relates to performance.

REFERENCES

Amabile, T. M., Conti, R., Coon, H., Lazenby, J., & Herron, M. (1996). Assessing the work environment for creativity. *Academy of Management Journal*, *39*(5), 1154-1184.

Askin, N., & Mauskapf, M. (2017). What makes popular culture popular? Product features and optimal differentiation in music. *American Sociological Review*, *82*(5), 910-944.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

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.

Corritore, M., Goldberg, A., & Srivastava, S. B. (Forthcoming). Duality in diversity: Cultural heterogeneity, language, and firm performance. *Administrative Science Quarterly.*

Cox, T. H., & Blake, S. (1991). Managing cultural diversity: Implications for organizational competitiveness. *Academy of Management Perspectives*, *5*(3), 45-56.

Dahlin, K. B., Weingart, L. R., & Hinds, P. J. (2005). Team diversity and information use. *Academy of Management Journal*, *48*(6), 1107-1123.

Detert, J. R., & Edmondson, A. C. (2011). Implicit voice theories: Taken-for-granted rules of self-censorship at work. *Academy of Management Journal*, *54*(3), 461-488.

De Vaan, M., Stark, D., & Vedres, B. (2015). Game changer: The topology of creativity. *American Journal of Sociology*, *120*(4), 1144-1194.

De Vaan, M., Stark, D., & Vedres, B. (2015). Game changer: The topology of creativity. *American Journal of Sociology*, *120*(4), 1144-1194.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*(2), 245-260.

Doyle, G., Goldberg, A., Srivastava, S., & Frank, M. (2017). Alignment at work: Using language to distinguish the internalization and self-regulation components of cultural fit in organizations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 603-612).

Fiol, C. M. (1994). Consensus, diversity, and learning in organizations. *Organization Science, 5*(3), 403-420.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117-132.

Ford, C., & Sullivan, D. M. (2004). A time for everything: How the timing of novel contributions influences project team outcomes. *Journal of Organizational Behavior*, *25*(2), 279-292.

Gersick, C. J. (1988). Time and transition in work teams: Toward a new model of group development. *Academy of Management Journal*, *31*(1), 9-41.

Gersick, C. J. (1989). Marking time: Predictable transitions in task groups. *Academy of Management Journal*, *32*(2), 274-309.

Gersick, C. J. (1991). Revolutionary change theories: A multilevel exploration of the punctuated equilibrium paradigm. *Academy of Management Review*, *16*(1), 10-36.

Giorgi, S., Lockwood, C., & Glynn, M. A. (2015). The many faces of culture: Making sense of 30 years of research on culture in organization studies. *The Academy of Management Annals*, *9*(1), 1-54.

Goldberg, A., Srivastava, S. B., Manian, V. G., Monroe, W., & Potts, C. (2016). Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *American Sociological Review*, *81*(6), 1190-1222.

Hallett, T., & Ventresca, M. J. (2006). Inhabited institutions: Social interactions and organizational forms in Gouldner's Patterns of Industrial Bureaucracy. *Theory and Society*, 35(2), 213-236.

Hambrick, D. C., Cho, T. S., & Chen, M. J. (1996). The influence of top management team heterogeneity on firms' competitive moves. *Administrative Science Quarterly*, 659-684.

Harding, D. J. (2007). Cultural context, sexual behavior, and romantic relationships in disadvantaged neighborhoods. *American Sociological Review*, *72*(3), 341-364.

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface-and deep-level diversity on work group cohesion. *Academy of Management Journal*, *41*(1), 96-107.

Hinds, P. J., & Mortensen, M. (2005). Understanding conflict in geographically distributed teams: The moderating effects of shared identity, shared context, and spontaneous communication. *Organization Science*, *16*(3), 290-307.

Horwitz, S. K., & Horwitz, I. B. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, *33*(6), 987-1015.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1-24.

Jackson, S. E., May, K. E., & Whitney, K. (1995). Understanding the dynamics of diversity in decision-making teams. *Team Effectiveness and Decision Making in Organizations*, *204*, 261.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly*, *44*(4), 741-763.

Joshi, A., & Roh, H. (2009). The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal*, *52*(3), 599-627.

Kearney, E., & Gebert, D. (2009). Managing diversity and enhancing team outcomes: the promise of transformational leadership. *Journal of Applied Psychology*, *94*(1), 77.

Kearney, E., Gebert, D., & Voelpel, S. C. (2009). When and how diversity benefits teams: The importance of team members' need for cognition. *Academy of Management Journal*, *52*(3), 581-598.

Khaire, M., & Wadhwani, R. D. (2010). Changing landscapes: The construction of meaning and value in a new market category—Modern Indian art. *Academy of Management Journal*, *53*(6), 1281-1304.

Kilduff, M., Angelmar, R., & Mehra, A. (2000). Top management-team diversity and firm performance: Examining the role of cognitions. *Organization Science*, *11*(1), 21-34.

Kirton, M. J. (Ed.). (1989). *Adaptors and innovators: Styles of creativity and problem solving*. Routledge.

Knight, D., Pearce, C. L., Smith, K. G., Olian, J. D., Sims, H. P., Smith, K. A., & Flood, P. (1999). Top management team diversity, group process, and strategic consensus. *Strategic Management Journal*, *20*(5), 445-465.

Kurtzberg, T. R. (2005). Feeling creative, being creative: An empirical study of diversity and creativity in teams. *Creativity Research Journal*, *17*(1), 51-65.

Lawrence, B. S. (1997). Perspective—The black box of organizational demography. *Organization Science*, *8*(1), 1-22.

Lewis, David. 1969. Convention: A philosophical study. Harvard University Press, Cambridge, MA.

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*. 66(3): 516-553.

Macy, M. W., Kitts, J. A., Flache, A., & Benard, S. (2003). Polarization in dynamic networks: A Hopfield model of emergent structure. *Dynamic Social Network Modeling and Analysis*, 162-173.

Martins, L. L., Schilpzand, M. C., Kirkman, B. L., Ivanaj, S., & Ivanaj, V. (2013). A contingency view of the effects of cognitive diversity on team performance: The moderating roles of team psychological safety and relationship conflict. *Small Group Research*, *44*(2), 96-126.

Martins, L. L., Schilpzand, M. C., Kirkman, B. L., Ivanaj, S., & Ivanaj, V. (2013). A contingency view of the effects of cognitive diversity on team performance: The moderating roles of team psychological safety and relationship conflict. *Small Group Research*, *44*(2), 96-126.

McGrath, J. E. (1991). Time, interaction, and performance (TIP) A Theory of Groups. *Small Group Research*, *22*(2), 147-174.

McPherson, J. M., & Rotolo, T. (1996). Testing a dynamic model of social composition: Diversity and change in voluntary groups. *American Sociological Review*, 179-202.

Michel, J. G., & Hambrick, D. C. (1992). Diversification posture and top management team characteristics. *Academy of Management Journal*, *35*(1), 9-37.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111-3119).

Mortensen, M., & Haas, M. R. (2018). Perspective—Rethinking Teams: From Bounded Membership to Dynamic Participation. *Organization Science*, *29*(2), 341-355.

Murnighan, J. K., & Conlon, D. E. (1991). The dynamics of intense work groups: A study of British string quartets. *Administrative Science Quarterly*, 165-186.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*(2), 166-180.

Olson, B. J., Parayitam, S., & Bao, Y. (2007). Strategic decision making: The effects of cognitive diversity, conflict, and trust on decision outcomes. *Journal of Management*, *33*(2), 196-222.

Pelled, L. H., Eisenhardt, K. M., & Xin, K. R. (1999). Exploring the black box: An analysis of work group diversity, conflict and performance. *Administrative Science Quarterly*, *44*(1), 1-28.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Phillips, K. W., & Loyd, D. L. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational Behavior and Human Decision Processes*, *99*(2), 143-160.

Scott, E. Page. 2007. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton, NJ: Princeton University Press.

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. (2017). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science*, *64*(3), 1348-1364.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, *10*(2), 175-195.

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, 273-286.

Swidler, A. 2001 Talk of Love: How Culture Matters. Chicago, IL: University of Chicago Press.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468-472.

Van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, *58*.

Weber, K. (2005). A toolkit for analyzing corporate cultural toolkits. *Poetics*, *33*(3-4), 227-252.

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400-415.

Williams, K. Y., & O'Reilly III, C. A. (1998). Demography and diversity in organizations: A review of 40 years of research. *Research in Organizational Behavior*, *20*, 77-140.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on the World Wide Web* (pp. 1445-1456). ACM.

Zenger, T. R., & Lawrence, B. S. (1989). Organizational demography: The differential effects of age and tenure distributions on technical communication. *Academy of Management Journal*, *32*(2), 353-376.

TABLES AND FIGURES

Table 1: Analogy Tasks Based on Word Embedding Model Trained on Entire Set of Teams' Slack Archives

Analogy task	Answer given by word embeddings model trained on team Slack archives
Bug - code = ?	Issue
Milestone + deliverable = ?	Sprint
Sprint - pressure = ?	Phase
Man + casual = ?	Dude
Instagram - photos = ?	Facebook
Machine - software = ?	Device
Machine + intelligent = ?	Brain
California - startup = ?	Australia
Human + desires + art = ?	Culture
Visual - creative = ?	Polish
Team - community = ?	@-tag

Table 2: Descriptive Statistics and Correlation Matrix

	Mean	St. Dev.	N	Fraction of milestones [illegible]	Discursive diversity	Team size	Price	Gender diversity	Country diversity	Role diversity	Fraction of team members [illegible]	Topical diversity	Sentiment	Subjectivity
Fraction of milestones delivered on-time*	0.000	1.000	117											
Discursive diversity	0.578	0.213	117	**0.22**										
Team size	8.111	4.267	117	0	**0.27**									
Price*	0.000	1.000	117	**0.24**	**0.36**	**0.29**								
Gender diversity	0.395	0.250	117	-0.1	-0.02	0.09	0.03							
Country diversity	0.597	0.196	117	-0.0.	0.05	**0.21**	0.05	0.1						
Role diversity	0.746	0.107	117	0.13	**0.36**	**0.56**	**0.2**	-0.07	**0.28**					
Fraction of team members active on Slack	0.282	0.952	117	0.17.	**-0.19**	**-0.49**	-0.06	**-0.23**	-0.1	**-0.26**				
Topical diversity	0.485	0.143	117	-0.02	**0.29**	**0.2**	0.09	0.01	**0.2**	**0.27**	-0.06			
Sentiment	0.057	0.030	117	-0.06	-0.05	-0.18	0.14	**0.19**	-0.01	-0.1	-0.07	**-0.27**		
Subjectivity	0.207	0.067	117	-0.14	**-0.38**	-0.17	-0.04	**0.22**	-0.02	-0.14	-0.04	**-0.26**	**0.74**	
Concreteness	2.550	0.050	117	0.14	**0.27**	-0.06	0.13	0.01	-0.08	-0.01	0.06	0.11	0.04	**-0.2**

** We present standardized versions of variables for the fraction of milestones delivered on time and price because they contain confidential information from the company that provided the data.*

*Note: Correlations significant at p<0.05 are shown in **bold**. For time-varying variables, weekly averages are shown.*

Table 3: T-Tests on Covariate Means Before and After Matching

	Estimate		Std. error		T-statistic		P-value	
Variable	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
Price	0.542	0.301	0.179	0.203	3.036	1.481	0.003*	0.142
Team size	0.341	0.068	0.183	0.14	1.864	0.489	0.065	0.626
Country diversity	-0.122	-0.205	0.185	0.184	-0.656	-1.115	0.513	0.267
Role diversity	-0.049	0.515	0.186	0.178	-0.262	2.898	0.793	0.005*
Gender diversity	-0.122	-0.141	0.185	0.182	-0.656	-0.774	0.513	0.441
Fraction of team members active on Slack	-0.261	-0.094	0.175	0.165	-1.489	-0.571	0.139	0.569

Table 4: Logistic Regressions of Timeliness (Hypothesis 1)

	Dependent variable:						
	Fraction of milestones delivered on time						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Role diversity	0.203*	0.117	0.112	0.112	0.110	0.108	0.107
	(0.100)	(0.107)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)
Topical diversity			0.261	0.253	0.272	0.201	0.188
			(0.647)	(0.647)	(0.650)	(0.653)	(0.659)
Discursive diversity		1.094*	1.056*	1.061*	1.049*	0.952*	0.945*
		(0.445)	(0.455)	(0.455)	(0.457)	(0.468)	(0.470)
Sentiment (var.)				-2.202			-3.345
				(7.072)			(7.961)
Subjectivity (var.)					-1.432		-1.054
					(7.949)		(8.854)
Concreteness (var.)						2.199	2.449
						(2.197)	(2.255)
Constant	0.656	0.098	-0.021	0.109	0.112	-0.327	-0.066
	(0.338)	(0.410)	(0.504)	(0.654)	(0.895)	(0.591)	(0.922)
Observations	112	112	112	112	112	112	112
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stratum weights	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Log Likelihood	-183.336	-180.256	-180.175	-180.126	-180.158	-179.666	-179.525
Akaike Inf. Crit.	386.672	382.512	384.349	386.252	386.317	385.333	389.051

Note: *p<0.05; **p<0.01; ***p<0.001

Table 5: Logistic Regressions of Timeliness of Individual Milestones for 14-Day Milestone Phases (Hypothesis 2)

Days from deadline	Log odds of discursive diversity	Std. Err.	P-value
-3	-0.657	0.292	0.025
-4	-1.298	0.271	0.000
-5	-2.098	0.076	0.000
-6	1.532	0.616	0.013
-7	2.166	0.563	0.000
-8	2.758	0.679	0.000
-9	3.757	0.828	0.000
-10	3.867	0.789	0.000
-11	4.003	0.447	0.000
-12	4.084	0.094	0.000
-13	3.549	0.233	0.000
-14	2.781	0.030	0.000

N = 101 milestone phases

Table 5 shows the log odds of discursive diversity in logistic regressions predicting the timeliness of individual milestones for 14-day milestone phases (n=101).

Figure 1: Discursive Diversity Across Team Life Stages

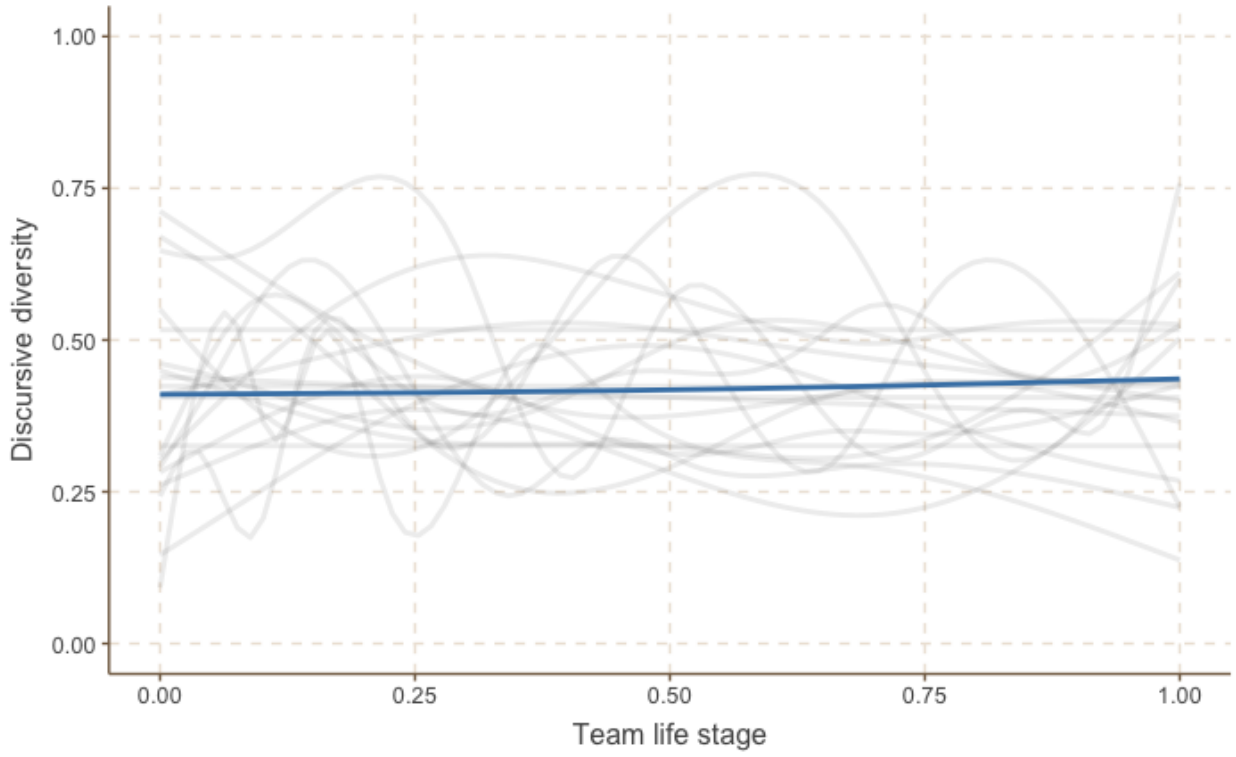

Figure 1 plots team discursive diversity as a function of team life stage, where life stages of 0 and 1 correspond to the start and end points of the project, respectively. Each of the grey lines represents one of 20 randomly sampled teams. The blue line represents the mean level of discursive diversity for all 117 teams at a given life stage. While survey-based approaches to measuring cognitive diversity might capture the time-aggregated mean, they would most likely fail to capture the variation in individual teams' discursive diversity over time.

Figure 2: Mean Levels of Discursive Diversity and Timeliness

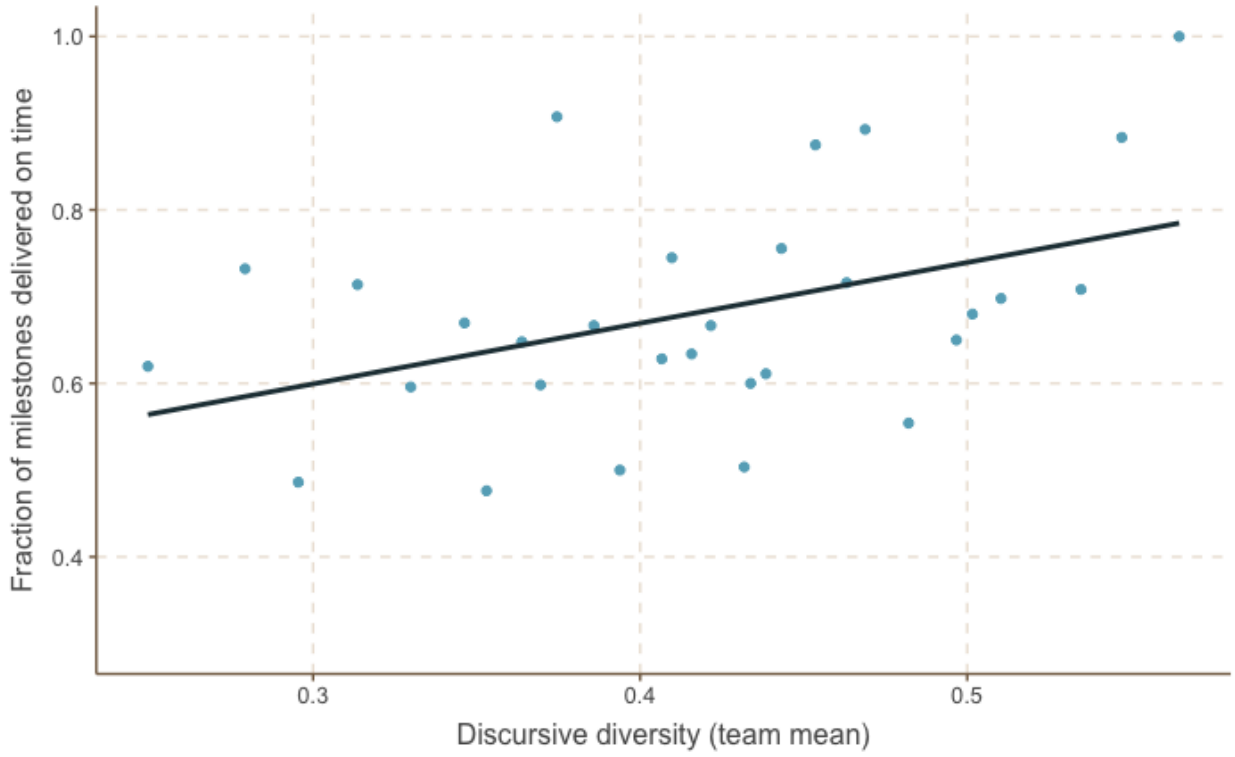

Figure 2 plots the fraction of milestones delivered on time as a function of teams' average daily level of discursive diversity for 112 teams that were matched using CEM. For the purposes of visualization, 112 teams were binned into 30 bins on discursive diversity. The black line represents best fit from logistic regression predicting timeliness (the fraction of milestones delivered on time) based on discursive diversity. The plot looks very similar when plotting the fraction of milestones delivered on time as a function of teams' average weekly level of discursive diversity.

Figure 3: Temporal Variation in Discursive Diversity's Effects on Timeliness (Hypothesis 2)

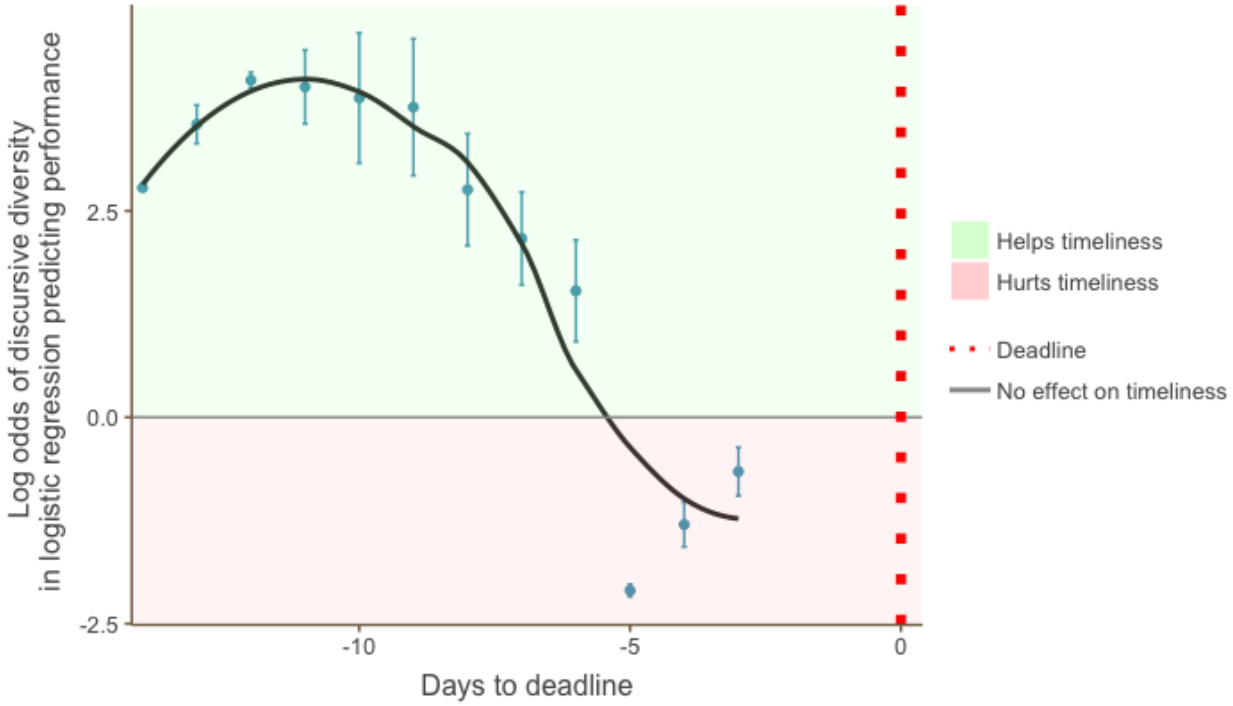

Figure 3 plots the impact (measured by the log odds) of discursive diversity on the probability of meeting an impending milestone deadline as a function of the number of days left until the deadline. The plot considers only milestone phases that were exactly 14 days in length (n=101) (the most common milestone phase duration in our data). For 14-day milestone phases, discursive diversity is associated with an increased probability (as indicated by positive log odds) of meeting the milestone on time up until 6 days before the deadline. Thereafter, discursive diversity reduces the probability of making the milestone on time. The same broad pattern of discursive diversity becoming detrimental to performance as deadlines draw nearer held for milestone phases of lengths other than 14 days, but not all coefficients were significant in those analyses due to loss of statistical power.