Quantifying the Causal Effects of Conversational Tendencies

JUSTINE ZHANG, Cornell University
SENDHIL MULLAINATHAN, Chicago Booth School of Business
CRISTIAN DANESCU-NICULESCU-MIZIL, Cornell University

Understanding what leads to effective conversations can aid the design of better computer-mediated communication platforms. In particular, prior observational work has sought to identify behaviors of individuals that correlate to their conversational efficiency. However, translating such correlations to causal interpretations is a necessary step in using them in a prescriptive fashion to guide better designs and policies.

In this work, we formally describe the problem of drawing causal links between conversational behaviors and outcomes. We focus on the task of determining a particular type of policy for a text-based crisis counseling platform: how best to allocate counselors based on their behavioral tendencies exhibited in their past conversations. We apply arguments derived from causal inference to underline key challenges that arise in conversational settings where randomized trials are hard to implement. Finally, we show how to circumvent these inference challenges in our particular domain, and illustrate the potential benefits of an allocation policy informed by the resulting prescriptive information.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing design and evaluation methods.*

Additional Key Words and Phrases: causal inference; conversations; counseling

ACM Reference Format:
Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the Causal Effects of Conversational Tendencies. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 131 (October 2020), 24 pages. https://doi.org/10.1145/3415202

1 INTRODUCTION

Conversations are central to the success of many consequential tasks. Understanding how to foster more effective discussions can guide computer-mediated platforms to better facilitate such endeavors as collaborating on large-scale projects [15, 26, 30, 33, 34], deliberating on law and policy [19, 37], informing and educating others [62, 69], or providing social support [11, 14, 55]. The growing availability of conversational data in these domains presents an opportunity to gain insights about what makes such discussions effective—a key step in improving these platforms.

One promising approach towards such insights is examining how people behave in more or less effective discussions. Indeed, past studies have highlighted various indicators of conversational behaviors that are tied to desired downstream outcomes such as successful persuasion [60, 68] or problem solving [15, 39], or improvements in emotional state [14, 55]. Conversational data was also used to characterize individuals in terms of their past conversational behaviors, providing rich signals of the roles they play in their interactions [67, 68] or their effectiveness in attaining

Authors' addresses: Justine Zhang, jz727@cornell.edu, Cornell University; Sendhil Mullainathan, sendhil@chicagobooth.edu, Chicago Booth School of Business; Cristian Danescu-Niculescu-Mizil, cristian@cs.cornell.edu, Cornell University.

2573-0142/2020/10-ART131 $15.00
https://doi.org/10.1145/3415202

conversational outcomes [1, 5, 9]. Translating these descriptive findings into *prescriptive* information, however, requires determining whether the relationships between conversational behaviors and outcomes is causal in nature.

Consider, for instance, the case of a psychological crisis counseling platform—a challenging, inherently interactional domain on which our work will focus. By analyzing logs of counseling conversations, the platform might identify particular patterns in how counselors behave during conversations where crises are successfully addressed, compared to during less successful interactions. For example, an analyst may observe that counselors who have more effective conversations also tend to exhibit more positive sentiment in their language [1, 47]. Does this imply that the platform should allocate more conversations to counselors who have the tendency to use more positive language?

Answering such prescriptive questions is a necessary prerequisite for ensuring that policies and interventions based on behavioral signals will have desired effects. This entails probing whether observed links between behaviors and outcomes are causal in nature, a difficult task especially in settings where running randomized trials is infeasible, and where behavioral dynamics are complex. Both of these traits are common to sensitive conversational domains such as crisis counseling.

In this work, we critically examine the task of drawing causal links between conversational behaviors and outcomes from observational data. Our particular aim is to concretely describe the challenges of this task and highlight cases where these challenges can be addressed. We approach this aim by way of analyzing a family of conversational settings, *goal-oriented asymmetric conversational platforms*, spanning domains like customer service, mental health counseling, interviewing and tutoring. In such settings, a platform has a dedicated roster of agents who seek to fulfill an outcome through having conversations with clients, such as an improvement in clients' mental well-being. Crucially, the platform has some leverage in allocating or training agents, but cannot specify the types of clients it serves. We focus on a particular type of policy that observational data could inform, and that is highly pertinent to these settings: allocating agents to upcoming conversations based on their observed past behaviors.

We first draw on the causal inference literature to provide a theoretical analysis of the inference problem involved [3, 53]. Using a potential outcomes framework, we articulate the causal effects we wish to estimate. Through formalizing the inference task, we highlight two key difficulties, one that inherits from the broader challenge of causal inference in observational settings, and another that directly derives from the interactive nature of conversations. This formalization also allows us to surface particular cases under which these challenges could be mitigated and to correspondingly propose solutions.

To empirically demonstrate the practical implications of our theoretical formulation, we instantiate these inference challenges and solutions in a large dataset of counseling conversations, obtained in collaboration with a text-based crisis counseling service. This highly consequential setting serves as a real-world example of the subclass of inference problems analyzed, and additionally illustrates the properties that enable these problems to be addressed. In this context, we show that accounting for the challenges we identified allows us to make more careful inferences than naive approaches, which often overestimate the strength of the causal relation between conversational behaviors and outcomes. We additionally probe the feasibility of an allocation policy that is based on these relations by simulating its implementation under idealized conditions, following prior work in the medical domain [31]. This simulation suggests that allocating counselors based on their conversational tendencies has the potential to improve the platform's effectiveness as long as spurious relations between tendency and outcome are discounted, and could motivate more realistic experimentation.

Our theoretical and empirical analyses focus on settings like the crisis counseling service as a more tractable case study, but serve to argue a broader point: transforming descriptive observations into prescriptive insights should be approached with care, especially in light of the particular challenges that conversational settings present. By explicitly articulating such challenges, we lay the groundwork for further efforts to analyze and address causal inference problems in a wider range of conversational settings.

2 BACKGROUND AND SCOPE

To clearly describe the task of making causal inferences about conversational behaviors, we focus on formally analyzing a narrower subset of such inference problems. In this section, we introduce and motivate our scope: we describe the family of conversational settings we will center our analyses around, along with the particular type of policy that we would like to inform. We also outline the causal inference challenges that we will later more rigorously take up.

Conversational setting. We analyze a category of conversational settings that we term *goal-oriented asymmetric conversational platforms*. Consider a platform that maintains a roster of *agents* who are expected to interact with incoming *clients*. First, the platform is *goal-oriented*: it has an overall objective that it seeks to use its agents to maximize. Second, it is *conversational* in the sense that agents work towards this objective by having conversations with clients, such that their behaviors within these conversations are consequential. Finally, the platform has an *asymmetric* degree of leverage: it can implement policies that affect its agents, but is unable to control its clients' characteristics.

This paradigm recurs across many often technologically-mediated domains like customer service [24, 41]—where sales representatives interact with customers, interviews—where interviewers interact with interviewees, and education—where teachers or tutors interact with students [20]. As an illustrative example that we later revisit in more depth, consider a crisis counseling helpline. The helpline employs a team of counselors who interact with individuals contacting the helpline in moments of mental crisis. The overall goal of this platform is to help these individuals in crisis; counselors work towards this goal by having conversations with them. The platform can select, train, or otherwise support its affiliated counselors. However, it would be infeasible and even unethical to restrict the types of people who seek help from it.

Allocation policy. We would like to examine how the platform can derive recommendations for managing its agents to better achieve its objective, subject to the limited influence it has over its clients. Here, our focus is on policies that impact how the platform *allocates* its agents. As a basic example, the platform may allocate more conversations to agents that it identifies as being more effective; as such, it may seek guidance on how best to select these effective conversationalists. Given the inherently conversational setting, we consider policies where agents are allocated on the basis of behaviors they exhibit over past conversations they've taken, which we refer to as behavioral *tendencies*. As such, we analyze when these aggregate tendencies—e.g., an inclination to use more positive language or to write longer messages—can be used by the platform to identify and hence allocate conversations to more effective agents. Intuitively, observing that certain behavioral signals are correlated with desired conversational outcomes would suggest that the platform should use tendencies inferred from these signals to allocate agents. The remainder of our work more rigorously examines this intuition.

We note that using behavioral tendencies to allocate agents is one of many policies that a platform could pursue. Here, we briefly outline some alternatives and motivate our particular focus. First, the platform may wish to allocate agents without accounting for their past behaviors—for instance, it may instead rely on past performance, or on demographic and personality attributes [6, 7, 29, 35, 66]. However, as noted in the introduction, past work illustrates that conversational behaviors can

provide rich signals of a conversation's outcome or an agent's characteristics; here, we specifically take up the potential usefulness of these signals in guiding concrete policies. We later empirically compare the effectiveness of conversational signals to these non-conversational attributes.

Second, more sophisticated allocation policies could extend the demonstrative approach considered here—for instance, the platform could match particular agents with conversations they are particularly well-suited for [36]. We leave an analysis of such policies to future work, noting that these more informed allocations require additional information (such as the nature of the client involved) which may not be readily available at the start of a conversation.

Finally, we contrast a policy of allocating agents with one that *trains* agents to adopt particular behaviors [2, 17, 57]. Both allocation and training-based policies could be informed by inferences about how behaviors and outcomes relate. We later revisit the training approach in the discussion (Section 5) to suggest that it shares the inference challenges that the allocation policy is subject to, but comes with additional difficulties as well, which we leave for future work.

Overview of inference challenges. As a precondition to implementing any allocation policy, the platform would need to ensure that the policy could actually have a desired effect. Concretely, we must consider a counterfactual question: if the platform had allocated another agent with a different tendency to a conversation, would the conversation have had a better outcome? While this question could in principle be addressed via randomized experiments, an experimental approach is often infeasible given the sensitivity of a conversational setting like counseling, and the difficulty of specifying treatments involving complex linguistic or interactional signals [18, 63]. Addressing such inherently counterfactual questions with observational data has been a core focus of causal inference (for surveys, see [3] and [53]). Such literature, however, has not dealt with the setting of conversations, which presents additional challenges that we identify and address in this work.

To outline the difficulties of the inference task in a conversational domain, consider a naive approach for relating conversational behaviors and outcomes: if we observe that good outcomes follow conversations where agents exhibit a certain behavior, we may naively infer that this behavior is a useful signal of effectiveness. For instance, suppose we find that client mood tends to improve after conversations involving agents who use language with a greater degree of positive sentiment. Such a finding could motivate us to allocate more positive agents to more future conversations.

At a high level, this initial approach suffers from a crucial pitfall: while an outcome may indeed arise as a result of an agent's behaviors, many *circumstantial* factors could also influence both the outcome and the nature of the conversation, and thus the behaviors that the agent exhibits. As such, our observations of the relation between a behavior and an outcome are confounded with circumstances of the conversation that neither the agents nor the platform can influence. For instance, an agent may say more positive things in a circumstance involving a congenial client who might also be more easily satisfied. However, in a situation involving a client with a genuinely difficult situation, a tendency for positivity may not even be appropriate, let alone effective. This means that a naive correlational approach cannot answer the counterfactual question posed above; in particular, the approach cannot inform us on how more positive agents would fare in conversations with less congenial clients.

3 FORMULATING THE INFERENCE TASK

We now proceed to more rigorously examine the entanglement between behavior, outcome and circumstance, focusing on the policy of allocating agents given their conversational tendencies. While the ideas we subsequently discuss are broadly relevant to other policies (such as those for training agents), we can more tractably address the inference task in the allocation policy, and hence present a formal analysis of this policy as an illustrative starting point.

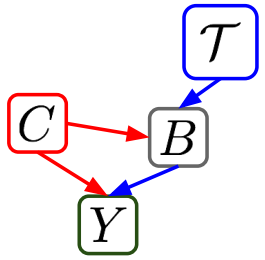

Fig. 1. Graphical representations of the key dependencies underlying the inference task, between tendency $\mathcal{T}$, outcome Y, behavior B and circumstance C. Our goal is to estimate the effect of tendency on outcomes (blue path), however the circumstances under which the behaviors and outcomes are observed confound this estimation (red arrows).

In particular, the allocation policy takes an aggregated view: the platform makes allocations based on how agents tend to behave over their past conversations. Intuitively, taking agent-level aggregates decouples our analyses from the circumstances of any one interaction: while an agent's behavior in a single conversation may be constrained by circumstantial peculiarities, over many conversations, their personal inclinations may materialize as conversational tendencies. Likewise, an agent may exhibit a systematic *propensity* to elicit certain outcomes, even if the outcome of a single interaction is contingent on its circumstance.

In what follows, we draw on the causal inference literature to formally examine the inference task underlying the allocation policy [3, 53]. First, we define this task in terms of the causal effect of allocation that we wish to estimate. We then discuss the challenges we face in quantifying this effect. We decompose these challenges into two key difficulties that stem from the *observational* nature of our data and the *interactional* nature of our conversational setting. We analyze each of these challenges by concretely identifying biases that arise in naive estimators of the effect of allocation, and then describe particular cases under which these biases can be addressed.

Inference task: estimating the allocation effect. Our goal is to evaluate the potential effectiveness of a policy that allocates agents to conversations, given their conversational tendencies. We now discuss the central measurement in this task, which corresponds to the counterfactual question introduced in the preceding section: given two agents J and K, who have different tendencies with respect to some behavioral signal (e.g., J tends to use more positive language than K), what is the effect of allocating one agent to a conversation versus the other, on a given outcome? We henceforth refer to this quantity as the *allocation effect*.

Under our observational approach, we wish to estimate the allocation effect from data on conversations that J and K have already taken. As such, we must use the data in two ways. First, we must use past observations to estimate the propensity of each agent to get a desired outcome (e.g., proportion of their clients who improved their mood). Second, we must estimate each agent's behavioral tendencies from their past conversations.[1]

In order for our estimate of the allocation effect to have a causal interpretation, we must ensure it can be directly ascribed to differences in the tendencies of J and K, rather than to differences in the circumstances of the conversations in which J and K's outcomes and behaviors were observed. As noted in the preceding section, these conversational circumstances can shape both the outcome of a conversation and an agent's behavior within the conversation, which thus become entangled. These

[1]Indeed, conversational data seldom comes with a priori labels of how agents tend to act; we may contrast this data-driven approach with self-reported indicators.

problematic dependencies are summarized in the graphical representation [43] depicted in Figure 1. We would like to estimate the effect of (allocating) tendencies $\mathcal{T}$ on outcomes Y (blue path); to this end, we must use behaviors B and outcomes Y observed under particular circumstances C. These circumstances can determine both behaviors and outcomes (red paths); our challenge is thus to somehow disentangle the effects of circumstances and tendencies.

Potential outcomes formulation. To formally highlight the biases that are incurred as a result of this entanglement, we mathematically express the allocation effect in terms of the potential outcomes framework [3, 53]. Let $\mathcal{T}$ be a random variable denoting a conversational tendency of agents, and suppose that agents J and K have different tendencies τ^J and τ^K. Let Y be a random variable denoting a conversational outcome. The allocation effect is then the expected difference in outcome if J, rather than K, is allocated to a conversation:

$$\mathcal{D}(\tau^J,\tau^K) = \mathbb{E}[Y \mid \mathcal{T} = \tau^J] - \mathbb{E}[Y \mid \mathcal{T} = \tau^K] \tag{1}$$

Let $\mathbb{D}(\tau^J,\tau^K)$ denote an estimate of $\mathcal{D}(\tau^J,\tau^K)$ from the data. Formally, this estimate has a causal interpretation if it is unbiased, i.e., $\mathbb{E}[\mathbb{D}(\tau^J,\tau^K)]=\mathcal{D}(\tau^J,\tau^K)$. Conversely, the estimate fails if it is contingent on the circumstances C under which the observed conversations occurred.[2]

As we have intuitively noted and as shown in Figure 1, such dependencies on C arise when we estimate Y with observed outcomes, and $\mathcal{T}$ with observed behaviors. We now proceed to articulate the challenges that are incurred from these dependencies. For each challenge, we provide an intuitive description supplemented with a graphical representation of the relationships between the variables involved [43], before drawing on potential outcomes arguments to formally express the corresponding biases [3, 53]. Our formal descriptions also point to particular settings with properties that enable us to mitigate these biases, and we discuss solutions that make use of these properties as well.

3.1 Estimating outcomes: bias from observed assignment

We first address the difficulties stemming from estimating agents' propensities for outcomes Y using our observations of their past conversations. To simplify the discussion, we provisionally suppose that we are given explicit labels of the agents' tendencies, returning to this point in the next subsection (3.2).

At a high level, our measurement of how tendencies relate to outcomes suffers from a problem that pervades observational studies: we can only observe outcomes in conversations that were actually *assigned* to agents exhibiting these tendencies. Here, we describe this problem in the context of conversations.

Let A denote the observed assignment—i.e., the matching between each agent J and their conversations in the data. The assignment mechanism potentially exposes different agents to contrasting circumstances: for example, agent J may be assigned to more challenging clients than K. As such, these assignment-induced differences in circumstance, rather than differences in the agents' tendencies, could drive observed differences in outcome. In this way, A skews our estimation of the allocation effect.

The graphical model depicted in Figure 2 highlights the problematic dependencies between tendency $\mathcal{T}$ and outcome Y, as indicated by the red edges: assignment A determines both $\mathcal{T}$ and C,

[2]Throughout, the notation we use adopts the following convention: uppercase denotes random variables (e.g., Y, $\mathcal{T}$ and $\mathcal{D}$ are random variables for conversational outcome and tendency, respectively), lowercase denotes realizations of these variables (e.g., τ^J is an observed value of $\mathcal{T}$), and empirical estimators are listed in blackboard bold (e.g., $\mathbb{D}$ is an empirical estimate of $\mathcal{D}$ based on the observed data).

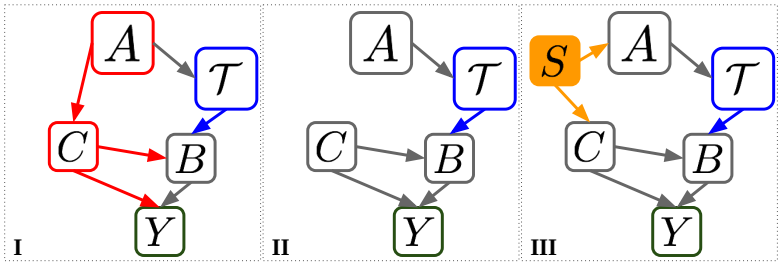

Fig. 2. Graphical representations of the dependence between assignment A and outcome Y through behavior B and circumstance C that result from the observational nature of our analyses, giving rise to the selection bias exposed in (3). **I**: the problematic pathways from A to Y; **II**: an idealized setting where conversations are randomly assigned to agents, in which the dependency is trivially broken; **III**: a scenario where assignment is governed by a set of observable *selection variables* S.

which in turn determine Y. As such, we cannot discount the effect of differences in assignment (red), beyond differences in tendency (blue), on the observed outcome.

Potential outcomes formulation. To surface the bias incurred from assignment, we formally examine the estimation of Y. As a first attempt, we can estimate the propensity of an agent J to get an outcome using the average outcome over their past conversations, denoted $\mathbb{Y}^J$. As such, we would measure $\mathcal{D}(\tau^J,\tau^K)$ as $\mathbb{D}(\tau^J,\tau^K) = \mathbb{Y}^J - \mathbb{Y}^K$.

We note that our empirical estimators are contingent on A, i.e., we can only observe J in the conversations in which they actually participated. As such,

$$\mathbb{E}[\mathbb{Y}^J] = \mathbb{E}[Y|\mathcal{T} = \tau^J, A = J]$$

Substituting this expression into the above equation for $\mathbb{D}(\tau^J,\tau^K)$, we see that our estimator of the allocation effect is biased:[3]

$$\begin{aligned}
\mathbb{E}[\mathbb{D}(\tau^J,\tau^K)] &= \mathbb{E}[\mathbb{Y}^J - \mathbb{Y}^K] \\
&= \mathbb{E}[Y|\mathcal{T}=\tau^J, A=J] - \mathbb{E}[Y|\mathcal{T}=\tau^K, A=K] \\
&= \mathbb{E}[Y|\mathcal{T}=\tau^J, A=J] - \mathbb{E}[Y|\mathcal{T}=\tau^K, A=J] \qquad (2) \\
&+ \mathbb{E}[Y|\mathcal{T}=\tau^K, A=J] - \mathbb{E}[Y|\mathcal{T}=\tau^K, A=K] \qquad (3)
\end{aligned}$$

The equations highlight that our observed difference could have two sources. The first (2) corresponds to the effect of varying the tendencies over a shared set of circumstances (i.e., that were assigned to J). This is the value we need to estimate in order answer the counterfactual question: what outcomes would have been attained had the conversations that were assigned to J been instead handled by an agent with a different tendency τ^K? The second (3) reflects the *selection bias* that arises because J and K were actually exposed to different circumstances via assignment, as illustrated in Figure 2I.

An idealized setting: random assignment. As with many causal inference questions, selection bias would be eliminated if agents were *randomly assigned* to conversations, and are hence exposed to the same distributions of circumstances. As such, observed differences in outcome could no longer be ascribed to assignment-induced differences in circumstance. Formally, random assignment

[3]In the last derivation we subtract and re-add the second term.

makes assignment and outcome independent for each agent (Figure 2II), such that the problematic term (3) trivially cancels out.

However, this selection bias remains in more realistic conversational settings, where assignment mechanisms are seldom random. In the extreme, if an agent *selects* their conversations, a record of positive conversational outcomes could be ascribed to picking clients who are easier to help, rather than having some replicable conversational proficiency. The problem persists beyond self-selection—e.g., agents who work during the day may encounter more congenial clients than those who work at night.

A limited solution: controlling for circumstance. We may try to mitigate selection biases by controlling for the conversational circumstances C, for instance by comparing $\mathbb{Y}^J$ and $\mathbb{Y}^K$ only over conversations that match on attributes of the circumstance, e.g., are about the same issue. Indeed, many prior studies of conversations have employed such techniques [14, 27, 42, 55, 59, 60, 70]. *Completely* controlling for circumstance certainly breaks the problematic pathway from A to Y: Figure 2I shows that the two variables are conditionally independent given C and $\mathcal{T}$ (formally written as $Y \perp\!\!\!\perp A \mid \{C, \mathcal{T}\}$).[4]

However, this approach is fundamentally limited: we can only control for the circumstantial attributes that we can observe. This leaves other important but inaccessible aspects (e.g., the client's mental state) unaccounted for.

Tractable setting: observed selection variables. We now describe a subset of settings under which this bias can be mitigated, involving assignment mechanisms with some additional structure. In particular, suppose that assignment is random up to a set of completely observable *assignment selection variables* S (Figure 2III, orange edges). As a natural example, consider conversational platforms where agents work during different *shifts*, and clients are randomly assigned to agents within each shift time. While different agents and clients may select different shifts, within a single shift these factors play no role in who gets assigned to whom; furthermore, for each conversation, the platform knows the shift in which it took place. Beyond shift times, other examples of selection variables include geographic location and organizational divisions like departments of a store.[5]

Importantly, conditioning on S breaks the pathway between A and Y; that is, Y and A are conditionally independent given S and $\mathcal{T}$ ($Y \perp\!\!\!\perp A \mid \{S, \mathcal{T}\}$). Controlling for selection variables can be seen as a special case of controlling for observable circumstantial attributes, where we know how these attributes S are related to the assignment mechanism. *Within each value of the selection variable*, our observations of agents' conversational outcomes are hence decoupled from circumstantial differences due to assignment. As such, we modify our estimator to first measure the allocation effect for a particular selection variable (e.g., within a shift), comparing outcomes attained by agents with tendencies τ^J versus τ^K only for conversations with that selection variable.

Formally, for a given selection variable s, denote the corresponding estimator of the allocation effect as $\mathbb{D}(\tau^J, \tau^K \mid S = s)$. By conditional independence, we have that:

$$\begin{aligned}
&\mathbb{E}[Y \mid \mathcal{T} = \tau^J, A = J, S = s] \\
&= \mathbb{E}[Y \mid \mathcal{T} = \tau^J, A = K, S = s] \\
&= \mathbb{E}[Y \mid \mathcal{T} = \tau^J, S = s]
\end{aligned}$$

[4] Conditional independence corresponds to the criterion of d-separation in the graphical representation [43].

[5] We are effectively using the assignment of agents as valid instrument, conditional on shift, for the kinds of conversation the client is exposed to [4, 8, 44].

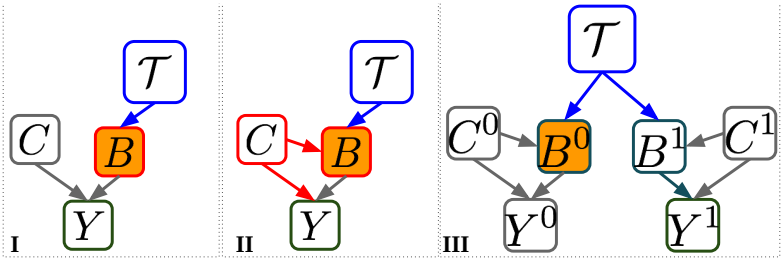

Fig. 3. Graphical representations of the entanglement between circumstances C, behaviors B and outcomes Y, giving rise to the bias in (6). **I**: dependencies in a non-interactional setting; **II**: problematic dependencies when B interacts with circumstances C that also shape Y; **III**: our approach, observing behaviors and outcomes on different splits of data.

Thus, after conditioning on S, the bias (3) cancels out. That is, among conversations with the same S, empirical differences in outcome are entirely driven by tendency:

$$\begin{aligned}&\mathbb{E}[\mathbb{D}(\tau^J,\tau^K|S=s)]=\\&=\mathbb{E}[Y|\mathcal{T}=\tau^J,S=s]-\mathbb{E}[Y|\mathcal{T}=\tau^K,S=s]\end{aligned} \tag{4}$$

Repeating this matching process across all S then yields an aggregate measurement of outcome differences arising from varied tendencies, rather than from differences in assignment.

3.2 Estimating tendencies: bias from interactional effects

We now address the difficulties stemming from estimating agents' tendencies $\mathcal{T}$ using our observations of their past behaviors. To simplify the discussion, we suppose that the difficulty in estimating outcomes, as described in the preceding section (3.1), has been fully addressed.

At a high level, the problem we face stems from the interactional nature of conversations: the behavior of an agent both shapes, and is constantly shaped by the behavior of the other participant. As such, our measurement of an agent's tendencies, and hence our inferences about the relation between tendency and outcome, is skewed by the circumstances that agents inevitably react to in conversations. At an extreme, we may observe that agents say "you're welcome" precisely after clients thank them. This does not necessarily mean that saying "you're welcome" is a behavioral inclination some agents have, beyond a reaction to the preceding interaction; it certainly does not follow that we should encourage more frequently saying "you're welcome".

An interactional problem. As a thought experiment, consider a *non-interactional* domain where an agent's behavior can affect an outcome without any interaction with the client—a "secret santa" paradigm where an agent, the gift-giver, has no back and forth with their recipient (Figure 3I). In this case an agent's behavior is purely a reflection of the agent's tendencies (e.g., an inclination for cheap gifts); and an empirical mismatch between $\mathcal{T}$ and B simply reflects the noise with which a tendency gives rise to a behavior. As we accrue more observations of the agent, we would expect such mismatches to diminish.

In contrast, in an interactional setting, these factors are problematically entangled (Figure 3II, red path): since the agent inevitably reacts to the client's behavior, B reflects C as well as $\mathcal{T}$. Furthermore, C can impact outcomes Y. An agent's observed behavior hence constrains the distribution of C that could have yielded our observed outcomes; as with nonrandom assignment, differences in observed outcomes once again could reflect differences in circumstance as well as in tendency.

Potential outcomes formulation. Formally, let B be a random variable denoting observed agent behaviors. We use an aggregate of J's past behaviors, denoted $\mathbb{B}^J$, to measure τ^J. Our empirical estimators are hence contingent on these observed behaviors:

$$\mathbb{E}[\mathbb{Y}^J] = \mathbb{E}[Y \mid \mathcal{T} = \tau^J, B = \mathbb{B}^J]$$

Again, we highlight the bias in estimator $\mathbb{D}(\tau^J,\tau^K)$:

$$\begin{aligned}
&\mathbb{E}[\mathbb{Y}^J - \mathbb{Y}^K] \\
&= \mathbb{E}[Y \mid \mathcal{T} = \tau^J, B = \mathbb{B}^J] - \mathbb{E}[Y \mid \mathcal{T} = \tau^K, B = \mathbb{B}^K] \\
&= \mathbb{E}[Y \mid \mathcal{T} = \tau^J, B = \mathbb{B}^J] - \mathbb{E}[Y \mid \mathcal{T} = \tau^K, B = \mathbb{B}^J] \qquad (5) \\
&+ \mathbb{E}[Y \mid \mathcal{T} = \tau^K, B = \mathbb{B}^J] - \mathbb{E}[Y \mid \mathcal{T} = \tau^K, B = \mathbb{B}^K] \qquad (6)
\end{aligned}$$

As before, two factors contribute to the observed difference in outcome. The first (5) arises from a difference in tendencies. The second (6), as we've described above and as depicted in Figure 3II, reflects a difference in circumstances, and is inherent to the interactional nature of conversations.

A limited solution: ignoring the interaction. The factor in (6) intuitively compounds as the conversation progresses and an agent's behavior becomes increasingly contingent on the circumstances. As such, we may seek to dampen this bias by only considering behaviors from the start of the conversation, before behavior and circumstance become tightly coupled. Indeed, prior work has taken this limited view of conversations [1, 70] with this confound in mind. However, insofar as this approach does not fully address the bias incurred by interaction, it also constrains the scope of the conversational tendencies we can study.

Tractable setting: separable sets of conversations. To factor out this interactional bias, we must decouple our observations of agent behaviors and outcomes from the conversational circumstances they are both tied to. We consider a simple fix: for each agent, we measure their behaviors over a *subset* of the conversations they've taken, and use a *separate* set of conversations to measure the outcomes they elicit.[6]

Formally, suppose we split each of B, Y, C into two random variables, one for each subset. As shown in Figure 3III, the only pathway connecting an agent's behaviors and outcomes *across* these splits is via their conversational tendencies. That is, B^0 and Y^1 are conditionally independent given $\mathcal{T}$ (i.e., $Y^1 \perp\!\!\!\perp B^0 \mid \mathcal{T}$), so

$$\mathbb{E}[Y^1 \mid \mathcal{T} = \tau^J, B^0 = \mathbb{B}^{J,0}] = \mathbb{E}[Y^1 \mid \mathcal{T} = \tau^J]$$

and the bias term (6) cancels out.

Such a solution is applicable to conversational platforms where agents take many conversations, and where different subsets of these conversations are separable from each other, as in a common scenario where clients contact a platform for ad-hoc purposes. We may contrast these conditions with settings in which clients influence each other or recur across multiple interactions.

[6]While we solve a different problem, our solution is analogous to separating train and test sets to mitigate overfitting [18]—here we "train" our measurements of tendencies and "test" their effects on separate data splits. Note that throughout, we use "subset" to refer to a collection of conversations, not to a subset of messages within a single conversation.

4 EMPIRICAL DEMONSTRATION

Having developed a general description of our inference task and its challenges, we now demonstrate these ideas empirically. In particular, we consider a real-world example of an asymmetric conversation platform: a large-scale crisis counseling service.

In what follows, we introduce this setting, describe the particular dataset we examine, and explain how it is illustrative of our theoretical formulation. We then study the allocation effect of a few simple tendencies, comparing naive estimators to approaches informed by our preceding analyses. Finally, we estimate the effects of a policy of allocating counselors via a simulated experiment, showing how a careful consideration of conversational tendencies can provide an informative starting point in evaluating this policy.

4.1 Setting: Crisis counseling conversations

The crisis counseling platform provides a free 24/7 service where *counselors*—playing the role of agents—have conversations via text message with clients in mental distress who contact the platform, henceforth *texters*. We accessed the complete collection of anonymized conversations in collaboration with the platform, Crisis Text Line,[7] and with IRB approval; counselors and texters have consented to make their data available for research purposes.

The counseling platform is a particularly consequential example of a goal-oriented asymmetric conversational platform. The platform's overall goal is to better support texters through their distress; counselors aim at this objective in each conversation they take with a texter. These conversations are challenging and complex, and are typically quite substantial, averaging 26 messages long. They give rise to a rich array of conversational behaviors and interactional dynamics which may impact a texter's experience [1, 71, 72].

Conversational clients: texters. Texters contact the platform with a variety of issues, ranging from depression to work problems to suicidal ideation. They encompass a broad range of demographic and geographical attributes; seasonal changes and current events can also shape the types of texters who contact the service. Crucially, the service is open for anyone to reach out to at any time. Conversations with these texters thus span a challenging diversity of circumstances.

Platform agents: counselors. Counselors with the service are dedicated volunteers who are selected and trained by the platform. The long-term nature of counselors' engagement with the platform underlines the practical relevance of an agent-level allocation policy. In taking a long-term view of counselor behavior, we focus on analyzing the subpopulation of 4,861 counselors who take at least 80 conversations. These counselors constitute 34% of the total population, and have taken a total of 1,180,473 conversations, or 83% of conversations on the platform to date. Henceforth, all of the statistics we report are computed over the first 80 conversations taken by each of these sufficiently prolific counselors.

Conversational outcomes. For the purposes of our present demonstration, we consider two complementary signals of a conversation's outcome that are used by the platform to assess the conversations that take place on it:

- **Texter rating**: After each conversation, texters are surveyed on whether or not the interaction was *helpful*. Out of 29% of conversations that receive ratings, 87% are positively rated. Prior computational analyses of counseling conversations have also used such survey responses as an indicator of conversation quality [1, 71].
- **Conversational closure**: Ideally, conversations *close* at a moment that feels appropriate for both the counselor and texter. However, not all texters remain engaged in a conversation. A

[7] Access to the data is by application, at https://www.crisistextline.org/data-philosophy/research-fellows/. The extensive ethical considerations, and policies accordingly implemented by the platform, are detailed in Pisani et al. [48].

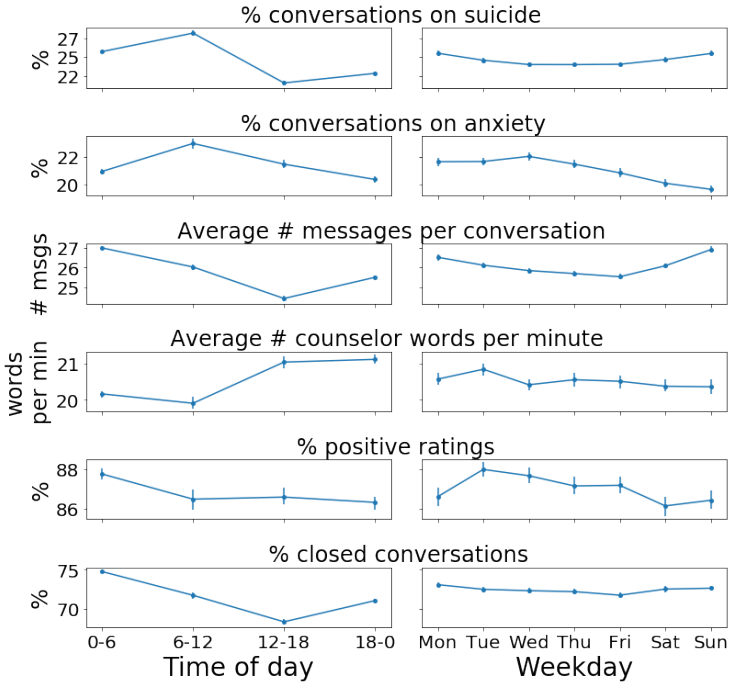

Fig. 4. Issue frequency, average conversation length, counselor speed, and outcomes across a day or week. Error bars represent bootstrapped 95% confidence intervals.

counselor who is faced with an unresponsive texter ends the conversation after following a standardized protocol specified by the platform.[8] 72% of conversations are properly *closed* in this sense, while the rest end in texter disengagement.

In general, evaluating the success of a counseling conversation is difficult [23, 61]. For instance, eliciting feedback from texters is challenging, as evidenced by the relatively low proportion of post-conversation surveys with responses; the ratings obtained may also reflect a biased sample of texters who decide to fill out the survey.[9] On the other hand, while the closure outcome is well-defined over all conversations, it is a less precise indicator—texters may disengage from a conversation for numerous reasons which may be extrinsic to the interaction, such as a low phone battery.[10] Given the focus of our work on rigorously probing potential causal relations between behavioral tendencies and outcomes, we leave the problem of developing richer and more reliable measures of conversational outcomes to other work. We also note that alternative outcomes would be still be subject to the circumstantial confounds detailed in our general description in Section 3.

Circumstance-based confounds. Intuitively, both outcomes are heavily dependent on the texter a counselor interacts with, and the context in which a conversation takes place. This underlines

[8]In particular, this standardization minimizes the impact of the counselor's inclinations in determining this outcome.

[9]We note that that there is an insignificant correlation (Kendall's tau = 0.02) between the propensity of a counselor to receive ratings, and their propensity for positive ratings (among their rated conversations), suggesting that at the counselor level, analyses focusing on counselors for whom enough ratings are observed would not be skewed towards counselors whose conversations are better- or worse-perceived.

[10]The potential for several factors beyond what's recorded in the data to relate to closure exemplifies the limited efficacy of controlling for observable attributes of the circumstance.

the need to address circumstantial factors when relating behavioral tendencies to outcomes. As a concrete demonstration of the salience of these circumstantial factors, we observe that the time at which a conversation takes place is related to the types of issues a texter experiences, a counselor's conversational behaviors, and the outcomes that result, as shown in Figure 4. For instance, the share of conversations involving suicidal ideation peaks in the morning and drops in the afternoon (28% vs. 22% of conversations); conversations are especially long on Sundays and short on Fridays (26.9 vs. 25.5 messages); more conversations are closed from midnight to 6 AM than from noon to 6 PM (75% vs. 68% of conversations).

4.1.1 Tractability. Thus far, we have suggested that the counseling platform is representative of the inference task and challenges we formally described. We additionally assume that the platform exhibits the properties under which these inference challenges can be mitigated; our framework and the validity of these assumptions were informed by interactions with the platform's staff and engineers. Nonetheless, the assumptions we make are necessarily simplifying, especially as we do not directly control how the platform operates. For the purposes of demonstration, we argue that these assumptions are well-founded, and revisit potential limitations in the discussion (Section 5).
Observed selection variables. We assume that the assignment of conversations to counselors is random up to the *shift times* that counselors sign up to take. As such, these shift times correspond to the fully-observable selection variables S with which the bias from assignment can be addressed (Section 3.1). This assumption reflects the platform's actual assignment process: while counselors can choose which shifts to sign up for, the platform assigns counselors to conversations randomly. For our demonstration we model shifts as temporal bins spanning the same 3-month window, day of week, and 6 hours of the day (e.g., Wednesdays from 12 to 6 AM, in January to March 2017).
Separable sets of conversations. We also assume that there are no dependencies between different conversations taken by a counselor, allowing us to address the bias from interaction (Section 3.2). In particular, given the platform's focus on providing support in acute crises, texters generally do not contact the service repeatedly; further, the platform does not deliberately assign repeat texters to the same counselor (i.e., in contrast to a therapy-oriented service).

4.2 Analysis: Relating tendencies and outcomes

In what follows, we use the counseling setting to empirically illustrate the inference challenges we formulated, as well as the solutions we proposed to address them. As a demonstration, we examine the extent to which a representative selection of conversational tendencies are related to the outcomes we've described, in terms of their allocation effect. We consider three approaches to estimating the strength of the relation between a conversational tendency and outcome. The estimates returned by each approach are denoted by different marker shapes in Figure 5 (△, □, ◯); stronger effects are indicated by points further from the vertical line (indicating no effect). The approaches are successively more rigorous in addressing the inference challenges, and contrasts in the effect sizes they estimate are visually represented as the different horizontal positions of the markers. In comparing between these estimators, we show how a more careful analysis informed by our causal arguments (represented as ◯) can distinguish between tendencies that could usefully guide an allocation policy, versus those that are related to outcome by virtue of circumstantial confounds (represented as △ and □), and that the platform may have less leverage over.
Conversational behaviors. Past work in counseling has suggested a range of conversational behaviors which relate to counseling effectiveness [21, 23, 38, 40, 49–52, 64]. Extending these efforts, recent computational studies have highlighted conversational features that could signal positive outcomes in counseling conversations [1, 10, 46, 47, 71], as well as other settings in the mental health domain like online support forums [14, 58] or longer-term therapy [13, 45, 56]. The

question of causality has largely been outside the scope of such studies;[11] here, we take up this question in terms of the allocation policy. For the purposes of demonstration, we focus on a small set of behaviors, which we selected as simple representatives of prevalent types of conversational behaviors considered in these past works. These behaviors are listed in Figure 5, along with studies which have demonstrated their correlations with mental health-related outcomes. In particular, conversation length, response length and response speed relate to the *fluency and pace* of the conversation [1, 13, 47]; sentiment is a frequently-cited attribute of the *style or tone* of an utterance [1, 13, 47, inter alia]; lexical similarity between utterances and linguistic coordination are often used to characterize *interactional* behaviors [1, 58] like adapting to a client's language or reflecting their concerns.[12]

Each of the behaviors we consider is potentially subject to the assignment- and interaction-based challenges we have described. As shown in Figure 4, they may be biased by a circumstantial factor like shift time. They may also reflect both the counselor's own conversational aptitude—e.g., their ability to manage conversational progress, maintain a helpful tone, and meaningfully respond to the texter—and the texter's inclinations—e.g., their responsiveness, emotional state, and openness to disclosing information. The subsequent analyses therefore clarify the extent to which the relations between these behaviors and outcomes, as surfaced by prior work, have causal interpretations in terms of the allocation effect.

Naive formulation: conversation-level effects. We first compare counselor behaviors in conversations that are rated positively versus in those rated negatively, as well as in conversations that are closed versus in those where the texter disengages. For each conversation-level behavior and outcome, these comparisons yield statistically significant differences (Mann-Whitney U test $p < 0.01$), echoing several correlations reported in prior work between behaviors and outcomes in individual conversations. As we have argued, the usefulness of these relationships in guiding policies is unclear, since they could reflect circumstantial factors that the platform cannot influence. For instance, the sentiment of counselor messages is significantly more positive in positively- versus negatively-rated conversations; this could reflect the benefits of an upbeat tone, or that distressed texters who are harder to help also tend to discuss less positive material. At the extreme, closed conversations are much longer than disengaged ones (28.4 vs. 20.0 messages per conversation on average), perhaps tautologically: disengaged conversations end prematurely by definition.

Counselor-level correlations ($\triangle$). To build up to a counselor-level approach that addresses the influence of circumstance, we first consider correlations between counselor-level aggregates of behavior $\mathbb{B}$,[13] and of outcome $\mathbb{Y}$ (computed as a counselor's proportion of positively-rated or closed conversations). This view corresponds to the counselor-level approach taken in Althoff et al. [1].[14]

To quantify the extent to which an aggregated behavior $\mathbb{B}$ relates to an outcome propensity $\mathbb{Y}$, we compute Kendall's tau correlations between $\mathbb{B}$ and $\mathbb{Y}$, depicted in Figure 5 as $\triangle$. At a high level, Kendall's tau compares the rankings of counselors according to $\mathbb{B}$ and according to $\mathbb{Y}$ by capturing

[11] In support forum settings, Choudhury and Kıcıman [14] and Saha and Sharma [55] similarly examine the causal effects of linguistic behaviors on outcomes such as the risk of suicidal ideation; the techniques they employ can be seen as controling for circumstance using observable attributes, which we contrast with our use of observed selection variables in Section 3.1.

[12] We measure a counselor's speed in a conversation as the number of words they write, per minute taken to reply to a texter. Following Althoff et al. [1], we measure sentiment as the VADER compound score of each message [25] and similarity as the cosine similarity between a counselor's message and the texter's preceding message; we obtain conversation-level measures of response length, sentiment and similarity by averaging over the counselor's messages in a conversation. As in Althoff et al. [1], we use the approach in [16] to measure coordination, noting that it produces a counselor-level, as opposed to a conversation-level score.

[13] With the exception of coordination, which is already a counselor-level property, we derive counselor-level aggregates by averaging a counselor's per-conversation behaviors, e.g., average sentiment.

[14] Note that Althoff et al. [1] only consider the top and bottom 40 counselors in terms of $\mathbb{Y}$, while we consider all counselors.

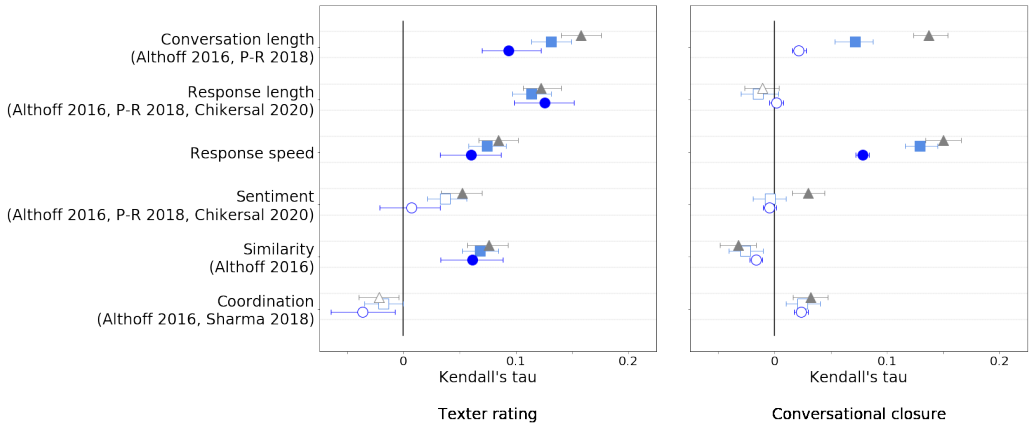

Fig. 5. Relation between counselor-level behavioral tendencies and outcomes, measured as Kendall's tau correlations, in increasingly controlled settings: △ correlates counselor behavior and outcome propensity; □ computes this correlation across temporally-interleaved splits of conversations; ◯ further controls for shift time, thus reflecting the allocation effect formulated in Equation 1 while accounting for the inference challenges described in Sections 3.1 and 3.2. Error-bars show bootstrapped 95% confidence intervals; shapes are filled for bootstrapped and Bonferroni-corrected $p < 0.01$. Abbreviated citations indicate studies that have demonstrated correlational relationships between the respective behaviors and outcomes.

the extent to which, within each pair of counselors, differences in $\mathbb{B}$ are in the same direction as differences in $\mathbb{Y}$. This mirrors our formulation of the allocation effect from Equation 1, which is likewise defined over pairs of counselors; here, however, we make the naive assumption that $\mathbb{B}$ and $\mathbb{Y}$ correspond to estimates of counselor tendency and outcome that can be meaningfully related (i.e., we ignore the two sources of bias).

Addressing bias from interaction (□). As described in Section 3.2, both $\mathbb{B}$ and $\mathbb{Y}$ are entangled with the circumstances of conversations by virtue of the interaction between counselors and texters. To mitigate the bias from interaction, we divide each counselors' conversations into two splits, such that split 0 consists of their even-indexed conversations (i.e., the second, fourth, sixth, …) and split 1 consists of their odd-indexed conversations. Using Kendall's tau, we compare the ranking of counselors according to their average behavior $\mathbb{B}^0$ over split 0 with their ranking based on their outcome propensity $\mathbb{Y}^1$ over split 1, depicted as □ in Figure 5. As such, $\mathbb{B}^0$ and $\mathbb{Y}^1$ correspond to estimates of counselor tendency and outcome which address this source of bias.

Addressing bias from assignment (◯). The relation between $\mathbb{B}^0$ and $\mathbb{Y}^1$ is still subject to biases incurred by the assignment of counselors and texters. As discussed in Section 3.1, we address this problem by controling on shift time, our observed selection variable. For each counselor J and shift s, we compute a shift-specific outcome propensity $\mathbb{Y}_s^{J,1}$ (again over split 1). For counselors J and K, we then compute the difference in outcome propensity over each shift they coincide on, $\mathbb{D}^1(\tau^J,\tau^K \mid S=s) = \mathbb{Y}_s^{J,1} - \mathbb{Y}_s^{K,1}$. To aggregate across shifts, we take $\mathbb{D}^1(\tau^J,\tau^K)$ as the average of $\mathbb{D}^1(\tau^J,\tau^K \mid S=s)$ weighted by the number of conversations taken by the least-active of the two counselors within each shift. Finally, we compute Kendall's tau between outcome differences from split 1 and behavioral patterns from split 0, shown in Figure 5 as ◯.[15]

[15]Our findings are qualitatively similar if we enforce that for each pair of counselors J and K considered in our measurement of the shift-controlled Kendall's tau statistic, the number of conversations they each take during shifts they are both in exceeds some minimum number; for the rating outcome, this minimum pertains to the number of rated conversations they

Having addressed the two sources of bias, the values represented as ◯ thus correspond to an unbiased estimate of the allocation effect of the conversational tendencies we've examined.

Results. In comparing different counselor-level approaches, we highlight the two sources of bias that are in play, and show how addressing them can moderate our understanding of how different tendencies and outcomes are related. This is depicted in Figure 5 as ◯s which are hollow—indicating no statistical significance—or closer to the vertical line (at Kendall's tau = 0) than corresponding △s or □s—indicating that the latter two approaches overestimated the effect size. For example, the large counselor-level effects of conversation length on closure (△) diminish drastically after addressing interactional bias (□), showing that length tautologically reflects closure. Further addressing the temporally-mediated assignment bias (◯) shows that this tendency does not have a significant effect on closure; the decreased effect size also suggests that the previously-observed relation may also have been contingent on shift time, echoing the across-shift variations in conversation length depicted in Figure 4.

These results distinguish between tendencies that are potentially useful in guiding allocation policies, and those that are correlated to outcome by way of circumstantial factors. While these correlations suggest that such tendencies could be highly informative of a conversation's circumstances, they do not translate to recommendations for how the platform should allocate counselors. For example, while conversations in which a counselor's utterances exhibited more positive sentiment resulted in better outcomes, this is largely due to circumstantial and interactional effects, perhaps reflecting texters who are easier to help, a priori any interaction. As such, we should not expect that allocating counselors with a tendency to use more positive language to a conversation will increase the likelihood of receiving a positive rating or closing properly. In contrast, allocating more conversations to counselors who tend to write longer messages or to better echo the texters (higher similarity) may be more promising in improving ratings.

4.3 Simulated experiment: Estimating the effects of an allocation policy

Having demonstrated how our framework can be used to probe the causal nature of the relation between behavioral tendencies and conversational outcomes, we now estimate the potential impact of an allocation policy that assigns counselors to conversations based on these tendencies. We do this by *simulating* an implementation of the allocation policy under idealized conditions, following prior work in the medical domain [31]. We note that this simulated experiment is only meant to serve as a feasibility check that could precede (but not stand in for) more costly real-world experimentation on the actual platform (e.g., via randomized controlled trials [3, 54]).

In practice, to implement and test a tendency-based allocation policy, the platform could proceed in two steps. First, it would identify counselors who exhibit conversational tendencies with a positive allocation effect on an outcome, using our framework to ensure that measurements of these effects aren't biased by the inference challenges we've described. We simulate this step by translating the empirical approach detailed in the previous section to train a *predictive model* that can then be used to identify counselors who are likely to be effective in future conversations (§4.3.1).

Second the platform would allocate counselors to incoming texters based on their predicted effectiveness, comparing the resultant outcomes to a control condition (e.g., outcomes in a random subset of shifts where counselors are still allocated randomly). We coarsely simulate this step by estimating the effects of a counterfactual re-allocation of counselors within each shift, based on their predicted efficiency (§4.3.2).

each take in such shifts. Choosing smaller thresholds potentially incurs noisier measurements of outcome propensity, while more restrictive cut-offs result in less statistical power; in the case of rating, our analyses would be limited to a potentially skewed sample of shifts and of counselors where enough ratings are obtained.

4.3.1 Training a predictive model. We first identify potentially effective counselors in terms of the rating and closure outcomes. Here, we translate the empirical approach from Section 4.2 into the setup of a predictive task; for the purposes of demonstration, we use the simple conversational tendencies from the preceding analysis as features, noting that future work could naturally consider richer representations of tendencies.

Task setup. Crucially, in modeling counselor effectiveness, we must ensure that the assignment and interaction-based inference challenges are both addressed. To do so for multiple tendencies at once, we integrate our inference solutions into the setup of a prediction task: given a counselor's behavioral tendencies exhibited in their first 40 conversations—i.e., their *past behavior*—we wish to predict their propensity to receive good ratings or close conversations in their next 40 conversations—i.e., their *future outcomes*. In other words, we *split* counselors' conversations into past and future conversations, relating tendencies estimated from one split to outcome propensities estimated from the other and thus mitigating bias from interaction.

To additionally address bias from assignment we formulate this prediction task as a *paired* task that matches counselors on our selection variable, shift. Concretely, for each pair of counselors that have future conversations in a given shift, our goal is to guess which counselor will get a higher proportion of good outcomes in that shift.

We divide the population of counselors into a train and test set comprising 50% of counselors each. We train the model for texter rating over 10,118 pairs spanning 280 shifts; for closure we train on 55,473 pairs across 329 shifts. We use SVM models with 10-fold cross-validation.

Model performance. Before applying these models to our simulated experiment, we report their performance for reference. The relative test set accuracy of the trained models corroborates the effect sizes we observed following our controlled approach (denoted by ◯ in Figure 5). In predicting *future* rating, features based on the minimalistic set of tendencies we used outperform a random (50%) baseline (Bonferroni-corrected binomial test $p < 0.001$) with an accuracy of 56.6%. In predicting future closure rate, these features get a lower accuracy of 50.8%—the poorer performance suggests that the apparently strong naive correlations (corresponding to △ in Figure 5, right) are contingent on the circumstantial confounds that our task setup addresses.[16] Therefore, we expect behavioral tendencies to be less informative in improving the closure outcome via an assignment policy, than in improving the rating outcome.

4.3.2 Simulated re-allocation. We now use observational data to *simulate* an allocation policy that assigns counselors that are predicted to be more effective—based on their behavioral tendencies exhibited in *past* conversations—to more *future* conversations in their shifts. Estimating the impact of this policy requires addressing the same counterfactual question introduced in Section 3—given counselors with different tendencies, what is the effect of allocating one counselor to a conversation, versus the other? Here, we adapt a procedure from prior work that simulates the effectiveness of policies in a medical domain [31].

Simplifying assumptions. Before detailing our simulated policy, we note that it draws on the key assumption that within-shift re-allocations are feasible. In practice, the logistical and ethical aspects of this assumption would need to be carefully considered: such a policy would hinge on the availability and willingness of the counselors, their capacity for taking more conversations, and the potentially detrimental burden they would incur from the extra load.[17] For the purposes of

[16] The relative magnitude of the feature weights learned by each model echo the ranking of the controled assignment effects depicted as ◯ in Figure 5. For instance, in the closure task, conversation length has the smallest weight while response speed has the largest (in spite of similar counselor-level correlations denoted by △).

[17] We do not consider cross-shift reallocations since those would potentially imply requiring counselors to work at times at which they are not available.

this demonstration, this assumption allows us to focus on gauging the allocation effect that we have explored in the preceding sections; as such, we highlight the causal inference problem while leaving important and complementary aspects of the policy to future work.

Counterfactual re-allocation. Given a shift, we focus on the subset of conversations taken by the test-set counselors during their 40th to 80th (i.e., future) conversations. We use our paired prediction models to produce rankings of these counselors based on tendencies observed in their first 40 (i.e., past) conversations, and in separate shifts from their 40th to 80th conversations. We consider a counterfactual scenario in which all conversations in this shift are taken by the top k% of counselors according to these predicted rankings, thus simulating allocating more conversations to them and making use of our assumption that counselors can be re-allocated within a shift.

To estimate the effect of this within-shift re-allocation, we compare the proportion of good outcomes over conversations taken by these predicted top counselors—the *counterfactual outcome*—to the actual proportion—the *realized outcome*. We macro-average these within-shift outcomes across all shifts, so that no single shift has a disproportionate impact on the estimate. To ensure that we have enough data to provide clear estimates, the statistics we subsequently report are taken over shifts with at least six counselors who each take at least three of their 40th to 80th conversations in that shift; for the rating outcome we further enforce that each of these conversations receives a rating from the texter. As such, we consider 55 shifts for rating and 138 for closure.[18]

Estimated effects. Figure 6 shows counterfactual outcomes, macroaveraged by shift, for different k (◯), compared to the realized outcome (dashed line). For texter rating, each counterfactual outcome improves upon a realized outcome. Pairing on shift, these differences are significant for each k (Wilcoxon $p < 0.01$, indicated as filled-in ●), suggesting that the counterfactual improvements occur across many different shifts: concretely, at $k = 25\%$, the counterfactual outcome improves over the realized one in 74% of shifts. This suggests that there is some promise in allocation policies that are informed by conversational tendencies, and could motivate more involved studies, such as those deploying experiments that more actively intervene in the platform.

For the closure outcome, the counterfactual scenario does not improve significantly over the realized one (Wilcoxon $p \geq 0.01$ for each, indicated as ◯). This further reinforces that the strong relationships between tendency and closure reflected in naive approaches (△ and □ in Figure 5), which do not have a causal interpretation, cannot usefully guide allocation policies. As such, more involved experimentation may be unwarranted.

4.3.3 Comparison to non-conversational information. As an alternative to gauging the effectiveness of conversational tendencies, the platform may wish to rely on other information that might more directly relate to outcomes. In particular, a counselor's **past outcomes**—i.e., their propensity to elicit positive ratings or to close conversations, as computed over their first 40 conversations—could be a strong signal of their future effectiveness. Here, we briefly evaluate the utility of these non-conversational signals in guiding the re-allocation.

Results. Following the same simulation procedure as before, we start by identifying potentially effective counselors on the basis of past outcomes. For rating, a predictive model based on past ratings gets an accuracy of 53.1%—while this outperforms the random baseline, it underperforms relative to the model trained on tendency (Binomial $p < 0.001$ for both). This suggests that past outcomes are indicative of future performance (having accounted for our inference challenges

[18] Small modifications to these parameters yield qualitatively similar results. In choosing these parameters, we acknowledge some tradeoffs: lowering these thresholds incurs some noise—concretely, the counterfactual performances would be taken over a fewer number of counselors and conversations, increasing the chance that the estimates are skewed by exceptionally good or bad predictions or outcomes. Raising these thresholds decreases the number of shifts considered, resulting in a loss of statistical power and potentially constraining the representativeness of the findings.

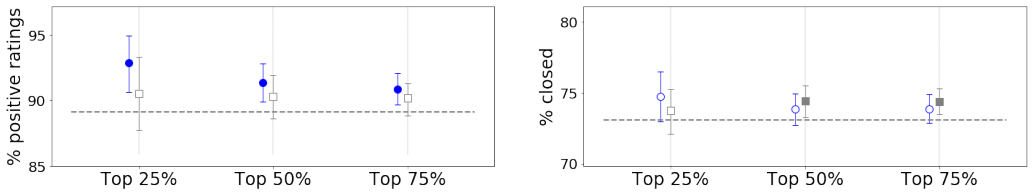

Fig. 6. Proportions of positive ratings (left) or closed conversations (right), macroaveraged over shifts, in a simulated counterfactual setting where the platform selects the top k% counselors on the basis of their future performance as predicted using past conversational tendencies ◯, or based on historical outcomes □. Error bars show bootstrapped 95% confidence intervals. The dashed line denotes the actual realized proportion of positive ratings or proper closures observed in the data. Shapes are filled in for Bonferroni-corrected Wilcoxon $p < 0.01$, comparing counterfactual and realized outcomes and pairing on shift.

in the prediction task setup). Nonetheless, the conversational tendencies we considered—while simplistic—are more informative than non-conversational indicators; the counselors' behavioral tendencies may be less noisy and more robust across circumstances than the past ratings they receive from different texters, especially given the low response rate.

Reflecting the relatively low prediction accuracies, the counterfactual outcomes from accounting for past ratings do not significantly improve upon realized outcomes (Wilcoxon $p \geq 0.01$, results macroaveraged over shift shown in Figure 6 left, as □). At each k, these outcomes are lower than the counterfactual outcomes attained using the tendency-based approach (comparing relative heights of the ● and □ markers), though none of these differences are statistically significant (Wilcoxon $p \geq 0.01$). This shows that there is some promise to going beyond non-conversational information, while motivating the need for richer conversational signals that could be more informative.

For closure, a predictive model based on a counselor's past ability to close conversations slightly outperforms the model using tendencies, with an accuracy of 51.6% ($p < 0.001$). We find that at $k = 50\%$ and 75%, the counterfactual outcomes from using past closure propensities do significantly exceed realized outcomes (Wilcoxon $p < 0.01$, Figure 6 right, indicated as ■).[19]

5 DISCUSSION

In this work, we considered the problem of translating observed relationships between conversational behaviors and outcomes into actionable insights. Through examining a particular policy—allocating agents in conversational platforms—that such observational analyses could inform, we formally described the inference task and inherent challenges involved, translating causal inference arguments to the domain of conversations. In the context of crisis counseling, we demonstrated the importance of properly addressing these challenges, but also the potential benefits of policies that are informed by careful analyses of conversations. Here, we describe how our particular work informs broader studies of computer-mediated settings where conversations play a central role. We also highlight some empirical and theoretical limitations that future work could take up.

The bulk of our work examines one type of conversational setting—goal-oriented asymmetric conversation platforms—and one type of policy—allocating agents. This focus enables us to develop theoretical descriptions that clearly highlight key aspects of the inference task—the relationships between behavior, circumstance, and outcome; it is also grounded in a socially consequential real-world setting, crisis counseling. We note, however, that this inference task is relevant to a broader range of conversational settings: like team discussions or public forms, where conversational roles

[19]We attribute the lower counterfactual estimate at $k = 25\%$ to noise, since fewer counselors are re-allocated at this setting.

may be more fluid and extend beyond that of agent and client, while different participants may have different goals in the conversation. A platform could also pursue a range of other policies, such as *training* effective conversational behaviors or deterring detrimental ones via practices like moderation. These related settings and policies inherit the challenges we have described: the underlying problems of causally relating behaviors and outcomes, and of addressing the inference challenges from assignment and interaction, continue to be salient beyond our particular focus.

We see our work as critically examining one important part of developing policies to improve conversational platforms: translating descriptive findings based on observational data into prescriptive information. Designing and implementing these policies requires a wealth of additional research. For instance, other studies could examine more intricate conversational behaviors and tendencies, viewing the rich types of characteristics explored in many past descriptive studies through a causal lens. A more nuanced understanding of conversational outcomes is needed to gauge the effectiveness of any policy. For example, in this setting, a complementary line of work could develop more informative ways to elicit texter feedback, seeking to better understand and mitigate the presently low response rate; understanding longer-term impacts of a counseling conversation would also be valuable.

Numerous factors relating to the implementation of a policy would need to be accounted for as well; as noted earlier, the estimated benefits of allocating more conversations to counselors must be weighed against the potential for the additional workload to strain the counselors' mental health and conversational effectiveness. Addressing these aspects is beyond the scope of our work, and we look to other studies of computer-mediated communication platforms for promising approaches [30, 32]. However, we emphasize that the core problem of measuring the causal effects of a policy is salient regardless of the extent to which other components of the policy are well-developed—such that identifying these effects, while not sufficient, is necessary in informing these policies.

5.1 Limitations

Our present study is subject to two broad types of limitations, relating to the extent to which our conceptual description is a good model of real-world communication platforms like the counseling service, and to aspects of the model that could be extended to encompass a broader range of settings.

Empirical limitations. Throughout our paper, we've argued that the type of conversation platform we theoretically examined, and the assumptions necessary to mitigate inference challenges, are represented in the crisis counseling setting and are realistic across a broader range of domains. As we've noted, we must inevitably make some assumptions about the nature of the counseling platform. Concretely, we supposed for sake of demonstration that assignment is random within shift and that dependencies do not exist between different conversations taken by a counselor. In practice, the platform's assignment procedure prioritizes clear cases of suicidal ideation, and more experienced counselors can take multiple conversations at once, which may induce cross-conversational dependencies. Future work could better account for factors such as these, that exist across other conversational settings. In particular, these efforts could investigate the extent to which the solutions we propose are sensitive to these exceptions, and could relax the assumptions that these exceptions challenge.

Theoretical limitations. Our theoretical formulation—of a particular inference problem in a particular type of conversational setting—could be relaxed or extended in several ways. In the purview of this subproblem, the solution we propose to mitigate the bias from assignment requires us to control for a selector variable such as shift, and hence restricts us from making comparisons or agent allocations across shifts, or from applying our approach to settings where selector variables are not fully observable. Future work could consider ways to relax this requirement, perhaps by using parametric models of the relation between shift, behavior and outcome.

We have examined a simple formulation of an allocation policy: discovering and hence allocating more conversations to the most effective agents within a shift. Future work could examine more sophisticated allocation policies, and the more complex causal inferences required to motivate them. For instance, finer-grained models of tendency and circumstance could point to policies that match agents with situations they are particularly well-suited to address, given their behavioral tendencies. A complementary line of work could more rigorously interrogate the assumption at the heart of our simulated re-allocation (§4.3.2) by accounting for the impact of increasing agents' loads, or by examining policies that are less contingent on this assumption.

Our methodology could also be extended to encompass a broader range of conversational paradigms beyond asymmetric settings, such those involving discussions in a team [6, 12, 22, 65]. Such extensions would need to account for additional properties of the domain, such as a more diverse and dynamic range of conversational roles, over which the platform might have varying degrees of leverage. In such contexts, the attributes of the individuals involved as gleaned from indicators such as personality types or demographic information, as well as how these attributes are combined, has been experimentally shown to potentially impact the effectiveness of a discussion [7, 66]. Here, it would be interesting to see how conversational tendencies could be used in concert with these non-interactional labels.

Our work is crucially limited to addressing the problem of agent-level allocation. As such, we've examined a coarser policy than that of training agents to adopt particular behaviors. Future work could take up the corresponding inference task: estimating how a change in an agent's behavior, once they are assigned to a conversation, affects the conversation's outcome. Intuitively, this task is more challenging to address: as previously noted, taking agent-level aggregates allows us to abstract away from the circumstances within a conversation in our analyses; in the training paradigm, circumstantial factors could be even more intricately entangled with behavior and outcome.

Our work addresses two key features of analyzing conversational data: that this data is often observational, and that it contains complex interactional dynamics. However, we leave an additional pillar of this setting open: conversations are *linguistic*. As such, the behavioral signals we glean from the data are necessarily low-dimensional representations [18, 28, 63] of abstract and perhaps more consequential conversational qualities. For instance, our formulation allows us to reason about the effect of allocating agents, and in the counseling setting we show that verbosity is a good *signal* of a counselor's effectiveness. However, unilaterally instructing to counselors that they increase the number of words they use may be ineffectual, if verbosity is simply a proxy for a tendency that is less straightforward to model, such as eloquence. This present limitation further constrains our ability to translate inferences we've made to policies such as training, and to make finer-grained statements about conversational behaviors and their impacts. Thus, there is ample opportunity for future work to address this gap, by way of more nuanced examinations of conversational behaviors and of their causal effects.

ACKNOWLEDGMENTS

The authors would like to thank the participants at the New Directions in Analyzing Text as Data conference (Fall 2019) and the Interdisciplinary Seminar in Quantitative Methods at the University of Michigan (Fall 2019), as well as Jonathan P. Chang, Caleb Chiam, Liye Fu and Lillian Lee for helpful discussions. This research would not have been possible without the support of Crisis Text Line, and we are particularly grateful to Robert Filbin, Christine Morrison, and Jaclyn Weiser for their valuable insights into the platform and for their help with using the data. The research has been supported in part by NSF CAREER Award IIS1750615 and a Microsoft Research PhD Fellowship. The collaboration with Crisis Text Line was supported by the Robert Wood Johnson Foundation; the views expressed here do not necessarily reflect the views of the foundation.

REFERENCES

[1] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-Scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* (2016).

[2] Allen D. Andrade, Anita Bagri, Khin Zaw, Bernard A. Roos, and Jorge G. Ruiz. 2010. Avatar-Mediated Training in the Delivery of Bad News in a Virtual World. *Journal of Palliative Medicine* (2010).

[3] Joshua D. Angrist and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

[4] Joshua D. Angrist and Jörn-Steffen Pischke. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives* (2010).

[5] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to Me: Foundations for Successful Individual-Group Interactions in Online Communities. In *Proceedings of CHI*.

[6] Markus Baer, Greg R. Oldham, Gwendolyn Costa Jacobsohn, and Andrea B. Hollingshead. 2008. The Personality Composition of Teams and Creativity: The Moderating Role of Team Creative Confidence. *The Journal of Creative Behavior* (2008).

[7] Julia Bear and Anita Woolley. 2011. The Role of Gender in Team Collaboration and Performance. *Interdisciplinary Science Reviews* (2011).

[8] Carlos Brito and Judea Pearl. 2012. Generalized Instrumental Variables. In *Proceedings of UAI*.

[9] Moira Burke, Elisabeth Joyce, Tackjin Kim, Vivek Anand, and Robert Kraut. 2007. Introductions and Requests: Rhetorical Strategies That Elicit Response in Online Communities. In *Communities and Technologies 2007*, Charles Steinfield, Brian T. Pentland, Mark Ackerman, and Noshir Contractor (Eds.). Springer.

[10] Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of ACL*.

[11] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. In *Proceedings of CHI*.

[12] Ziqiang Cheng, Yang Yang, Chenhao Tan, Denny Cheng, Yueting Zhuang, and Alex Cheng. 2019. What Makes a Good Team? A Large-Scale Study on the Effect of Team Composition in Honor of Kings. In *Proceedings of WWW*.

[13] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. In *Proceedings of CHI*.

[14] Munmun De Choudhury and Emre Kıcıman. 2017. The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk. In *Proceedings of ICWSM*.

[15] Justin Cranshaw and Aniket Kittur. 2011. The Polymath Project: Lessons from a Successful Online Collaboration in Mathematics. In *Proceedings of CHI*.

[16] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of WWW*.

[17] Orianna DeMasi, Marti A. Hearst, and Benjamin Recht. 2019. Towards Augmenting Crisis Counselor Training by Improving Message Retrieval. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology* (2019).

[18] Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. How to Make Causal Inferences Using Texts. (2018).

[19] Jane E. Fountain. 2003. Prospects for Improving the Regulatory Process Using E-Rulemaking. *Commun. ACM* (2003).

[20] Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. 1995. Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology* (1995).

[21] Shane Haberstroh, Thelma Duffey, Marcheta Evans, Robert Gee, and Heather Trepal. 2007. The Experience of Online Counseling. *Journal of Mental Health Counseling* (2007).

[22] Randall S. Hansen. 2006. Benefits and Problems with Student Teams: Suggestions for Improving Team Projects. *Journal of Education for Business* (2006).

[23] Clara E. Hill and Emilie Y. Nakayama. 2000. Client-centered Therapy: Where Has It Been and Where Is It Going? A Comment on Hathaway (1948). *Journal of Clinical Psychology* (2000).

[24] Yuheng Hu, Ali Tafti, and David Gal. 2019. Read This, Please? The Role of Politeness in Customer Service Engagement on Social Media. In *Proceedings of HICSS*.

[25] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of ICWSM*.

[26] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. In *Proceedings of CSCW*.

[27] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the Crowd: What Do People React to in Online Discussions?. In *Proceedings of EMNLP*.

[28] Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *Proceedings of ACL*.

[29] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W. Malone. 2017. What Makes a Strong Team?: Using Collective Intelligence to Predict Team Performance in League of Legends. In *Proceedings of CSCW*.

[30] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of CSCW*.

[31] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction Policy Problems. *American Economic Review* (2015).

[32] Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.

[33] Walter S. Lasecki, Samuel C. White, Kyle I. Murray, and Jeffrey P. Bigham. 2012. Crowd Memory: Learning in the Collective. In *Proceedings of CHI*.

[34] Kurt Luther, Kelly Caine, Kevin Ziegler, and Amy Bruckman. 2010. Why It Works (When It Works): Success Factors in Online Creative Collaboration. In *Proceedings of GROUP*.

[35] I. Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven Dow. 2016. Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams. In *Proceedings of CSCW*.

[36] Keith Maki, Michael Yoder, Yohan Jo, and Carolyn Rosé. 2017. Roles and Success in Wikipedia Talk Pages: Identifying Latent Patterns of Behavior. In *Proceedings of EMNLP*.

[37] Brian McInnis, Dan Cosley, Eric Baumer, and Gilly Leshed. 2018. Effects of Comment Curation and Opposition on Coherence in Online Policy Discussion. In *Proceedings of GROUP*.

[38] Brian L. Mishara, François Chagnon, Marc S. Daigle, Bogdan Balan, Sylvaine Raymond, Isabelle Marcoux, Cécile Bardon, Julie K. Campbell, and Alan D. Berman. 2007. Which Helper Behaviors and Intervention Styles Are Related to Better Short-Term Outcomes in Telephone Crisis Intervention? Results from a Silent Monitoring Study of Calls to the U.S. 1-800-SUICIDE Network. *Suicide & Life-Threatening Behavior* (2007).

[39] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *Proceedings of NAACL*.

[40] John C. Norcross and Michael J. Lambert. 2018. Psychotherapy Relationships That Work III. *Psychotherapy* (2018).

[41] Grant Packard, Sarah G. Moore, and Brent McFerran. 2018. (I'm) Happy to Help (You): The Impact of Personal Pronoun Use in Customer–Firm Interactions. *Journal of Marketing Research* (2018).

[42] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. Emojis on Twitter: A Causal Inference Approach. In *Proceedings of OSSM*.

[43] Judea Pearl. 1995. Causal Diagrams for Empirical Research. *Biometrika* (1995).

[44] Judea Pearl. 2013. On the Testability of Causal Models with Latent and Instrumental Variables. In *Proceedings of UAI*.

[45] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology* (2003).

[46] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and Predicting Empathic Behavior in Counseling Therapy. In *Proceedings of ACL*.

[47] Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the Quality of Counseling Conversations: The Tell-Tale Signs of High-Quality Counseling. In *Proceedings of LREC*.

[48] Anthony R. Pisani, Nitya Kanuri, Bob Filbin, Carlos Gallo, Madelyn Gould, Lisa S. Lehmann, Robert Levine, John E. Marcotte, Brian Pascal, David Rousseau, Shairi Turner, Shirley Yen, and Megan L. Ranney. 2019. Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line. *Journal of Medical Internet Research* (2019).

[49] Tom Pyszczynski, Kathleen Holt, and Jeff Greenberg. 1987. Depression, Self-Focused Attention, and Expectancies for Positive and Negative Future Life Events for Self and Others. *Journal of Personality and Social Psychology* (1987).

[50] Derek Richards and Ladislav Timulak. 2012. Client-Identified Helpful and Hindering Events in Therapist-Delivered vs. Self-Administered Online Cognitive-Behavioural Treatments for Depression in College Students. *Counselling Psychology Quarterly* (2012).

[51] Carl R. Rogers. 1957. The Necessary and Sufficient Conditions of Therapeutic Personality Change. *Journal of Consulting Psychology* (1957).

[52] Stephen Rollnick and William R. Miller. 1995. What Is Motivational Interviewing? *Behavioural and Cognitive Psychotherapy* (1995).

[53] Paul R. Rosenbaum. 2010. *Design of Observational Studies*. Springer.

[54] Donald B. Rubin. 2007. The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. *Stat. Med.* (2007).

[55] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *Proceeding of ICWSM*.
[56] Jessica Schroeder, Jina Suh, Chelsey Wilks, Mary Czerwinski, Sean A. Munson, James Fogarty, and Tim Althoff. 2020. Data-Driven Implications for Translating Evidence-Based Psychotherapies into Technology-Delivered Interventions. In *Proceedings of PervasiveHealth*.
[57] Craig S. Schwalbe, Hans Y. Oh, and Allen Zweben. 2014. Sustaining Motivational Interviewing: A Meta-Analysis of Training Studies. *Addiction* (2014).
[58] Eva Sharma and Munmun De Choudhury. 2018. Mental Health Support and Its Relationship to Linguistic Accommodation in Online Communities. In *Proceedings of CHI*.
[59] Dhanya Sridhar and Lise Getoor. 2019. Estimating Causal Effects of Tone in Online Debates. In *Proceedings of IJCAI*.
[60] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. In *Proceedings of WWW*.
[61] Terence J. G. Tracey, Bruce E. Wampold, James W. Lichtenberg, and Rodney K. Goodyear. 2014. Expertise in Psychotherapy: An Elusive Goal? *The American Psychologist* (2014).
[62] Yi-chia Wang, Robert Kraut, and John M. Levine. 2012. To Stay or Leave? The Relationship of Emotional and Informational Support to Commitment in Online Health Support Groups. In *Proceeding of CSCW*.
[63] Zhao Wang and Aron Culotta. 2019. When Do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception Using Individual Treatment Effect Estimation. In *Proceedings of AAAI*.
[64] Harry Weger, Gina R. Castle, and Melissa C. Emmett. 2010. Active Listening in Peer Interviews: The Influence of Message Paraphrasing on Perceptions of Listening Skill. *International Journal of Listening* (2010).
[65] Mark E. Whiting, Allie Blaising, Chloe Barreau, Laura Fiuza, Nik Marda, Melissa Valentine, and Michael S. Bernstein. 2019. Did It Have To End This Way?: Understanding The Consistency of Team Fracture. In *Proceedings of CHI*.
[66] Anita Woolley and Thomas Malone. 2011. What Makes a Team Smarter? More Women. *Harvard Business Review* (2011).
[67] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia. In *Proceedings of ICWSM*.
[68] Diyi Yang and Robert E. Kraut. 2017. Persuading Teammates to Give: Systematic versus Heuristic Cues for Soliciting Loans. In *Proceedings of CSCW*.
[69] Diyi Yang, Miaomiao Wen, and Carolyn Penstein Rosé. 2015. Weakly Supervised Role Identification in Teamwork Interactions. In *Proceedings of ACL*.
[70] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.
[71] Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proceedings of ACL*.
[72] Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding Your Voice: The Linguistic Development of Mental Health Counselors. In *Proceedings of ACL*.

Received January 2020; revised June 2020; accepted July 2020