Memory & Cognition (2022) 50:407–424
https://doi.org/10.3758/s13421-021-01208-8

Memory for syntactic differences in mental illness descriptions

Emily N. Line[1] · Samantha Roberts[2] · Zachary Horne[3]

Accepted: 6 July 2021 / Published online: 7 September 2021

Abstract

The American Psychiatric Association recommends that practitioners discuss mental illnesses using person-first, or comparatively *state-based* language, rather than trait-based language. The aim of this initiative is to both avoid treating the symptoms of an illness as a defining characteristic of the people who experience these symptoms and to reduce the stigmatization of mental illness. However, some of the implications of these initiatives have not been tested. Here, we investigate one of these implications—people's memory for changes in syntactic constructions in descriptions of mental illness. In three experiments, we observed that people form similar representations of state- and trait-based passages as reflected by their performance in two recognition tasks and a free-recall task. However, a fourth experiment suggested that participants' memories of the exact syntax they read are not so degraded that they are unable to recover what they read when explicitly prompted. Altogether, these results suggest that some aspects of the person-first language initiative are likely to be transient.

Keywords State-based language · Person-first language · Memory · Stigma

Introduction

Nineteen percent of people in the United States report having some kind of mental illness in a given year (Substance Abuse and Mental Health Services Administration, 2018) and the negative effects of mental illnesses are exacerbated by the pervasive stigma surrounding them (e.g., Link, Phelan, Bresnahan, Stueve, & Pescosolido, 1999). One source of mental health stigma is the use of certain kinds of language (e.g., the slur "crazy"), and even more subtly, certain syntactic constructions ("has depression" vs. "is experiencing the symptoms of depression"). The emerging literature on the impact of syntactic constructions on the perception of mental illness, which we detail below, has led to calls for the use of so-called person-first language. The person-first language initiative emphasizes that practitioners should focus on the patients themselves rather than their illness by avoiding labels corresponding to their illness. For example, practitioners should describe a patient as "having schizophrenia," rather than as being "schizophrenic". Person-first language and related language-based initiatives have been widely studied, but it is unknown what the long-lasting cognitive impact of person-first language is—to what extent does shifting syntactic constructions in discussing mental health affect our episodic memories of a person experiencing the symptoms of mental illness? Here, we examine this question in four preregistered experiments. We argue that although person-first language may yield more positive self-reported attitudes towards people experiencing mental health issues, aspects of these effects may be transient. People may interpret person-first constructions as nonetheless indicative of a mental illness.

A key aim of person-first language is to reconstruct descriptions of mental illness by changing their syntactic structure so that they do not identify a person with their mental illness. The person-first language initiative has been widely discussed and adopted in clinical psychology, but the (assumed) cognitive underpinnings have received less attention. In particular, we propose that foundational research on episodic memory and language comprehension can also shed light on the merits—and potential problems—with this initiative.

It has been established that memory for the meaning of an event (e.g., a sentence, passage) tends to be more accurate

✉ Emily N. Line
neuline2@illinois.edu

[1] University of Illinois, Urbana-Champaign, Champaign, IL, USA

[2] Arizona State University, Phoenix, AZ, USA

[3] University of Edinburgh, Edinburgh, UK

and longer lasting than memory for the structure of that event (e.g., the syntactic form of sentences in a passage: Brewer, 1977; Deese, 1959; Loess, 1967; Loftus, Miller, & Burns, 1978; Roediger & McDermott, 1995; Sachs, 1967; Sulin & Dooling, 1974). For instance, when asked to recall a passage of text after a delay, people's memories tend to reflect their interpretations of a passage rather than the exact sentences they read (Bransford & Franks, 1971). Van Dijk and Kintsch (1983) provide a framework for understanding these effects. Under this framework, representations for passages can be conceived of as occurring at three levels: the surface level, the text or proposition level, and the situation model (see also Fletcher & Chrysler, 1990; Fletcher, 1994, for further expansion on the topic). The surface level representation refers to the structure of a sentence, the proposition level concerns the relationships between arguments and predicates, and the situation model is a reader's interpretation of the text. Importantly, each level does not affect memory in the same way, especially after a delay (Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Brewer, 1977; Deese, 1959; Loess, 1967; Loftus et al., 1978; Roediger & McDermott, 1995; Sachs, 1967; Sulin & Dooling, 1974).

The applied implications of the intuitive result that memory for the meaning of a text (i.e., the situation model) is more accurate than memory at the surface or proposition level are significant. For example, changing a word that updates the situation model when questioning an eyewitness of an accident can lead to different recollections of the accident (Loftus & Palmer, 1974), but subtle syntactic shifts, those that change the surface aspect of a sentence, do not have correspondingly substantial effects. Along these lines, educational psychologists have found that substituting more familiar words with identical meaning can improve memory for content on chemistry tests, but shifts in syntactic structure (e.g., passive to active voice) do not affect students' performance (Cassels & Johnstone, 1984). Although syntax can be attended to and remembered in some contexts where syntax is particularly important to the topic (e.g., in people's memory for poetry), in everyday prose the situation model dominates what people remember (Tillmann & Dowling, 2007).

To date, clinicians have primarily focused on what person-first language appears to accomplish in immediate settings without considering how robust these shifts will be across time. Key findings regarding surface, proposition, and situation model representations suggest potential limitations in the long-term efficacy of person-first language initiatives. In particular, we expect that, as seen in other applied settings, meaning rather than syntax will dominate what people remember about descriptions of mental illness. To convey the significance of this possibility, we will first review some of the central results and primary motivations for the person-first language initiative.

Stigma and the person-first language initiative

Stigma has a pervasive effect on the way people with mental illnesses are treated, impacting how people seek treatment and care for themselves (Corrigan, Druss, & Perlick, 2014). For example, people with mental illness are seen by others to be more dangerous, even when outside factors do not support these stereotypes (Link, Phelan, Bresnahan, Stueve, & Pescosolido, 1999). Mental illness stigma also impacts the ability to find a job or housing, and can act as a barrier to higher education (Wahl, 1999). Strikingly, over a third of Americans report that they would not interact socially with a person who has depression, schizophrenia, or is substance-dependent (Martin, Pescosolido, & Tuch, 2000). This is notable given that large proportions of the population report having a mental illness (National Institute of Mental Health, 2018). These and similar effects have led medical professionals to seek means for reducing the stigmatization of mental illness.

Medical professionals have pursued several means for reducing the stigma surrounding mental health. For example, education about mental illness and contact with people who have mental illnesses can reduce negative misconceptions (Corrigan et al., 2014; Corrigan, Larson, Sells, Niessen, & Watson, 2006; Papish, Kassam, Modgill, Vaz, Zanussi, & Patten, 2013). Anti-stigma campaigns like England's Time to Change campaign (2019) or the National Alliance on Mental Illness's StigmaFree campaign (2019) have also raised awareness of the prevalence of mental illness (Substance Abuse and Mental Health Services Administration, 2018). The aim of these initiatives is to show that mental health issues are common, thus shifting the perception that people with mental illness should be feared or avoided.

A more recent means for reducing stigmatization is addressing how *labeling* impacts the public's perception of mental illness. How a group is labeled can cause people to draw sharp distinctions between in-group and out-group members, leading to discrimination and loss of status (e.g., Blaska, 1993; Broyles, Binswanger, Jenkins, Finnell, Faseru, Cavaiola, et al., 2014; Granello & Gibbs, 2016; Kelly et al., 2015; Link & Phelan, 2001; Link, 1987). Further, people prefer to be referred to by person-first terms rather than state-based labels (Pivovarova & Stein, 2019). In light of these and related findings, the American Psychiatric Association (APA; 2013) has established mental health language guidelines for clinicians, practitioners, and journalists. Specifically, the APA proposes that one way to reduce the stigma of mental illness is by using person-first language. This initiative has proved influential: institutions including the American Psychological Association's (2010) publication manual now recommend the use of person-first language. The person-first language initiative has even led to the enactment of the People First Respectful Language Modernization Act of 2006, which requires all new and revised laws in the United States to use person-first language when referring to people with mental illnesses and disabilities.

The aims of the person-first language initiative are to reduce the stigmatization of mental illness. This general goal can be broken down into two distinct aims: First, the person-first language initiative is intended to provide a more humanizing experience for patients when practitioners are interacting directly with people experiencing the symptoms of mental illness (Broyles et al., 2014; Kelly, Wakeman, & Saitz, 2015). Second, the initiative is intended to transform the "normative labeling mindset" so that people are identified with something more than just their mental illness. In encouraging person-first language use, researchers and clinicians have suggested that stereotypes and stigma of mental illness will decrease (Blaska, 1993; Granello & Gibbs, 2016). The present research focuses on one aspect of the second aim of this initiative.

Person-first, state-based language, and representation of syntactic structure

The person-first distinction is an example of a more general linguistic distinction—the trait-state distinction. Going forward, we will discuss the person-first language initiative in these terms because it highlights the ways in which this initiative connects to stigma-reducing language initiatives in other areas of psychology. Under this more general framework, we can understand the APA guidelines as a recommendation for using comparatively more *state-based* language (e.g., "Smith is experiencing the symptoms of schizophrenia", rather than *trait-based* language (e.g., "Smith is a schizophrenic"; see Gelman & Heyman, 1999; Link, 1987, for more discussion of this distinction). State- and trait-based language is most appropriately conceived of as a continuous distinction rather than a clean dichotomy. For example, the linguistic construction "Smith is a schizophrenic" is unambiguously trait-based because it identifies Smith with their mental illness ("is a"). In contrast, "Smith is experiencing symptoms of schizophrenia" is a purely state-based construction because it describes a mental illness as something Smith is *currently* experiencing, but hopefully will not always experience. However, other examples of (purportedly) appropriate person-first language provided by the APA also include constructions like, "Smith is a person *with* schizophrenia" (American Psychiatric Association, 2013). This example falls between the two ends of this state-trait continuum because it does not label Smith, but it nonetheless ties a noun to them (see Fig. 1). Under the state-trait framework, trait-based language includes both common nouns (e.g., "schizophrenic") and diagnostic possessive phrases (e.g., "has schizophrenia"), both of which have been shown to lead people to infer stable traits rather than transient states (see Gelman & Heyman, 1999; Horne & Cimpian, 2018, for a related discussion of this distinction).

This literature allows us to better understand the link between the person-first language initiative and similar state-based initiatives in developmental and social psychology. Several studies provide support for the state-based initiative as a means for reducing stigma surrounding mental illness. For example, clinicians and the general population assign different levels of responsibility to people labeled with a common noun term relating to a substance abuse disorder (e.g., is a heroin addict) compared to when someone is described *having* a substance abuse disorder (i.e. using a diagnostic possessive phrase; Kelly & Westerhoff, 2010a; b). Beyond the mental health domain, research with adults and children suggests subtle syntactic cues cause people to think that a behavior or disease is stable and long-lasting. Describing a behavior with trait-based rather than state-based language tends to cause people to think that these diseases are stable traits of the people exhibiting these behaviors (e.g., Gelman & Heyman, 1999; Reynaert & Gelman, 2007). Furthermore, trait-based language is often interpreted as a normative generic statement, which can lead to stereotyping (e.g., Wodak, Leslie, & Rhodes, 2015). Trait-based language has also been shown to increase essentialist thinking—the folk psychological theory that internal essences make things the kind of thing that they are (e.g., Bastian & Haslam, 2006; Gelman & Heyman, 1999; Prentice & Miller, 2007; Rhodes, Leslie, Yee, & Saunders, 2019). Importantly, essentialist thinking might be linked to stigmatization and stereotyping (especially when nouns are used; see Howell & Woolgar, 2013; Howell, Ulan, & Powell, 2014), including the stereotyping of mental illness

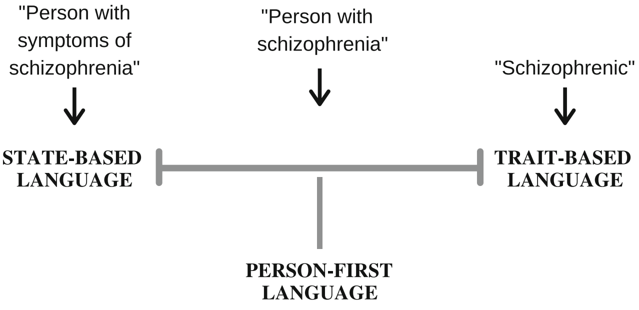

Fig. 1 A simplified depiction of the state-trait language continuum. Person-first language falls between state-based language and trait-based language because it allows for diagnostic possessive phrases, which permit trait-based inferences

(for a discussion of this possibility, see Prentice & Miller, 2007). Together, this research suggests describing an individual with state-based language may lead both clinicians and the public at large to make more appropriate inferences about mental illness, which may reduce stigma surrounding mental illnesses (see Bastian & Haslam, 2006; Bastian & Haslam, 2007; Pauker, Ambady, & Apfelbaum, 2010; Prentice & Miller, 2007; Yzerbyt, Corneille, & Estrada, 2001, for similar findings about race and gender).

Although the state-based language initiative may provide a more humanizing experience for people with mental illness and reduce stigma in inferential contexts, existing research has not determined if there is a lasting cognitive impact of using subtly different language to describe a mental illness. For example, it is unclear how state-based language impacts the way people remember people described using different syntactic constructions. Shifting linguistic constructions may reduce the dehumanization of people that have mental health disorders initially—a clear achievement in itself—but it is nonetheless an open question what the long-lasting cognitive impact of these cues are. Does using this language systematically affect how we remember someone described as experiencing the symptoms of a mental illness?

It is clear that state- and trait-based descriptions differ on the surface and proposition levels (Van Dijk & Kintsch, 1983). Statements which only differ in their surface or even propositional features can be difficult to distinguish, depending on the context and syntactic constructions (see Fletcher, 1994; Bransford & Franks, 1971, for a discussion on variables influencing surface and proposition level memory). This may be particularly likely if a sequence of state-based propositions is generalized to a single statement using the disorder label as a superset. For example, the statements “Bob has flashbacks” and “Bob has nightmares” can be generalized to “Bob has post-traumatic stress disorder” (Van Dijk, 1980; Van Dijk & Kintsch, 1983).

Whether state- and trait-based descriptions differ at the situation level is an open question. For example, consider the phrases “Sally is sad all the time” and “Sally has depressive disorder”. Interpreted very broadly, these two phrases could induce similar situation models—both about Sally and her mental health. However, among practitioners, and arguably the general public, the phrase “Sally is sad all the time” should not *imply* any mental diagnosis. The person-first language initiative distinguishes state- and trait-based language to avoid identifying a person with a disorder label. But this distinction is likely to be inconsistent with key findings on semantic representation (Collins & Loftus, 1975; Helbig, 2006). Still, it is an empirical question whether sensitivity to the state-trait distinction in inferential tasks (i.e., when people are asked to explicitly consider the implications of each statement) are *remembered* in a way that respects the state-trait distinction. Differences in situation model representations are the easiest of the three to recognize or identify in memory tasks (Fletcher & Chrysler, 1990), which would predict that if these descriptions differ at the situation level; people’s memory for state- versus trait-based descriptions would be robust.

The syntactic differences between state- and trait-based language may have further implications for our understanding of memory and representation. Syntactic structure affects our interpretation and comprehension of sentences when forming a model of a sentence (Gernsbacher, 1985; Givón, 1992; Becker, Ferretti, & Madden-Lombardi, 2013; Magliano & Schleich, 2000). For example, syntactic aspects of language can affect the perception of the passage of time and activation of working memory (Becker et al., 2013; Magliano & Schleich, 2000). Magliano and colleagues (2000) find events described with the imperfective aspect of verbs are associated with slower decay rates of working memory activation than events described with the perfective aspect.

Much of the literature regarding the grammatical aspects of language and the situation model look specifically at narratives (Magliano & Schleich, 2000; Becker et al., 2013; Givón, 1992), where the sequence and duration of events can play an important role in the construction of the situation model and working memory. Still, we can evaluate how grammatical differences between state- and trait-language, which only differ at the sentence-level, affect the situation model. Symptoms (state-based) can be described using the imperfective aspect of verbs (e.g., “Bob was having nightmares”, “Sally was feeling sad”) but they can also be described using the perfective aspect (e.g., “Bob had nightmares” and “Sally felt sad”). The use of a disorder label (trait-based) usually requires the use of perfective aspect (e.g., “Bob had post-traumatic stress disorder”, “Sally had depression”, whereas it is less natural to say that “Sally was having depression”). If state-based descriptions use the imperfective aspect of verbs while the corresponding trait-based passages use the perfective aspect of verbs, this could result in different situation models as well.

It should be noted that in the case of discussing illness and symptoms of disorders, regardless of the grammatical aspects of the construction, it seems less likely that people would view symptoms as more “ongoing” than the disorder itself. Someone described as having depression would not lead people to form a situation model of the event that is “complete”. The nature of depression, and the other disorders we use as examples, are often ongoing, even lasting a lifetime.

Together, research on episodic memory and language comprehension suggest that even if clinicians use state-based language, it is possible these relatively subtle syntactic shifts do little to impact people’s recollections of information they've read or heard. This possibility would indicate that using state-based language to reduce the stigma surrounding mental illness may not have lasting effects on how people with symptoms of mental illness are remembered. This possibility would not necessarily negate the immediate benefits of state-based language initiatives, including its impact on how people with mental illnesses feel when interacting with clinicians. However, it would contextualize the size of the impact that these initiatives could plausibly have.

Representational structure

Effects predicted by the person-first language initiative—for example, robust memory for slightly different syntactic constructions—depend on an assumption about the structure of our concepts. Figure 2 depicts two possible representational structures. In Panel A of Fig. 2, the concept DEPRESSION is the parent of three semantic units—symptoms of depression. Under this structure, activation of DEPRESSION excites the semantic units (e.g., the symptoms) connected to the parent concept. For example, if we learn that someone is depressed, we would infer that they might have certain symptoms (e.g., loss of interest). In Panel B of Fig. 2, the structure is quite different because individual semantic units themselves can also activate DEPRESSION. For example, if we learn someone has lost interest in their hobbies and is in a sad mood, we might infer that they have depression. It is well understood that for at least some concepts, semantic units can activate their parents (Collins & Loftus, 1975); in this case, activation of units representing the presence of certain symptoms activate DEPRESSION. However, in inferential (e.g., Horne & Cimpian, 2018) or category formation contexts (e.g., Rhodes, Leslie, & Tworek, 2012), rather than during memory tasks, activation between parent and children may not be symmetrical—people infer that parents entail their children, but children do not necessarily entail their parents. One question then is what state-based language initiatives assume about the structure of concepts *at recall*—to our knowledge, no memory models would predict the uni-directional effects which have been observed in inferential tasks (Panel A Horne & Cimpian, 2018; Rhodes et al., 2012). If saying someone "has depression" versus "is experiencing the symptoms of depression" changes how we remember them, then this suggests a theory of representational structure. Specifically, this would suggest Panel A is a better depiction of at least some of our concepts. In contrast, if the mere mention of symptoms can lead people to remember that a person was described as *having depression*, then this would be more consistent with an undirected cyclic graph (Panel B; Collins & Loftus, 1975).

These theoretical considerations motivated the design of Experiments 1 and 2, and our caution in interpreting the results of Experiment 4 (discussed further below). Under both graphs, activation from trait concepts (i.e., the parent unit DEPRESSION) flows to activation of state concepts (i.e., units representing symptoms of depression). Namely, if someone has an anxiety disorder, we realize they will experience the symptoms of this disorder (see Fig. 2). All else being equal, it has been established that strong syntactic entailments of this sort are very likely to induce memory errors (Powell, Horne, Pinillos, & Holyoak, 2015; Gentner, 1981). Consequently, if participants incorrectly remembered that the passage said, "Hayden was experiencing the symptoms of generalized anxiety disorder" when in the trait-based language condition, it is possible this be could be because of the semantic entailment between traits and states, which would be an altogether unsurprising relative to the finding that participants remember a stronger assertion (e.g., "Hayden has generalized anxiety disorder") when only a weaker assertion is made ("Hayden is anxious all the time"). This consideration led us to design Experiments 1 and 2 focusing on whether people remember that a trait statement was made in the passage they read—do participants make a stronger inference about the protagonist in a story than what they read?[1]

Before proceeding, we first sought to replicate the entailment effects observed in other domains (Horne & Cimpian, 2018; Rhodes et al., 2012), effects which have only been indirectly tested in the literature on person-first language. Ensuring trait-to-state entailment effects replicate in an inferential task allows us to rule out alternative explanations for the possibility that people form the same situation model of trait- and state-based descriptions. For example, perhaps the person-first language distinction is illusory and thus, of course, people recall state-based descriptions as trait-based descriptions.

Testing prior inferential trait-state entailment effects

Participants Participants were 526 Mechanical Turk workers who participated in a short task after completing the tasks described in Experiment 4.[2] They were paid $2.00 for their participation in both tasks.

Materials and procedure

Following prior work (Gelman & Heyman, 1999; Horne & Cimpian, 2018), participants were presented with if–then

[1] We note that the designs of Experiments 3 and 4 do not depend on this assumption.

[2] The tasks participants completed for Experiment 4 are not directly related to the task for this study, nor were any instructions or information from the Experiment 4 task necessary to complete this task.

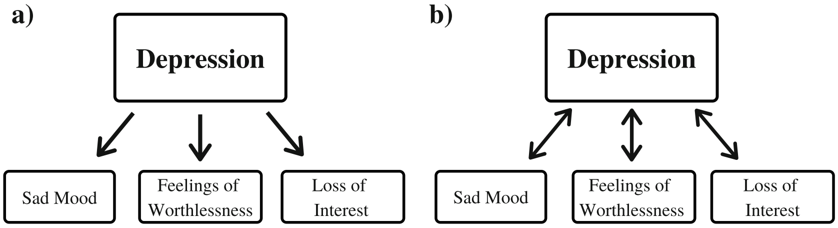

Fig. 2 Possible representational structures of the concept DEPRESSION. a) A directed acyclic graph in which activating the parent, DEPRESSION, excites its children, the symptoms of DEPRESSION. b) An undirected cyclic graph in which activation of semantic units (e.g., symptoms) can excite the unit DEPRESSION

statements in random order to examine the perceived entailment relationships between trait and state descriptions. For example, participants were presented with the statements, “If you have a depressive disorder, then you will frequently feel unhappy” (Condition = disorder entails symptoms) and “If you frequently feel unhappy, then you have a depressive disorder” (Condition = symptoms entail disorder). There were two entailments for four disorders (two entailment statements × four disorders = eight total statements). After each statement, they judged how much they agreed with the statement on a six-point Likert scale (1 = Strongly Disagree, 6 = Strongly Agree).

Results

We predicted that participants would agree with disorder-entails-symptoms statements more strongly than symptoms-entail-disorder statements. Figure 3 shows the proportion of responses at each Likert point for both statement types. Because of the ordinal nature of our data, we fit a cumulative mixed-effects model and found that participants were more likely to judge that someone with depression (i.e., a trait) was more likely to experience the symptoms of depression than that someone experiencing the symptoms of depression (i.e., state) had the disorder, $b = 0.79$, 95% CI [0.63 to 0.95], odds ratio = 2.20. Details about this model are located in the Appendix.

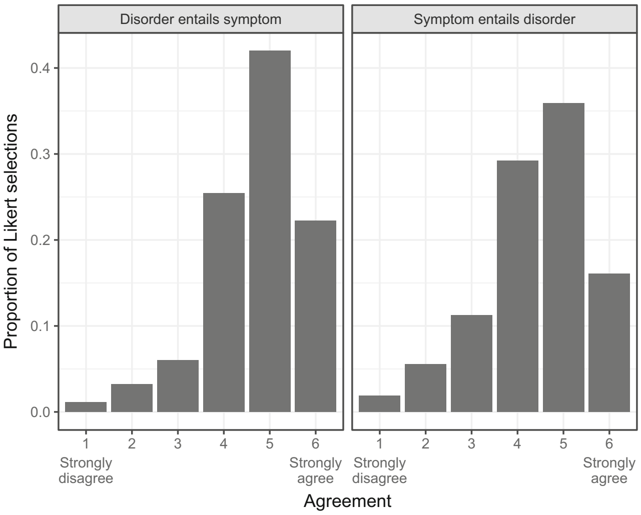

Fig. 3 A frequency plot of Likert selections of statements that a disorder entails symptoms versus symptoms entail a disorder

This result confirms the key assumption that people distinguish the entailment relationships between trait- and state-based statements in inferential tasks. Is this inference respected when people are asked to remember what they read about someone who exhibits symptoms? We examined this question in Experiments 1–4.

Experiments 1 and 2

In Experiments 1 and 2, we tested how participants remember protagonists described using state rather than trait-based language using a false-recognition paradigm. We examined participants’ memory for a passage that described a protagonist using either state or trait-based language, predicting that participants would exhibit higher rates of false recognition in the state-based language condition. That is, participants would incorrectly remember a protagonist as having a disorder even when they were only described as exhibiting symptoms of that disorder. This fact would speak to the question of what the lasting cognitive impact of subtle syntactic cues are on memory for mental illness descriptions.

Experiment 1

Preregistration We preregistered our data collection plan and analytic syntax; the raw data and code used to generate our analyses and figures are also available on the Open Science Framework: https://osf.io/rz2hk/.

Participants Participants in Experiment 1 were 113 Mechanical Turk workers who were paid $0.60 for participating in a 5-min study. Because we were unsure of the most probable effect sizes in this study, we aimed to recruit 100 participants based on a principled optional stopping procedure (Rouder, 2014). Unlike uncorrected p values, Bayes factors are unaffected by the stopping rule, allowing us to stop data collection once sufficient evidence has accumulated for either the alternative or null hypothesis. We tested whether the Bayes factor for our predicted hypothesis (namely, the interaction between condition and question type) provided strong support for the alternative or the null. If either BF_{10} or BF_{01} was greater than 20 (that is, the data strongly supported the null or alternative hypothesis), we would stop data collection. If the analysis did not provide clear support for either the null or alternative hypotheses, we would continue collecting data, adding 100 participants at a time, until there was clear evidence in one direction or we reached 400 participants. Because of constraints on how Mechanical Turk allows for the posting and acceptance of participation in the study, we ended up with a total sample size of 113 participants rather than our target of 100 participants. These participants were collected all at once. Further data collection was unnecessary because the Bayes factor for our primary parameter of interest exceeded 20.

Materials and procedure

Design

Experiment 1 was a two-way within-subjects design. Participants read four disorder passages (detailed below), and each passage was presented with either state- or trait-based language. Our experimental design guaranteed that each disorder appeared in both conditions, but once a version of the disorder passage was presented, it would not appear again in another condition.

Disorder passages Participants read a series of passages describing a person with either trait-based or stated-based language. Participants read about four protagonists, each of whom exhibited symptoms of or had a mental illness—depression, anxiety, post-traumatic stress disorder (PTSD), and obsessive-compulsive disorder (OCD). These specific mental illnesses were chosen as representative examples of common mental illnesses. The language used to describe each disorder was counterbalanced and randomized in a within-subjects design. Each participant received two state-based and two trait-based language passages in a random order. Our primary dependent variable was participant judgments of whether trait-based language occurred in the passage, in turn allowing us to compute how often participants correctly judged that a trait-based statement appeared in the passage they read. Thus, if a participant responded that trait-language appeared in the trait condition, this response was scored as correct. If a participant judged that trait-based language was used in the state condition, then this response was scored as incorrect.

It bears mentioning that although our materials cleanly map onto the state-trait distinction, they do not cleanly map onto the person-first distinction. However, as we have suggested, this distinction is better conceived of as a continuum where person-first language is a comparatively more state-based than it is trait-based language. Furthermore, the most common mental illnesses do not have felicitous identity constructions: For example, it is ungrammatical to say that someone "is a depressive" or "is an anxious". Although it is clearly grammatical to say that someone "is depressed" this construction means something different because it does not strongly entail stability of symptoms, at least according to the person-first framework. These considerations led us to contrast purely state-based constructions with *comparatively* more trait-based constructions.

The within-subjects nature of our design led us to take additional precautions so that participants would not infer the purpose of the experiment. First, they were instructed that this was a memory task and that they needed to pay attention to what they read—nothing about the task suggested that they should do anything other than try to remember as much as they could. Second, we included additional filler content to each disorder passage in order to make it difficult for participants to determine the purpose of our study. For example, each passage also stated the race of the protagonist, which was randomized and counterbalanced across vignettes (see Table S1 of the SOM). Participants were also tested on their memory of details such as race, to serve as filler questions.

Distractor passages After reading each disorder vignette, participants read a short news article, describing a current and controversial topic. These articles were chosen to be easy to read, engaging, and—again—make participants uncertain about the main purpose of our study (see Table S2 of the SOM). Participants read one news article after each disorder vignette before completing the recognition portion of the study. Participants finished reading each news article in about 90 s.

Recognition task After reading both a disorder vignette and a distractor vignette, participants were given a recognition task to examine their memory for the information they read (see Table S3 of the SOM). The task proceeded as follows: First, participants were presented with a statement (e.g., Passage 2 said, "The news article said North Korea would hold a parade..."). Participants were then asked whether that sentence appeared in the passage they read. After answering three questions about the distractor passage, participants answered three questions about the disorder passage. All of these questions followed the schema that one statement concerned an unrelated fact about the protagonist (e.g., Passage 1 said, "Hayden owns a golden retriever"), one statement concerned the race of the protagonist (e.g., Passage 1 said, "Hayden is an African American"), and one concerned the mental health of the protagonist. Of interest, was the statement about whether the protagonist had a mental illness (e.g., Passage 1 said, “Hayden has generalized anxiety disorder”). Participants were never asked the state-based version of the question by design (i.e., whether the passage said “Hayden was experiencing the symptoms of generalized anxiety disorder”) for the reasons discussed in the **Representational structure** section above.

Altogether, half of the statements in the recognition task were true and the other half were false. The order of these statements was randomized and their assignment to a given condition was counterbalanced. In total, participants answered 24 questions testing their memory (4 disorder passages × 3 memory questions = 12 questions) + (4 distractor passages × 3 memory questions = 12 questions). The primary dependent variable was whether, for a given disorder passage, participants correctly chose that trait-based description was present or absent. We compared the proportion of correct responses for disorder questions to control questions to ensure that some passages were not simply easier to remember than others.

Results

Figure 4 displays recognition performance across condition and question in Experiments 1 and 2. First, the figure indicates participants that followed the instructions to focus on remembering as much as they could about each passage—we found that people correctly remembered the answers to the control questions (that is, all those questions that did not concern mental illness) across both conditions. Furthermore, we found that they correctly remembered that the protagonist in the vignette had a mental illness in the trait-based condition. Consistent with our hypothesis, participants’ recognition memory performance was very poor in the state condition. To test this prediction formally, we performed Bayesian logistic mixed-effects modeling using the R package `brms` (Bürkner, 2017). We regressed accuracy of responses on Condition (0 = Trait, 1 = State) and question type (0 = Control, 1 = Disorder), and the interaction between these predictors, allowing for the effects and interaction of condition and question type to vary for each participant (i.e., the maximal model given our design; see Barr, Levy, Scheepers, & Tily, 2013). For brevity, we specify the model in `brms` syntax:

```
Recognition Accuracy ~
Condition*Question + (1 +
Condition*Question|Subject)
```

Bayesian statistical models formulate model parameters (which are unknown) as probability distributions wherein the joint probability distribution of the data, $\vec{y}$ (which is known), and model parameters $\vec{\theta}$ are computed via the prior probability of $\vec{\theta}$ and the probability of $\vec{y}$ given $\vec{\theta}$ (Schoot, Depaoli, King, Kramer, Kaspers, & Tadesse, 2021).

$$p(y, \theta) = p(y|\theta) \times p(\theta).$$

This formula states that the joint probability of y and θ equals the probability of y given θ (i.e., “the Likelihood”) multiplied by the probability of θ (i.e., “the Prior”). Sticking

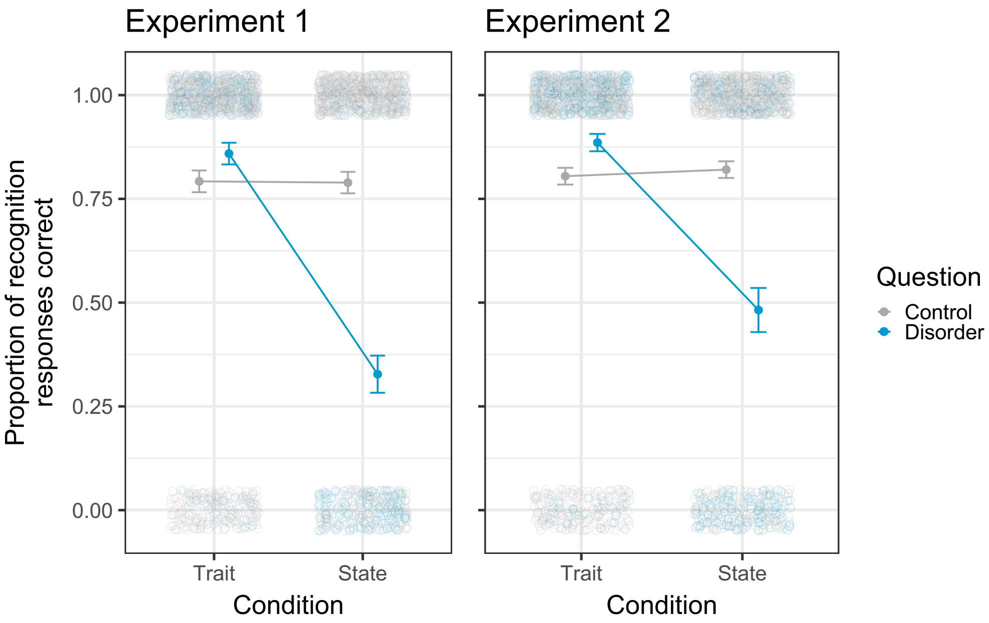

Fig. 4 Model-predicted proportion of responses that participants correctly recognized as appearing in the vignette they read for control and disorder questions in the trait and state conditions in Experiments 1 (*left panel*) and 2 (*right panel*). *Error bars* are ±1 within-subjects standard errors of the mean, and raw data (1 = Correct, 0 = Incorrect), is jittered to avoid overplotting

with convention, y denotes the data and θ denotes a set of parameters of a distribution (e.g., μ and σ are parameters of the normal distribution). The equivalence stated above can be derived from Bayes' theorem, which is used to calculate the posterior probability of $\vec{\theta}$ given $\vec{y}$:

$$p(\theta|y) \propto p(y|\theta) \times p(\theta).$$

This formula states that the probability of θ given y (the "Posterior") is proportional to the Likelihood and the Prior. For instance, if we wanted to compute the posterior probability of a mean given the data, we would need to compute the probability of the data given the mean and the prior probability of the mean.

$$p(\mu|y) \propto p(y|\mu) \times p(\mu).$$

To model the joint probability distribution of participants' responses [i.e., $p(y, \theta)$], we need to set priors over the possible effects each model parameter θ could have on the response variable y. We specified the following regularizing priors over the possible effects each parameter could have on the response variable (e.g., Bürkner, 2017; McElreath, 2016).

$$\begin{aligned} \alpha &\sim \mathcal{N}(0.00, 0.50) \\ \text{All } \beta &\sim \mathcal{N}(0.00, 0.50) \\ \sigma &\sim \mathcal{N}(1.00, 3.00) \\ \Omega_k &\sim LKJ(1.20) \end{aligned}$$

This model revealed the predicted interaction between condition and question type, $b = -2.53$, 95% CI [-3.13 to -1.93], $BF_{10} > 1000$. These findings are consistent with the prediction that although state-based language may impact how people conceive of someone's mental illness, the language may not induce long-lasting changes in the situation model of the passage they read. When people are presented with information which states that a protagonist exhibits certain symptoms—symptoms that are consistent with the presence of a mental illness—people responded that the protagonist has this illness. This finding is consistent with the undirected cyclic graph shown in Fig. 2.

Experiment 2

In Experiment 1, we found that people incorrectly remember that someone has a mental illness when they are only described with symptoms of that disorder (and not the disorder itself). However, it is also possible that the strength of the effect we observed in Experiment 1 is exaggerated by an artifact of the design. Specifically, the language in the state condition (e.g., "is extremely *anxious*") semantically overlaps, so to speak, with the language in the trait condition (e.g., "has an *anxiety disorder*"). The language in the state condition may act as a lure that drives up false recognition due to the similarity between the language in the passage and the language in the question. To address this possibility, in Experiment 2, we removed the shared features between the passage and the recognition question. This change allowed us to determine whether presenting symptoms associated with a disorder is sufficient to lead people to remember a protagonist as having a mental illness.

Preregistration We preregistered our data collection plan and analytic syntax. The raw data and code used to generate our analyses and figures are also available on the Open Science Framework: https://osf.io/rz2hk/.

Participants Participants in Experiment 2 were 112 Mechanical Turk workers who were paid $0.60 for participating in the study. Participants from Experiment 1 were prevented from participating in Experiment 2. Because Experiment 1 provided some guidance for the effect sizes we were likely to observe, we performed a power analysis based on the effect size we observed in Experiment 1. Our power analysis suggested we would only need 12 participants to observe an effect of this magnitude with 99% statistical power. This sample sized seemed implausibly small, and because we conjectured the effects in Experiment 2 were likely to be smaller after removing the artifact of our design, we aimed to collect the same number of participants we collected in Experiment 1.

Procedure The procedure for Experiment 2 was identical to Experiment 1. Experiments 1 and 2 differed in only one respect: we removed surface-level similarities from the vignette in the state condition and the question participants were asked in the recognition task. For example, rather than saying that "Nicole feels anxious all the time," the vignette said that "Nicole feels overwhelmed and panicked all the time". As another example, rather than stating that "Lola feels obsessive about carrying out these compulsions" the vignette stated that "Lola feels the need to wash her hands all the time to get rid of these thoughts" (see Table S4 of SOM for full materials).

Results

The same priors and regression model were fit in Experiment 2 as Experiment 1 (see Appendix for complete list of models and priors). Figure 4 (right panel) displays the proportion of correct responses in recognition phase across condition and question. Even with only indirect cues, participants exhibited high rates of false recognition in the state condition—people remembered that the protagonist in the passage had a mental illness when this was never stated, $b = -2.21$, 95% CI [-2.75 to -1.65]. This result suggests the mere mention of symptoms associated with a disorder may be sufficient to cause people to incorrectly respond that someone has a disorder.

Experiment 3

In Experiments 1 and 2, we observed that people appear to incorrectly remember that someone has a mental illness when they are only described with symptoms of that disorder. The purpose of Experiment 3 was to identify whether we could observe analogous effects using a free recall task. In this experiment, we instructed participants to "write down everything they could remember about the passages they read". We then coded their free responses to determine whether they recalled the protagonist had a disorder even when this information was not presented in the passage. The free response format removed possible demand characteristics or suggestibility introduced in the previous experiments.

Method

Participants Participants in Experiment 3 were 246 Mechanical Turk workers who were paid $2.00 for participating in the study. We increased the number of participants we recruited because we anticipated that even using a within-subjects design, our statistical power would be reduced because of the statistical noise introduced by a free recall task. We also increased participants' compensation because of the additional time the study required. Participants from Experiments 1 and 2 were prevented from participating in Experiment 3. There were 237 participants included in our analyses after excluding participants who incorrectly answered questions checking their attention. These exclusions were in accordance with our preregistration, but including all participants in our analyses does not impact our model parameter estimates. The raw data and code for Experiment 3 are available on the Open Science Framework: https://osf.io/rz2hk/.

Procedure The procedure for Experiment 3 was similar to Experiments 1 and 2. Participants read the series of passages describing a person with either trait-based or state-based language. After each passage, participants read a distractor passage and answered questions about it thereafter. Then participants were asked to write down everything they remembered about the disorder passage (labeled as Passage 1 in the experiment). As in Experiments 1 and 2, participants read a total of four disorder passages along with their corresponding distractor passages.

Coding free responses Participants in Experiment 3 were instructed to write down everything they remembered about each disorder passage. First, we coded participants' responses for whether a disorder itself (e.g., depression) and symptoms were mentioned in participants' free responses (denoted "disorder language"). When coding their responses, the possible disorder language categories were: the disorder was mentioned (ignoring whether symptoms were also mentioned), only symptoms were mentioned, or neither the disorder nor symptoms were mentioned. We then calculated a measure of "accuracy" based on these codes as follows:

> If a participant's response included trait-based language and they were in the trait condition, their response was scored as **correct**.
>
> Similarly, if a participant read a state-based passage and their response included only state-based (but not trait-based) language, their response was scored as **correct**.
>
> Conversely, if a participant's response included only state-based language and they were in the trait condition, their response was scored as **incorrect**.
>
> If a participant read a state-based passage and their response included trait-based language, their response was scored as **incorrect**.

There were very few responses that did not fall in one of these categories, so these responses were omitted in our analyses but their inclusion does not impact our parameter estimates.

Second, we counted how many times the disorder and/or the symptoms were mentioned (denoted "count language"; see SOM for full coding guide). When determining the count of disorder and/or symptoms language used, 1 was added to the count every time the disorder label or a distinct symptom was mentioned. For example, the following response given by a participant, "Tyler is suffering from post-traumatic stress disorder and has flashbacks and lives in a metropolitan area with his roommate," was given a count of 2 because both the disorder and the symptom of flashbacks were included in this participant's response.

Authors E. Line and Z. Horne coded participants' responses. They were blinded to the condition a given response was in and participants' responses were randomized using an R script prior to either coders seeing the raw data. This was done by assigning each response a unique identification number. The responses and the identification numbers were pulled from the rest of the data set and arranged by identification number. The coders only saw the responses themselves along with their corresponding identification number when they coded responses, thus blinding them to the condition of each response. After responses had been coded, the coded variables were matched back into the full data set.

The entire dataset was coded by first author, E. Line. After coding this data, reliability was computed by comparing the E. Line's codes to Z. Horne's codes, who coded 20% of the data (Syed & Nelson, 2015). There was strong agreement in codes of disorder language, 96% agreement and Cohen's $\kappa = .84$. The codes of our secondary variable of interest, count language, were strongly correlated, $r = .8$.

Third, we addressed the possibility that the state-based passages are inherently more difficult to recall in general by calculating the proportion of a participant's typed non-disorder language that matched the original passage. We computed the mathematical intersection of participants' typed-responses to the original passage (minus the disorder-related language) using the R package `stringr`. This package matches the strings (e.g., words, phrases, etc.) in participants' free responses to the strings in the original passage. We then divided the number of matched words by the total number of unique words in the original passage. Repeated words in either passage were not counted more than once.

Results

We hypothesized that if participants equate a state-based description with a statement of the disorder itself, we should see less accurate recall of the disorder passage in the state versus trait condition. Specifically, we predicted participants in the state condition to be more likely to recall that a disorder was mentioned in the passage—an error of commission—than would participants in the trait condition fail to recall the disorder—an error of omission.

Figure 5 displays the proportion of correct recall of disorder language across condition. We performed Bayesian logistic regression, predicting correct recall of disorder language on Condition (0 = Trait, 1 = State), allowing for the main effects of condition to vary for each participant. Priors and model syntax are located in the Appendix. This analysis revealed that the proportion of accurate responses is higher in the trait condition than in the state condition, $b = -0.83$, 95% CI $[-1.54$ to $-0.18]$. Participants exhibited more errors-of-commission in the state condition than they did errors-of-omission in the trait condition. This result suggests the mention of symptoms associated with a disorder can lead participants to interpret, and in turn recall, that someone has a disorder when only the symptoms are mentioned. This result supports our findings in Experiments 1 and 2.

The design of this experiment also provides some insight into participants' situation models of trait and state passages. Although participants were instructed to write down exactly what they read, many participants included interpretations of the passages that were not explicitly stated. For example, a participant stated that "Hayden lived with his parents *because* he was depressed" while the passage only explicitly stated that Hayden was depressed *and* that he lived with his parents. Setting anecdotal observations aside, our primary findings support the hypothesis that the situation models of state- versus trait-based descriptions are more similar than advocates of the person-first initiative might hope.

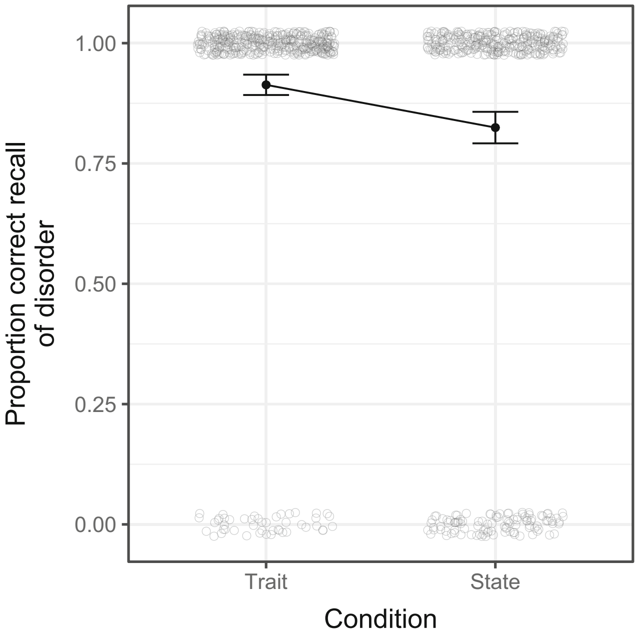

Fig. 5 Model-predicted proportion of responses in which participants correctly used state- or trait-based language based on the condition of the passages read in Experiment 3. *Error bars* are ±1 within-subjects standard errors of the mean, and raw data (1 = Correct, 0 = Incorrect), is jittered to avoid overplotting

Nonetheless, this set of analyses highlights one limitation of the free recall paradigm. The central comparison in this study is whether there are more errors-of-commission—recalling trait language in the state condition—to errors-of-omission—failing to recall trait language that did appear in the trait condition. Although this aspect of our design makes it more difficult to directly compare the conditions, we reasoned that this comparison would provide a conservative test of our hypothesis. On average, we expected that errors-of-omission should be more common than errors-of-commission, particularly among a participant population primarily motivated to complete the study as quickly as possible (read, Mechanical Turk workers). Still, to address the limitations inherent to this analysis, we conducted several follow-up analyses to evaluate the robustness of this result.

First, we ran an automated string-matching analysis to test how well the typed responses from participants matched the disorder sentence in the original passage. We predicted that even using this automated method, we would see a lower match in participants' typed responses in the state condition than in the trait condition. While this finding on its own would not indicate the *source* of the mismatch between responses and the original passage (the disorder sentence could be inherently more difficult to recall in the state versus trait condition) a lower match between response and original would provide converging evidence with the analysis reported above. To test this, we performed an exploratory analysis by fitting a beta regression model predicting percentage match between typed response and the disorder language in the original passage based on condition. Consistent with the results of our main analysis, we found the percentage match was credibly lower in the state condition than trait condition, $b = -0.09, 95\%$ CI [-0.18 to 0.00].

Next, we fit two models aimed at comparing how many disorder-related words (i.e., count language) were used in both conditions. We predicted that there would be fewer disorder words in the trait than state description because the label is shorthand, as it were, for a longer description. However, there was severe and unanticipated range restriction in the count language dependent variable—nearly all responses were either counts of 1 or 2. A Poisson regression model suggested there was no difference in disorder count words between the two conditions, contrary to our initial hypothesis, $b = -0.04, 95\%$ CI [-0.07 to 0.15].

This could be a genuine null effect or the result of range restriction, so we sought to resolve this by exploring whether there was evidence that participants treat traits, either implicitly or explicitly, as equivalent with longer descriptions of symptoms. In an exploratory analysis, we modeled how the length of participants' responses in the state condition differed as a function of whether they (incorrectly) recalled the label as being present or not. We predicted that within the state condition, participants who recalled a label would provide shorter responses than participants who did not. Our conjecture was that participants who recalled the disorder would truncate their responses because the label "stands in" for the symptoms. Consistent with this prediction, we found that when participants incorrectly recalled that the disorder was present in the state condition they also typed shorter responses, $b = -0.14, 95\%$ CI [$-0.21, -0.07$]. These results provide preliminary evidence that participants' recollections were truncated because trait labels stand in for the symptoms.

As noted above, it is also possible that it is inherently more difficult to remember state-based passages. If this is correct, we should observe that condition also predicts recall accuracy for non-disorder language. To test this, we modeled the proportion of matched responses to the non-disorder-related passages as a function of condition. We observed that there were no differences in participants' recall for non-disorder sentences for state versus trait passages, $b = -0.01, 95\%$ CI [-0.07 to 0.06] providing evidence against the hypothesis that the state-based passages were more difficult to remember in general.

Together, our preregistered analyses and further exploratory modeling provide converging evidence for our hypothesis: Although state- and trait-based statements have distinct entailments which participants recognize in inferential tasks, their memory and recollection of state-based passages suggest that the mention of symptoms leads them to infer the presence of the disorder.

Experiment 4

Experiment 4 consisted of a forced-choice version of the task used in Experiment 2. Participants were asked to identify which of two sentences, state or trait, they read in a passage. The design of Experiment 4 was otherwise identical to Experiment 2. We predicted that this task would be comparatively easy for participants and that it may identify a boundary condition on effects we observed in Experiments 1 through 3—how poor is a participant's memory when prompted with the very explicit option of choosing between an excerpt they did see and an excerpt they did not? We were unsure how likely this would be given the ease of this task. For this reason, we anticipated the effect was likely to be small, so we recruited a considerably larger sample in Experiment 4.

While we anticipated greater uncertainty with this design, we preregistered the hypothesis that there would be higher error rates in the state condition than the trait condition. Stated another way, participants would remember a person described as experiencing symptoms of a disorder as having the disorder, as we observed in Experiments 1—3.

It is worth pausing to note the limitations on the inferences one could draw from the design of Experiment 4. As outlined in the representational structure section of the paper, from an entailment perspective, falsely recognizing a trait sentence was present when in the state condition is not the same error as falsely recognizing a state sentence was present in the trait condition. As we found, people believe that a disorder entails the symptoms of a disorder more than they believe symptoms entail a disorder—this is a key assumption of the person-first language initiative. Thus, if a participant in the trait condition erroneously chooses the state response option, while still an error, in that the sentence was not verbatim presented, the state response option is entailed by the trait passage and strong entailments are already known to induce errors of this sort (Powell et al., 2015; Gentner, 1981). In contrast, if a participant in the state condition erroneously chooses the trait response option, this is a different kind of error, one which is incompatible with people's entailment judgments in inferential tasks. Consequently, direct comparisons of error rates across conditions need to be interpreted with caution because a null-effect could be the product of two different kinds of errors occurring simultaneously.

Participants Participants were 602 Mechanical Turk workers who were paid $2.00 for participating in the study. After completing the main task, participants completed the entailment rating task described in the **Representational structure** section of the paper, so we increased participant compensation. We increased the number of participants for Experiment 4 compared to the sample sizes in Experiments 1–3 because we anticipated a smaller condition difference in this task. There were 526 participants included in our analyses after excluding participants who incorrectly answered questions checking their attention. These exclusions were in accordance with our preregistration, but including all participants in our analyses does not impact our model parameter estimates. Mechanical Turk workers who had previously participated in Experiments 1–3 were prevented from participating in this study. The raw data and code used to generate our analyses and figures are also available on the Open Science Framework: https://osf.io/rz2hk/.

Procedure The procedure for Experiment 4 closely followed the procedures in the previous experiments. Participants read the series of passages describing a person with either trait-based or state-based language. After each disorder passage, participants read a distractor passage and answered questions about it immediately afterwards. Then, participants were asked to identify which of two sentences they had read in the passage. One sentence included the disorder label and the other sentence included a symptom description which was non-redundant with the symptoms included in the trait-passage. Specifically, both passages listed symptoms of the disorder, but the state passage included an additional symptom that was not included in the trait passage to ensure the passages were the same length. For example, while a trait passage said "Nicole has an anxiety disorder", the state passage said "Nicole frequently feels overwhelmed and panicked". These two phrases were then used as the two response options for the anxiety disorder passage.

Results

We hypothesized that condition (trait vs. state) would predict the proportion of correct responses, where this proportion would be lower in the state condition than the trait condition. To test this hypothesis, we regressed correct recognition on Condition (0 = Trait, 1 = State), allowing for the effect of condition to vary for each participant. The priors and model specification for this analysis are located in the Appendix.

Contrary to our hypothesis, we observed no difference between conditions in accuracy of response: a Bayesian logistic mixed-effects model indicated nearly identical performance in both conditions $b = -0.12$, 95% CI [−0.49 to 0.20]. Participants correctly identified the passage they read approximately 92% of the time. Figure 6 shows the proportion of accurate and inaccurate responses in each condition.

Experiment 4 identifies a boundary condition on the effects we observed in Experiments 1—3. In Experiments 1 and 2, we observed that participants were more likely to incorrectly select that the disorder was present in the passage they read in the state condition. We observed even stronger evidence for this tendency in Experiment 3: participants who were asked to recall everything they remembered about the passage verbatim made more errors in the state condition than the trait condition, specifically recalling the disorder was present in the state condition more often than participants in the trait condition omitted this information. These studies suggest participants may

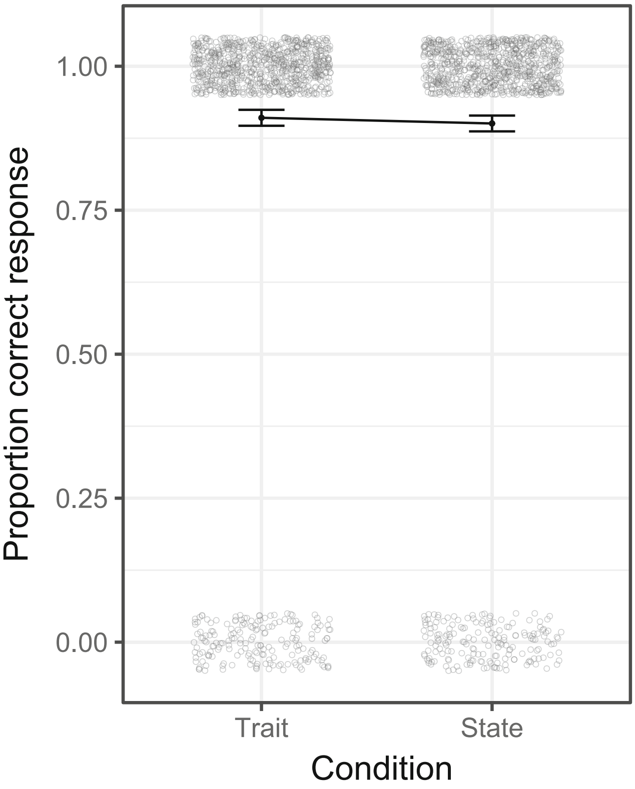

Fig. 6 Model-predicted proportion of responses in which participants correctly chose state- or trait-based excerpts based on the condition of the passages read in Experiment 4. *Error bars* are ±1 within-subjects standard errors of the mean, and raw data (1 = Correct, 0 = Incorrect), is jittered to avoid overplotting

be interpreting state-based descriptions as indicative of the presence of a disorder, a tendency we see manifested in their memory. However, Experiment 4 demonstrates that when participants are explicitly presented with the exact statements they read, they make few errors and error-rates do not differ across conditions. While these findings do not impugn the results of the previous three studies, they do suggest that people's memories are not so error prone that they cannot recognize the exact wording they were presented with after a short delay.

Still, it is important to acknowledge that participants are making distinct kinds of errors when they incorrectly select a state response option while in the trait condition versus a trait response option while in the state condition, and so this null effect needs to be interpreted with some caution.

Discussion

What is the lasting cognitive impact of using state-based language to describe people experiencing the symptoms of mental illness? In Experiments 1 and 2, we found people incorrectly judge that a passage indicated a protagonist had a disorder when only the symptoms of this disorder were stated. In Experiment 3, we found that when people freely recall passages in which a protagonist is described as having symptoms of a disorder, they recall the protagonist had a disorder even when this information wasn't present. Still, Experiment 4 identified a boundary condition on these effects: when participants were forced to choose whether state or trait-based language was presented in the passage they read, they made a similar number of errors, suggesting people's memories of a brief passage are not so degraded that they cannot recover what they read when they are explicitly prompted.

As in prior research, the present findings provide some evidence people remember the meaning of the passage better than they remember the syntactic structure of the passage itself (Deese, 1959; Roediger & McDermott, 1995; Loess, 1967; Loftus et al., 1978; Loftus & Palmer, 1974; Sachs, 1967; Sulin & Dooling, 1974). State- and trait-based descriptions of mental illness may have more similar situation models than the person-first language initiative may assume (Van Dijk & Kintsch, 1983), although the results of Experiment 4 indicate that with explicit prompting people can recover exactly what they read. Experiment 3 provided the clearest evidence for the claim that the situation models of state- and trait-based descriptions are similar: participants recalled that a disorder was stated in the passage even when this information was not present.

The results of Experiments 1–3 speak to the extent at which people's memories can distort the description of a person, which may pose problems for clinicians aiming to avoid identifying those suffering with mental illness with their disorder. The present research provides some evidence that people revert to trait-based interpretations of what they read, leading them to label people who are merely experiencing symptoms of a mental illness as actually having a mental illness. The tendency to incorrectly label people with a mental illness when they were only described as having symptoms could lead to increased stigma surrounding the person in question (Link & Phelan, 2001), though further research is needed to test this possibility. This finding is all the more notable given that our conceptual replication indicated people readily distinguish between the entailment relationship between disorders and symptoms.

Our results may raise a question about whether in other domains there could be long-lasting effects of small syntactic changes. For example, Bryan and colleagues (2011) found that trait-based versus state-based constructions of sentences about the importance of "being a voter" or "voting" resulted in larger voter turnout. Although our findings may seem to be in tension with some of these effects, there are several differences between these cases that distinguish them. For example, Bryan and colleagues (2011) hypothesize that trait versus state language changes how people see themselves, rather than how they remember an unknown protagonist. In turn, these subtle manipulations may allow for long-lasting impact because at the time the "voter versus voting" question is read, it shifts the reader's self-conception. In this way, some syntactic shifts may have lasting efficacy because they do not *hinge* on memory for sustaining those shifts.

This point highlights other aspects of the scope of our findings. In all four experiments, we used diagnostic possessive phrases (e.g., "Nicole has anxiety") as the trait-based condition and symptom-description phrases (e.g., "Nicole feels anxious all the time" or "Nicole feels overwhelmed and panicked all the time") as the state-based condition. This allowed us to include common mental illnesses such as depression and anxiety which do not have felicitous trait-based labels (e.g., "is a depressive"). Consequently, both of our conditions are shifted, so to speak, towards state-based language on the state-to-trait continuum (depicted in Fig. 1). The person-first language initiative recommends people avoid using common noun phrases (e.g., "Smith is a schizophrenic"), instead recommending people use diagnostic possessive phrases or other more state-based constructions (e.g., "Smith has schizophrenia"). However, we nonetheless found that even when participants were presented with state-based language they recalled *comparatively* more trait-based language. This effect may cut against one aspect of the underlying motivation for using person-first language. Consequently, these experiments provide some evidence that typical semantic influences on episodic memory are likely to impact how people remember subtly different constructions of mental illness descriptions.

Future directions and limitations

It is possible that language changes reinforced over time could make a difference in people's interpretation and recollection of people with symptoms of mental illness. The person-first language initiative may accompany an overall cultural shift around mental illness (Office of the Surgeon General (US); Center for Mental Health Services (US); National Institute of Mental Health (US), 2001). Not only does the person-first initiative recommend changing how clinicians talk about mental illness (i.e., syntactic constructions), it also admonishes against slang terms (e.g., "crazy" or "paranoid") and emphasizes a focus on the abilities and independence of people with mental illnesses (American Psychiatric Association, 2013). Because anti-stigma initiatives like the person-first language initiative may be continuously reinforced, it is possible this would lead people to better recall subtle changes in descriptions of people with mental illnesses, ultimately overcoming our tendency to equate state- and trait-based descriptions after a delay. This would indicate lengthy interventions might be needed to yield sustained changes in people's memories of those exhibiting symptoms of mental illness. However, future research would be needed to determine how shifting the norms of how we talk about mental illness impacts memory.

It also must be acknowledged that some mental illnesses may seem more salient than others and in turn could affect how syntactic constructions are remembered. The stimuli from these experiments focused on common mental illnesses. Several studies suggest certain severe mental disorders, such as personality disorders, have higher degrees of stigma associated with them (see Sheehan, Nieweglowski, & Corrigan, 2016; Aviram, Brodsky, & Stanley, 2006; Dickerson, Sommerville, Origoni, Ringel, & Parente, 2002), which perhaps could interact with the strength of syntactic manipulations. As a consequence, stimuli focusing on severe mental illnesses, such as borderline personality disorder or schizophrenia, may affect the strength of the effects we observed in Experiments 1–3. Along these lines, the duration of time between participants reading the target passage and answering questions regarding the passage was identical across all studies. It is possible that incorporating different time delays in this design could reveal boundaries or further generalizations for the effects observed here. For instance, it is possible we would observe the predicted condition effect in Experiment 4 if the delay was longer between reading the passage and the recognition task. Further research is necessary to resolve this issue.

Nonetheless, our findings suggest that under some conditions, people fail to remember subtle syntactic differences, instead falsely remembering a protagonist has a mental illness even when they were only described as experiencing symptoms of this illness. We observed this effect using both proximal and remote cues, and in a recall task which was free of task demands by design. Consequently, although state-based language initiatives have been adopted to lessen people's tendency to stigmatize people with mental illnesses, our studies suggest making syntactic changes in the way we talk about mental illness may not always yield the lasting effects we hope they would.

Appendix

Entailment replication

```
Likert Response ∼ Statement Type + (1
+ Statement Type|Subject), family =
''cumulative'', link = ''logit''
```

$$\begin{aligned}
\tau_1 &\sim \mathcal{N}(-2.51, 1.00)\\
\tau_2 &\sim \mathcal{N}(-1.73, 1.00)\\
\tau_3 &\sim \mathcal{N}(-1.09, 1.00)\\
\tau_4 &\sim \mathcal{N}(0.00, 1.00)\\
\tau_5 &\sim \mathcal{N}(1.50, 1.00)\\
\text{All } \beta &\sim \mathcal{N}(0.00, 1.00)\\
\sigma &\sim \mathcal{N}(1.00, 2.00)\\
\Omega_k &\sim LKJ(1.00)
\end{aligned}$$

Experiment 3 models and priors

Primary analysis:

```
Recall Accuracy ∼ Condition + (1
+ Condition|Subject), family =
''bernoulli'', link = ''logit''
```

$$\begin{aligned}
\alpha &\sim \mathcal{N}(1.12, 1.00)\\
\text{All } \beta &\sim \mathcal{N}(0.00, 1.00)\\
\sigma &\sim \mathcal{N}(1.00, 1.00)\\
\Omega_k &\sim LKJ(1.20)
\end{aligned}$$

Exploratory analysis using string matching:

```
Percent Match for Disorder Language ~
Condition + (1 + Condition|Subject),
family = ''beta'', link = ''logit''
```

$$\alpha \sim \mathcal{N}(-1.00, 2.00)$$
$$\text{All } \beta \sim \mathcal{N}(0.00, 1.00)$$
$$\sigma \sim \mathcal{N}(1.00, 2.00)$$
$$\text{Shape parameter } \phi \sim \gamma(0.01, 0.01)$$
$$\Omega_k \sim LKJ(1.20)$$

Secondary analysis of counts of disorder language:

```
Count of Disorder Language ~ Condition
+ (1 + Condition|Subject), family =
''poisson'', link = ''log''
```

$$\alpha \sim \mathcal{N}(1.00, 1.00)$$
$$\text{All } \beta \sim \mathcal{N}(0.00, 0.50)$$
$$\sigma \sim \mathcal{N}(0.50, 0.50)$$
$$\Omega_k \sim LKJ(1.20)$$

Exploratory analysis of response length for trials where disorder was remembered or not in state passages only:

```
Response Length ~ Remembered Disorder
+ (1 + Remembered Disorder|Subject),
family = ''poisson'', link = ''log''
```

$$\alpha \sim \mathcal{N}(3.00, 0.50)$$
$$\text{All } \beta \sim \mathcal{N}(0.00, 0.50)$$
$$\sigma \sim \mathcal{N}(0.20, 0.20)$$
$$\Omega_k \sim LKJ(1.20)$$

```
Percent Match on Non-Disorder Language
~ Condition + (1 + Condition|Subject),
family = ''beta'', link = ''logit''
```

$$\alpha \sim \mathcal{N}(0.00, 2.00)$$
$$\text{All } \beta \sim \mathcal{N}(0.00, 1.00)$$
$$\sigma \sim \mathcal{N}(1.00, 2.00)$$
$$\text{Shape parameter } \phi \sim \gamma(0.01, 0.01)$$
$$\Omega_k \sim LKJ(1.20)$$

Experiment 4 models and priors

```
Recognition Accuracy ~ Condition +
(1 + Condition|Subject), family =
''bernoulli'', link = ''logit''
```

$$\alpha \sim \mathcal{N}(1.10, 1.00)$$
$$\beta \sim \mathcal{N}(0.00, 0.25)$$
$$\sigma \sim \mathcal{N}(1.00, 1.00)$$
$$\Omega_k \sim LKJ(1.20)$$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-021-01208-8.

Open Science Practice Statement We preregistered the data collection plan and analytic syntax for both experiments. The raw data and code used to generate our analyses and figures are also available on the Open Science Framework (Experiment 1: https://osf.io/cfgpx/; Experiment 2: https://osf.io/mvy7p/, Experiment 3: https://osf.io/kpfae/, Experiment 4: https://osf.io/d8n4r/). The final analysis scripts, data, and Supplementary Online Materials (SOM) can be found at https://osf.io/rz2hk/).

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (DSM-5®). American Psychiatric Pub.

American Psychological Association. (2010). *Publication manual of the American Psychological Association*, (6th ed.). Washington: American Psychological Association.

Aviram, R., Brodsky, B., & Stanley, B. (2006). Borderline personality disorder, stigma, and treatment implications. *Harvard Review of Psychiatry*, *14*(5), 249–256.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, *42*(2), 228–235.

Bastian, B., & Haslam, N. (2007). Psychological essentialism and attention allocation: Preferences for stereotype-consistent versus stereotype-inconsistent information. *The Journal of Social Psychology*, *147*(5), 531–541.

Becker, R. B., Ferretti, T. R., & Madden-Lombardi, C. J. (2013). Grammatical aspect, lexical aspect, and event duration constrain the availability of events in narratives. *Cognition*, *129*(2), 212–220.

Blaska, J. (1993). The power of language: Speak and write using "person-first". *Perspectives on Disabilities*, 25–32.

Bransford, J., & Franks, J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*(4), 331–350.

Brewer, W. (1977). Memory for the pragmatic implications of sentences. *Memory & Cognition*, *5*, 673–678.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Broyles, L., Binswanger, I., Jenkins, J., Finnell, D., Faseru, B., Cavaiola, A., et al. (2014). Confronting inadvertent stigma and pejorative language in addiction scholarship: A recognition and response. *Substance Abuse*, *35*(3), 217–221.

Bryan, C., Walton, G., Rogers, T., & Dweck, C. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(31), 12653–12656.

Cassels, J., & Johnstone, A. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, *61*(7), 613.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Corrigan, P., Druss, B., & Perlick, D. (2014). The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest, 15*(2).

Corrigan, P., Larson, J., Sells, M., Niessen, N., & Watson, A. (2006). Will filmed presentations of education and contact diminish mental illness stigma? *Community Mental Health Journal*, *43*(2), 171–181.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Dickerson, F., Sommerville, J., Origoni, A., Ringel, N., & Parente, F. (2002). Experiences of stigma among outpatients with schizophrenia. *Schizophrenia Bulletin*, *28*(1), 143–155.

Fletcher, C. (1994). Gernsbacher, M. A. (Ed.) *Levels of representation in memory for discourse*. Cambridge: Academic Press.

Fletcher, C. R., & Chrysler, S. T. (1990). Surface forms, textbases, and situation models: Recognition memory for three types of textual information. *Discourse Processes*, *13*(2), 175–190.

Gelman, S., & Heyman, G. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science*, *10*(6), 489–493.

Gentner, D. (1981). Integrating verb meanings into context. *Discourse Processes*, *4*(4), 349–375.

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, *17*(3), 324–363.

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*.

Granello, D., & Gibbs, T. (2016). The power of language and labels: "The mentally ill" versus "people with mental illnesses" *Journal of Counseling and Development*, *94*, 31–39.

Helbig, H. (2006). Knowledge representation and the semantics of natural language. Springer.

Horne, Z., & Cimpian, A. (2018). Subtle syntactic cues affect intuitions about knowledge: Methodological and theoretical implications for epistemology. In Lomborzo, T., Knobe, J., & Nichols, S. (Eds.) *Oxford studies in experimental philosophy*, (Vol. 2, pp. 7–41). Oxford: Oxford University Press.

Howell, A., Ulan, J., & Powell, R. (2014). Essentialist beliefs, stigmatizing attitudes, and low empathy predict greater endorsement of noun labels applied to people with mental disorders. *Personality and Individual Differences*, *54*(2), 33–38.

Howell, A., & Woolgar, S. (2013). Essentialism and compassion: Predicting preference for noun labels applied to people with mental disorders. *Personality and Individual Differences*, *54*(1), 87–91.

Kelly, J., Wakeman, S., & Saitz, R. (2015). Stop talking 'dirty': Clinicians, language, and quality of care for the leading cause of preventable death in the United States. *The Journal of American Medicine*, *128*(1), 8–9.

Kelly, J., & Westerhoff, C. (2010a). Does our choice of substance-related terms influence perceptions of treatment need? An empirical investigation with two commonly used terms. *Journal of Drug Issues*, *40*, 805–818.

Kelly, J., & Westerhoff, C. (2010b). Does it matter how we refer to individuals with substance-related problems? A randomized study with two commonly used terms. *International Journal of Drug Policy*, *21*, 202–207.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*(2), 133–159.

Link, B. (1987). Understanding labeling effects in the area of mental disorders: An assessment of the effects of expectations of rejections. *American Sociological Review*, *52*, 96–112.

Link, B., & Phelan, J. (2001). Conceptualizing stigma. *Annual Review of Sociology*, *27*, 363–385.

Link, B., Phelan, J., Bresnahan, M., Stueve, A., & Pescosolido, B. (1999). Public conceptions of mental illness: Labels, causes, dangerousness, and social distance. *American Journal of Public Health*, *89*(9), 1328–1333.

Loess, H. (1967). Short-term memory, word class, and sequence of items. *Journal of Experimental Psychology*, *74*, 556–561.

Loftus, E., Miller, D., & Burns, H. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology*, *4*, 19–31.

Loftus, E., & Palmer, J. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585–589.

Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes*, *29*(2), 83–112.

Martin, J., Pescosolido, B., & Tuch, S. (2000). Of fear and loathing: The role of 'disturbing behavior,' labels, and causal attributions in shaping public attitudes toward people with mental illness. *Journal of Health and Social Behavior*, *41*(2), 208–223.

McElreath, R. (2016). *Statistical rethinking*. Boca Raton: CRC Press.

National Alliance on Mental Illness (2019). StigmaFree. https://www.nami.org/stigmafree

Office of the Surgeon General (US); Center for Mental Health Services (US); National Institute of Mental Health (US) (2001). *Mental health: Culture, race and ethnicity: A supplement to mental health: A report of the surgeon general*. Rockville, Substance Abuse and Mental Health Services Administration (US).

Papish, A., Kassam, A., Modgill, G., Vaz, G., Zanussi, L., & Patten, S. (2013). Reducing the stigma of mental illness in undergraduate medical education: A randomized controlled trial. *BMC Medical Education* 13.

Pauker, K., Ambady, N., & Apfelbaum, E. (2010). Race salience and essentialist thinking in racial stereotype development. *Child Development*, *81*(6), 1799–1813.

People First Respectful Language Modernization Act of 2006 (2006). Pub. L. No. B16–0665.

Pivovarova, E., & Stein, M. (2019). In their own words: Language preferences of individuals who use heroin. *Addiction, 114*(10).

Powell, D., Horne, Z., Pinillos, N. Á., & Holyoak, K. J. (2015). A Bayesian framework for knowledge attribution: Evidence from semantic integration. *Cognition*, *139*, 92–104.

Prentice, D., & Miller, D. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science*, *16*(4), 202–206.

Reynaert, C., & Gelman, S. (2007). The influence of language form and conventional wording on judgments of illness. *Journal of Psycholinguistic Research*, *36*(4), 273–295.

Rhodes, M., Leslie, S., Yee, K., & Saunders, K. (2019). Subtle linguistic cues increase girls' engagement in science. *Psychological Science*, *30*(3), 455–466.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology*, *21*(4), 803–814.

Rouder, J. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308.

Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437–442.

Schoot, R., Depaoli, S., King, R., Kramer, B., Kaspers, M., & Tadesse, M. (2021). Van de Bayesian statistics and modelling. *Nature Review Methods Primers, 1*(1).

Sheehan, L., Nieweglowski, K., & Corrigan, P. (2016). The stigma of personality disorders. *Current Psychiatry Reports, 18*(11).

Substance Abuse and Mental Health Services Administration. (2018). *Key substance use and mental health indicators in the united states: Results from the 2017 national 1079 survey on drug use and health*. Rockville: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.

Sulin, R., & Dooling, D. (1974). Intrusion of a thematic idea in retention of prose. *Journal of Experimental Psychology*, *103*(2), 255–262.

Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, *3*(6), 375–387.

Tillmann, B., & Dowling, W. J. (2007). Memory decreases for prose, but not for poetry. *Memory & Cognition*, *35*(4), 628–639.

Time to Change (2019). About mental health. https://www.time-to-change.org.uk/about-mental-health

Van Dijk, T. A. (1980). The semantics and pragmatics of functional coherence in discourse. *Speech act theory: Ten years later* 49–65.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Wahl, O. (1999). Mental health consumers' experience of stigma. *Schizophrenia Bulletin*, *25*(3), 467–478.

Wodak, D., Leslie, S., & Rhodes, M. (2015). What a loaded generalization: Generics and social cognition. *Philosophy Compass*, 625–635.

Yzerbyt, V., Corneille, O., & Estrada, C. (2001). The interplay of subjective essentialism and entitativity in the formation of stereotypes. *Personality and Social Psychology Review*, *5*(2), 141–155.