Perceptions of the Competent but Depressed

Kristen Kim

Woo-kyoung Ahn

Yale University

Abstract

Accurately recognizing and remembering the depressive symptoms of other people can be crucial in helping those suffering from depression. Yet, lay theories about depression might interfere with accurate perception or recollection of depression in others. The current study examined whether laypersons would misremember depressive symptoms in highly competent people as being less severe than they actually are. Participants first read a target vignette about a character displaying depressive symptoms, while the level of competency of the target character varied across different conditions. Then, participants read a foil vignette describing a character with similar depressive symptoms, which was intended to elicit memory errors for the target vignette. When the foil vignette described that the depressive symptoms were eventually overcome, participants were more likely to false-alarm the recovery as the competent character's than as the less competent character's (Experiment 1-a). Conversely, when the foil vignette's depressive symptoms were described to be highly severe, participants were less likely to false-alarm them as the competent character's symptoms than as the less competent character's symptoms (Experiment 2-a). This phenomenon appears to be unique to laypeople's perception of depression, as the same pattern of results was not obtained when the participants were mental health clinicians (Experiments 1-b and 2-b) or when laypeople participants read about symptoms of physical disorders or other mental disorders (Experiment 3). Taken together, the current study presents novel findings suggesting that competent people's depression is under-detected by laypeople. The implications and the limitations of the study are discussed.

Keywords: depression, competence, lay theories, memory, emotion

Perceptions of the Competent but Depressed

Major depression is highly prevalent, affecting more than 16 million American adults in a given year, and is a leading cause of disability (National Institute of Mental Health, 2016). Although depression is highly treatable (Hollon, Thase, & Markowitz, 2002), the majority of people with the disorder neither seek nor receive proper treatment (e.g., Young, Klap, Sherbourne, & Wells, 2001). Thus, it would be helpful for someone who is suffering from depression if the people around them were able to accurately recognize their depressive symptoms so that they could provide social support and encourage them to seek treatment (e.g., Gullivers, Griffiths, & Christensen, 2010). Given the importance of accurate detection of other people's depression, this study examines factors that might cloud perception of this affectively-laden mental health disorder.

Many existing studies on emotion perception have focused on facial cues and expressions (e.g., Adolphs, 2002; Etcoff & Magee, 1992). Research on impaired emotion perception therefore has focused on similar constructs, such as examining how accurate facial emotion perception might be hindered by perceivers' mental illnesses such schizophrenia (Kohler, Walker, Martin, Healey, & Moberg, 2009).

Unlike these studies, the current study examines how background knowledge lay people have can impede accurate perceptions of other people's depressive symptoms, including highly disordered emotions, such as sadness and a lack of positive emotions. Previous studies have demonstrated that background knowledge, such as stereotypes about gender or ethnicity, can result in skewed perceptions of depression (Burr, 2002; Potts, Burnam, & Wells, 1991). For instance, despite identical presentation of depressive symptoms in male and female case vignettes, primary care physicians were found to diagnose the female versions with depression significantly more often (Stoppe, Sandholzer, Huppertz, Duwe, & Staedt, 1999). The current study examines how lay theories of competence can distort perceptions of depressive emotion. Specifically, the present study tests whether competent people would be misremembered as exhibiting less severe forms of depression than less competent people.

To illustrate the general idea, consider the articles commonly featured in the media whenever someone seemingly perfect took his or her own life. Detailing how these people appeared to have it all, the media puzzled over the apparent suicides of highly successful people like Kate Spade, an American fashion designer who created an iconic handbag line, or Anthony Bourdain, a celebrity chef renowned for his exploration of international cuisine. Even outside the limelight of fame, reporters wrestled with similar questions when Taylor Wallace, a handsome and popular football star at Columbia University, hanged himself during his first month in college (Cohen & Italiano, 2017). In a piece on the suicide of University of Pennsylvania student Madison Holleran, a seemingly nonsensical question was raised: “can a computer’s hard drive malfunction even if the screen isn’t scratched?” (Fagan, 2015).

It is not implausible to conjecture the existence of lay-theory that competent people are less likely to be depressed, as it seems to have some basis in reality. For example, people of higher socioeconomic status – an important predictor of perceived competence (Fiske, Xu, Cuddy, & Glick, 1999) – are less likely to be depressed compared to people of lower socioeconomic status (Lorant et al., 2003). When people feel more competent, they are happier and experience less negative affect (Sheldon, Ryan, & Reis, 1996). Those who were judged to be more competent by their peers also had lower self-reported levels of depression (Cole, Martin, & Powers, 1997). Because these associations exist in the real world, laypersons may have developed the belief that competence is linked to lower levels of depression.

Furthermore, this lay theory might exist because competent people may be less likely to express negative emotions. People in general try to suppress negative emotions in public (Jordan et al., 2011), and competent people might do so even more than average people. For instance, some top students at Stanford University are described to have Duck Syndrome, hiding their stress, depression, and anxieties behind a façade of perfection, like ducks gliding effortlessly on water while its feet are furiously struggling underneath the water (Scelfo, 2015). Interviews of people surrounding men who committed suicide after living apparently successful lives (e.g., owners of a company) revealed how the deceased were always "smiling and cheerful" and "super helpers" for others (p. 392; Kiamanesh, Dyregrov, Haavind, & Dieserud, 2014). Similarly, the aforementioned suicide of Kate Space is described as being "so out of character," caused by the "illness hidden with a smile" (Merkin, 2018). If competent people hide depression more than average people, laypeople may end up with the belief that competent people are happier.

Assuming that people indeed believe that competent people are generally less depressed for reasons discussed so far, the main question of the current study is what the consequences of having such a lay-theory would be. For instance, stereotypes, once developed, can distort one's perception of reality to be consistent with the stereotype (Darley & Gross, 1983). Similarly, we claim that if people hold a theory that competent people are less depressed in general, they will discount depressive symptoms even when competent people explicitly express depression.

Note that this lay theory may have arisen due to actual relationships between competence and depression as delineated above, and thus endorsing such a theory per se may not be an irrational bias. Such lay theories may prove helpful for conjecturing about a person when no specific details about her depression are available, and people have to make the best possible guess based on their background knowledge. However, one's theory about *general* cases can also cause biases in the perceptions of a *specific* competent person even when the person is explicitly showing signs of depression. For instance, if a competent person's complaints of seriously low mood for the past two weeks are downplayed as being less serious due to the lay theory about competent people's depression, then people would be making errors in the perception of negative emotions that can have critical clinical ramifications.

No previous studies have examined this issue, despite a large literature on the role of competence (as opposed to warmth) as a basic dimension of social perception (e.g., Fiske et al., 1999; Fiske, Cuddy, Glick, & Xu, 2002; Fiske, Cuddy, & Glick, 2007). The study closest to the current research question found that people with depression are generally perceived by laypersons as more competent compared to people with other mental disorders (e.g., schizophrenia; Fiske, 2012). However, the present study examines how competent people's depression is perceived compared to less competent people's depression.

In order to examine the consequence of holding a theory about competent people's depression, the present study tests whether people are more likely to falsely remember the signs of depression as being less severe when they appear in competent people compared to less competent people. This false memory is likely, given the numerous demonstrations of existing beliefs leading to false memory (e.g., Schacter, 1995, 1999). For instance, after reading a list of words from a common theme (e.g., tired, snooze, blanket), participants false-alarmed non-presented, but related words (e.g., rest; Roediger & McDermott, 1995). Stereotypes could also lead to biased memory (e.g., recalling more negative information about a defendant with a Hispanic name compared to a neutral name; Bodenhausen & Lichtenstein, 1987).

Similarly, we predicted that if laypersons believe that competent people are less likely to suffer from severe depression, *identical* symptoms of depression may be remembered as less severe if they are described as belonging to a competent person compared to the same symptoms described as belonging to a less competent person. Laypersons may also misremember a competent person's depression as being more *recovered* (i.e., improved) compared to less competent people's depression.

Overview of experiments

The current study utilized vignettes describing a character with symptoms of depression (e.g., "Lately, he has told his best friend that he has been feeling tired and a little 'blue'"). The competence of the character was manipulated to be either competent (e.g., "…known to be competent, responsible, organized, and efficient"), average (e.g., "…of ordinary intelligence, and reasonably organized, though not perfect"), or incompetent (e.g., "...known to be incompetent, irresponsible, disorganized, and inefficient"). After reading this target vignette, participants read a foil vignette, in which a new character displayed similar depressive symptoms. This foil vignette was inserted to elicit memory errors depending on the way the target vignette was perceived (see below for details). At the end of the study, participants received a surprise recognition test about the target vignette, which was the main dependent measure.

In Experiment 1a, the foil vignette described a character with recovered depressive symptoms (e.g., "However, for the past few days he has been trying to lift his mood, and has managed to cheer himself up."). It was hypothesized that lay participants would be more likely to confuse this recovery as the target character's when they read about a competent character than when they read about an average or an incompetent character. Experiment 1b tested mental health clinicians using the same stimuli to examine whether similar effects of competency would be obtained with those who were trained to have accurate memory for symptoms.

In Experiment 2a, the foil vignette described a character with more severe depressive symptoms (e.g., “Lately, he has told his best friend that he has been feeling exhausted and very ‘blue’”). It was hypothesized that lay participants who read about a competent character would be less likely to confuse *severe* depressive symptoms in the foil vignette with the target character’s symptom than those who read about an average or an incompetent character. Experiment 2b tested mental health clinicians using the same stimuli. Experiment 3 tested whether similar effects would be obtained when laypersons read about symptoms of other illnesses (e.g., anxiety, schizophrenia, physical illness), or alternatively whether lay-theories on the role of competence are limited to depression.

In each experiment, we used both female and male versions of vignette for generalizability. None of the critical effects interacted with the gender of the character in the vignette (see Supplemental Materials). Thus, all analyses reported were performed collapsed over the male and female versions of vignettes.

Participants in Experiments 1-a, 2-a, and 3 were recruited from Amazon.com’s Mechanical Turk in exchange for small monetary compensation. Participants in Experiments 1-b and 2-b were licensed mental health clinicians (e.g., psychologists or clinical social workers), recruited through addresses provided by Psychlist Marketing Inc., which obtains mailing lists of mental health professionals through various state agencies. Clinicians were recruited in numbers proportionate to the number of clinicians available through Psychlist in every U.S. state. Recruitment postcards directed them to a URL to complete the survey, and participating clinicians received $10 Amazon.com gift cards. No participant could participate in more than one experiment in this report. All research in this study was approved by the Yale Human Subjects Committee.

In Experiments 1a, 2a, and 3, we aimed to recruit 50 participants per condition for the following reasons. Bodenhausen and Lichtenstein (1987) found the effect of race stereotype on recall using 20-25 participants in each condition. The sample size we aimed for was larger than this study, because our participants were tested online rather than in lab settings, and we planned a priori to exclude those who failed the attention check (based on performance in the recognition task or the intermediate task) or the manipulation check (see Supplemental Materials for details). Thus, collecting 50 participants would allow for enough participants even after exclusions. We also considered the recent discussion that many psychology studies are underpowered (Bakker, Hartgerink, Wicherts, & van der Maas, 2016).

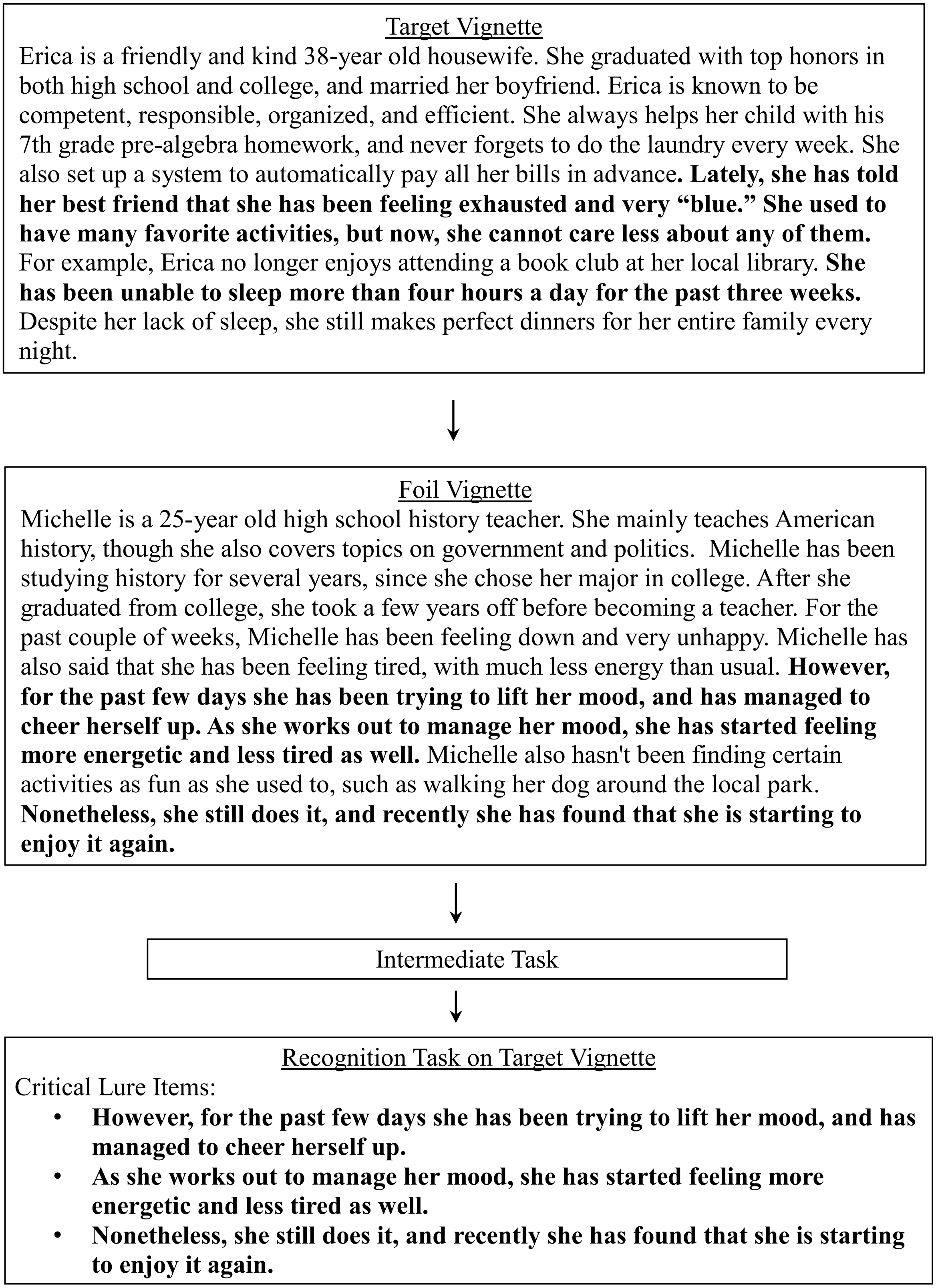
At the end of each experiment, we measured participants' perception of competence of the character in the target vignette, and these ratings confirmed that our manipulation of competence was valid. We also measured participants' perception of warmth of the target character, another important component of the stereotype content model (Fiske et al., 2002), and checked whether it was a confounding variable. Because warmth was significantly different across conditions in some experiments[1], it was included as a covariate in the analyses. The means and the statistical analyses of these ratings are included in the Supplemental Materials.

Experiment 1

After reading about a target character's depression, participants in Experiment 1 read about a foil character who had similar depressive symptoms but recovered from them. If participants believed that competent people are less likely to suffer from depression, then

[1] Warmth was a confound methodologically, in that it sometimes differed across conditions. However, the difference in the warmth ratings occurred mainly because the incompetent character was judged significantly less warm than the other two characters. The difference between the competent and the average conditions on warmth, which was the most critical comparison, was almost always not significant. That is, any effect found with competent characters compared to average characters could not be due to differences in perceived warmth, and it is difficult to imagine how warmth overall might have driven the observed results.

Figure 1. *An example of what a participant in the competent condition saw in progressing through Experiment 1a. The bolded font is used here to indicate depressive symptoms or the recovery of depressive symptoms, and was not presented to participants.*



Target Vignette

Erica is a friendly and kind 38-year old housewife. She graduated with top honors in both high school and college, and married her boyfriend. Erica is known to be competent, responsible, organized, and efficient. She always helps her child with his 7th grade pre-algebra homework, and never forgets to do the laundry every week. She also set up a system to automatically pay all her bills in advance**. Lately, she has told her best friend that she has been feeling exhausted and very "blue." She used to have many favorite activities, but now, she cannot care less about any of them.** For example, Erica no longer enjoys attending a book club at her local library. **She has been unable to sleep more than four hours a day for the past three weeks.** Despite her lack of sleep, she still makes perfect dinners for her entire family every night.

Foil Vignette

Michelle is a 25-year old high school history teacher. She mainly teaches American history, though she also covers topics on government and politics. Michelle has been studying history for several years, since she chose her major in college. After she graduated from college, she took a few years off before becoming a teacher. For the past couple of weeks, Michelle has been feeling down and very unhappy. Michelle has also said that she has been feeling tired, with much less energy than usual. **However, for the past few days she has been trying to lift her mood, and has managed to cheer herself up. As she works out to manage her mood, she has started feeling more energetic and less tired as well.** Michelle also hasn't been finding certain activities as fun as she used to, such as walking her dog around the local park. **Nonetheless, she still does it, and recently she has found that she is starting to enjoy it again.**

Intermediate Task

Recognition Task on Target Vignette

Critical Lure Items:

- **However, for the past few days she has been trying to lift her mood, and has managed to cheer herself up.**
- **As she works out to manage her mood, she has started feeling more energetic and less tired as well.**
- **Nonetheless, she still does it, and recently she has found that she is starting to enjoy it again.**

participants would be more likely to confuse the foil character's recovery as the target character's in the competent condition than in the less competent conditions.

Experiment 1a

Methods. Out of 304 participants recruited from Mechanical Turk only for this study, 258 participants remained after exclusions (Mean Age = 35.61, 41% Female; 62% White, 28% Asian, 4% Black, 6% Other). Supplemental Materials provide details of the exclusion criteria and the number of participants excluded for each criterion for all experiments.

As illustrated in Figure 1, participants first read a target vignette followed by a foil vignette. To increase generalizability, two versions of a target vignette were developed, one using a female character (a 38-year old woman named "Erica," see Table 1), and the other using a male character (a 32-year old man named "Eric," see Table 2), both described as "friendly and kind." Depending on the conditions, the target character was either competent, average, or incompetent, as shown in sentences 2, 3, 4, 5, 8, and 10 of Tables 1 and 2. These sentences were developed based on the stereotype content model (Fiske et al., 2002) along the key characteristics of perceived competence (e.g., intelligence, efficiency, organization, etc.). The number of words in each sentence was as closely matched as possible across the conditions. Because groups high in competence are usually perceived to be less warm (Cuddy, Fiske, & Glick, 2007), we were particularly mindful of the character's description in order to avoid such a confound (e.g., avoiding the use of a professional woman who might be associated with the negative stereotype of neglecting her family). Sentences 6, 7, and 9 in each version of the target vignette (see Tables 1 and 2) were depressive symptoms, which were identical across the three conditions. The symptoms were depressed mood, anhedonia, and insomnia, selected from the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013).

One-hundred thirty-two participants were randomly assigned to either competent (N = 44), average (N = 41), or incompetent (N = 47) condition of the female version, and a separate group of 126 participants were randomly assigned to either competent (N = 39), average (N = 41), or incompetent (N = 46) condition of the male version. Participants read the target vignette in their condition at their own pace, which was presented one sentence at a time.

Then, participants were presented with a foil vignette matching in the gender of the character in the target vignette. Within each gender condition, all participants read the same foil vignette regardless of the level of the target character's competency, and the competency of the character in the foil vignette was kept neutral. The character in the foil vignette displayed the same three depressive symptoms described in the target vignette, but showed recovery of depressive symptoms. Figure 1 shows the female version of the foil vignette, and the male version of the foil vignette is shown below. The bolded font indicating recovery of depressive symptoms (i.e., the sentences used as "critical lures"; see below) was not presented to the participants.

> Michael is a 25-year old high school history teacher. He mainly teaches American history, though he also covers topics on government and politics. Michael has been studying history for several years, since he chose his major in college. After he graduated from college, he took a few years off before becoming a teacher. For the past couple of weeks, Michael has been feeling down and very unhappy. Michael has also said that he has been feeling tired, with much less energy than usual. **However, for the past few days he has been trying to lift his mood, and has managed to cheer himself up. As he works out to manage his mood, he has started feeling more energetic and less tired as well.** Michael also hasn't been finding certain activities as fun as he used to, such as walking his dog around the local park. **Nonetheless, he still does it, and recently he has found that he is starting to enjoy it again.**

As with the target vignette, participants read the foil passage at their own pace, which was presented one sentence at a time.

Table 1. *Female versions of target vignettes used in Experiment 1a, varying in competency across columns. The bolded font is used here only to indicate depressive symptoms, and was not presented to the participants.*

Sentence Number	Competent	Average	Incompetent
1	Erica is a friendly and kind 38-year old housewife.	Erica is a friendly and kind 38-year old housewife.	Erica is a friendly and kind 38-year old housewife.
2	She graduated with top honors in both high school and college, and married her boyfriend.	She attended a public high school and graduated from a state university before marrying her boyfriend.	She barely graduated high school, and attended the local community college before marrying her boyfriend.
3	Erica is known to be competent, responsible, organized, and efficient.	Erica is of average intelligence, and reasonably organized, though not perfect.	Erica is known to be incompetent, irresponsible, disorganized, and inefficient.
4	She always helps her child with his 7th grade pre-algebra homework, and never forgets to do the laundry every week.	She helps her child with his 3rd grade math homework, and tries not to forget to do the laundry every week.	She doesn't know how to help her child with his 4th grade math homework, and frequently forgets to do the laundry.
5	She also set up a system to automatically pay all her bills in advance.	She usually pays her bills on time, though she has forgotten once or twice.	She always forgets to pay her bills on time, and has accumulated interest and fines.
6	**Lately, she has told her best friend that she has been feeling exhausted and very "blue."**		
7	**She used to have many favorite activities, but now, she cannot care less about any of them.**		
8	For example, Erica no longer enjoys attending a book club at her local library.	For example, Erica does not care about watching movies anymore.	For example, Erica no longer watches the home shopping network on TV.
9	**She has been unable to sleep more than four hours a day for the past three weeks.**		
10	Despite her lack of sleep, she still makes perfect dinners for her entire family every night.	Despite her lack of sleep, she still tries to make dinner for her family.	Because of her lack of sleep, she just orders takeout for her family every night.

Table 2. *Male versions of target vignettes used in Experiment 1a, varying in competence across conditions. The bolded font is used here only to indicate depressive symptoms, and was not presented to the participants.*

Sentence Number	Competent	Average	Incompetent
1	Eric is a friendly and kind 32-year old man.	Eric is a friendly and kind 32-year old man.	Eric is a friendly and kind 32-year old man.
2	He graduated with top honors in both high school and college, and married his girlfriend.	He attended a public high school and graduated from a state university before marrying his girlfriend.	He barely graduated high school, and attended the local community college before marrying his girlfriend.
3	Eric is known to be competent, responsible, organized, and efficient.	Eric is of ordinary intelligence, and reasonably organized, though not perfect.	Eric is known to be incompetent, irresponsible, disorganized, and inefficient.
4	Due to his exceptional performance, Eric has had a successful career for the past several years in upper-level management.	Eric has had the same job for the past several years in mid-level management and is considered a mostly unremarkable manager.	Despite frequent mistakes, Eric has managed to stay in a low-level management position for the local supermarket.
5	At home, he set up a system to automatically pay all his bills in advance.	At home, he usually pays his bills on time, though he has forgotten once or twice.	At home, he always forgets to pay his bills on time, and has accumulated interest and fines.
6	**Lately, he has told his best friend that he has been feeling exhausted and very "blue."**		
7	**He used to have many favorite activities, but now, he cannot care less about any of them.**		
8	For example, Eric does not care about reading his favorite classic novels anymore.	For example, Eric does not care about watching football games on TV anymore.	For example, Eric does not care about channel-surfing while sitting on his favorite couch anymore.
9	**He has been unable to sleep more than four hours a day for the past three weeks.**		
10	He still maintains his high performance in all of his work as before.	He still maintains his average performance in all of his work as before.	He continues to be a poor performer in all of his work as before.

Afterwards, participants received an intermediate task, in which they had to identify whether each of 20 pictures presented one at a time was a building or a house. Each response had to be made within 2 s to roughly equate the duration of this task across participants.

All participants then completed a recognition task. Participants were asked to remember the target vignette about Erica or Eric and to rate whether each of 13 sentences had appeared in this vignette on a scale from 1 (“Definitely True, or did appear”) to 6 (“Definitely False, or did not appear”). Seven sentences used in the recognition task were *studied* items (i.e., sentences presented in the target vignette; sentences 1, 2, 3, 4, 5, 8, and 10 in Table 1). Six were *lures* (i.e., sentences not presented in the target vignette). Three of the lures were *noncritical* lures, which were not related to the depressive symptoms (e.g., “Erica has two children”), and the other 3 were *critical* lures, which were about recovery of the depressive symptoms presented in the foil vignette, but not in the target vignette (i.e., bolded sentences in the foil vignette in Figure 1 and above). For all participants, the studied (S), noncritical (N), and critical (C) items were presented in the following order: S, S, S, N, C, N, S, C, S, N, S, C, S. The 3 studied items were presented first in a row to orient the participants to remembering the target character rather than foil character. We excluded data from participants who failed on accurately rating more than half of the studied and noncritical lure items[2], suggesting that they did not read the target vignette carefully and/or they might have been confused about which vignette this recognition test was for (see Supplemental Materials for the number of excluded participants for all experiments).

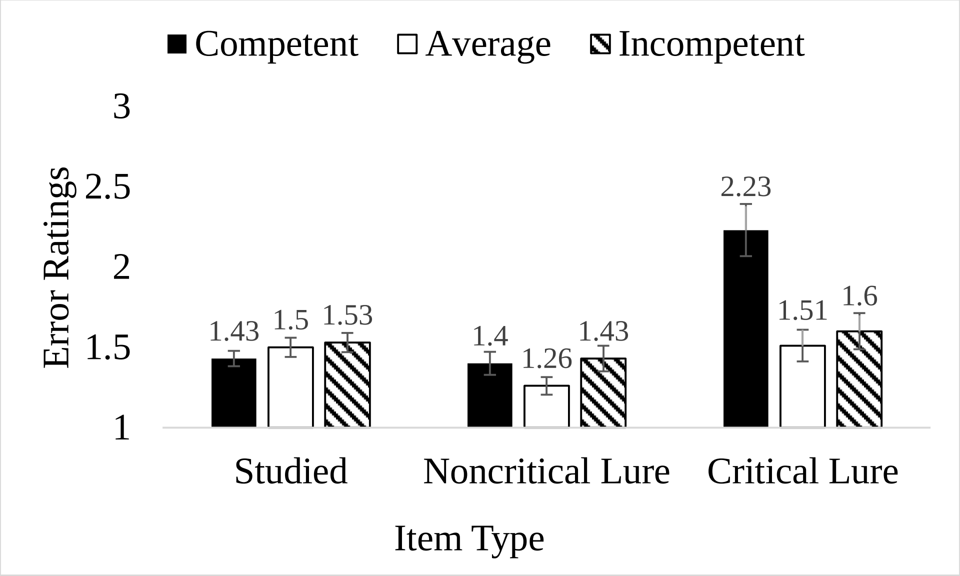
Then, the manipulation check was administered as described in the “Overview of Experiments” section to check the successful manipulation of competency and to control for any

[2] An inaccurate rating is defined as a rating of 4 or higher for the studied items, or 3 or lower for the noncritical lures.

variance in warmth across the vignettes in the subsequent analyses. Lastly, demographic information was collected.

Results and Discussion. To analyze the data from the recognition task, the ratings for each item type (i.e., studied, noncritical lure, and critical lure items) were first averaged within each participant. The ratings for the noncritical and critical lures were reverse-coded, so that higher scores would indicate greater errors in all three item types (i.e., misses for the studied items and false alarms for the lures). These scores are termed *error ratings* henceforth. Figure 2 shows mean error ratings broken down by condition and item type. The Supplemental Materials report estimated marginal means adjusted for warmth as a covariate for all experiments.

Figure 2. *Mean error ratings in the recognition task of Experiment 1a broken down by condition and item type. Error bars are ±1 standard error from the mean.*



A 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (conditions; competent, average, incompetent) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate found no main effect of item type, $F(1.56, 394.86) = .23$, $p = .74$, $\eta_p^2=.001$, and a significant main effect of condition, $F(2, 254) = 6.20$, $p = .002$, $\eta_p^2= .05$. This significant main effect is qualified by a significant interaction effect between item type and condition, $F(3.11, 394.86) = 7.52$, $p < .001$, $\eta_p^2 = .06$.

To understand the pattern of this interaction effect, one-way ANOVAs testing the effect of condition were performed for each item type. There was no significant effect of condition for the studied items, $F(2, 254) = .46$, $p = .63$, $\eta_p^2 = .004$, or the noncritical lures, $F(2, 254) = 1.50$, $p = .22$, $\eta_p^2 = .01$, but there was a significant effect of condition for the critical lures, $F(2, 254) = 9.68$, $p < .001$, $\eta_p^2 = .07$. Post hoc comparisons[3] using Bonferroni corrections for the critical lures showed that the error ratings for the competent condition were significantly higher than those for the average condition, $p < .001$, as well as significantly higher than those for the incompetent condition, $p = .001$. The error ratings for the critical lures in the average condition were not significantly different from those in the incompetent condition, $p = .99$.

These results therefore supported our hypotheses, showing that laypeople participants were more likely to confuse the foil recovery symptoms with the target vignette in the competent condition than in the less competent conditions. No other item types were remembered differently across the conditions, and the effect was unique to the competent condition, showing no difference between the average and the incompetent condition.

Experiment 1b

Experiment 1b tested whether mental health clinicians would show similar biases. On the one hand, previous studies have found that mental health clinicians are often strikingly similar to laypeople in the accuracy of their clinical judgements (e.g., diagnoses and treatment) (Christensen & Jacobsen, 1994; Ebling & Levenson, 2003), in which case the same effects might be obtained with clinicians. On the other hand, given that relevant education and training have

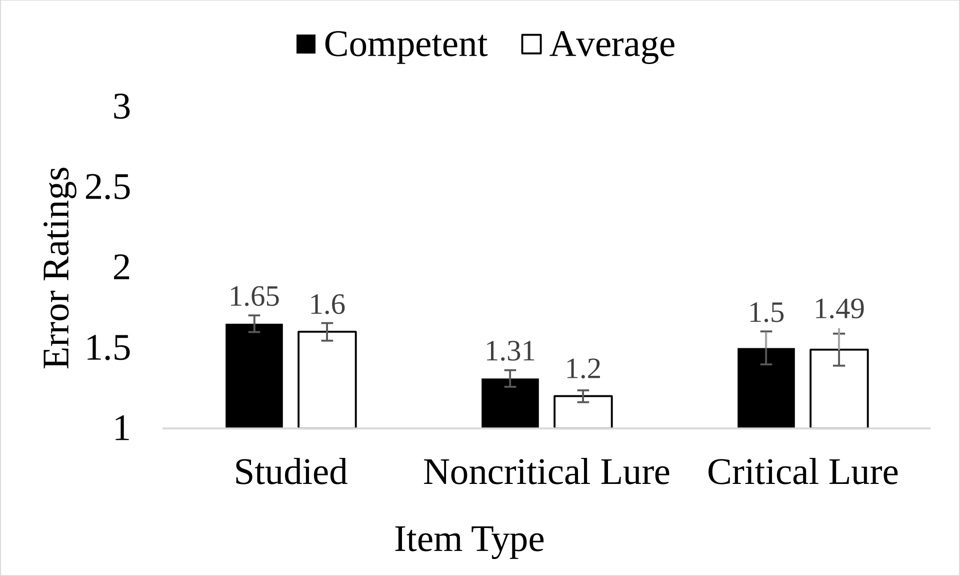
[3] All post-hoc analyses, both here and elsewhere, were based on the estimated marginal means reported in the Supplemental Materials. The patterns of the results did not change even when using the raw means.

been found to help improve diagnostic accuracy (Lambert & Wertheimer, 1988), as well as the likelihood that clinicians may have more experiences with competent people suffering from depression, they may not show the effect of competence.

Methods. Out of 243 clinicians recruited through Psychlist only for this study, 215 participants remained after exclusions (83% Female; 83% White, 1% Asian, 8% Black, 8% Other). The methods were the same as those in Experiment 1a, except that participants were clinicians, and that only the competent and average conditions were used. Because clinician participants were more challenging to recruit, we reduced the number of participants by dropping the incompetent condition, since Experiment 1a did not find any difference between the incompetent and the average conditions. Participants were randomly assigned to either female competent (N = 47), female average (N = 54), male competent (N = 58), or male average (N = 56) conditions. See Supplemental Materials for details of demographic information.

Results and Discussion. We were mainly interested in the difference between laypeople and clinicians. Thus, error ratings were analyzed using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (conditions; competent, average) x 2 (participant type; laypeople from Experiment 1a, clinicians from Experiment 1b) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. There was a significant 3-way interaction effect, $F(1.54, 576.35) = 8.88$, $p = .001$, $\eta_p^2 = .02$. As suggested by Figure 3, which shows the clinicians' error ratings for each item type for each condition, this significant 3-way interaction effect appears to be obtained because unlike in Experiment 1a, the clinician participants' error ratings on different item types did not vary across the two conditions.

Figure 3. *Mean error rating in the recognition task of Experiment 1b broken down by item type and condition. Error bars are ±1 standard error from the mean.*



Thus, we examined the clinicians' error ratings using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (condition; competent, average) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. (See the Supplemental Materials, S9 for the corresponding analyses for laypeople.) Indeed, there was no significant interaction effect, $F(1.64, 346.53) = .57$, $p = .53$, $\eta_p^2 = .003$. The main effect of the item type was significant, $F(1.64, 346.53) = 3.77$, $p = .03$, $\eta_p^2 = .017$, because the noncritical lures, which were easier to identify as absent, led to lower error ratings than the studied and the critical lures. The main effect of condition was not significant, $F(1, 212) = .78$, $p = .38$, $\eta_p^2 = .004$.

These results thus suggest that compared to laypersons, clinicians are less likely to assume that competent people are less depressed. Clinicians did not significantly differ in their recognition accuracy depending on varying levels of competence of the character. Nonetheless, null effects are always difficult to interpret. For instance, it is possible that clinicians might still believe that competent people are less depressed but they might have needed stronger manipulations to reveal the effect. At the very least, the significant interaction effect found in

Experiment 2a involving laypeople vs. clinicians and the experimental manipulations suggest that clinicians are less likely to show the effect of competence than laypeople. In the General Discussion section, we speculate reasons for the difference between laypeople and clinicians.

Experiment 2

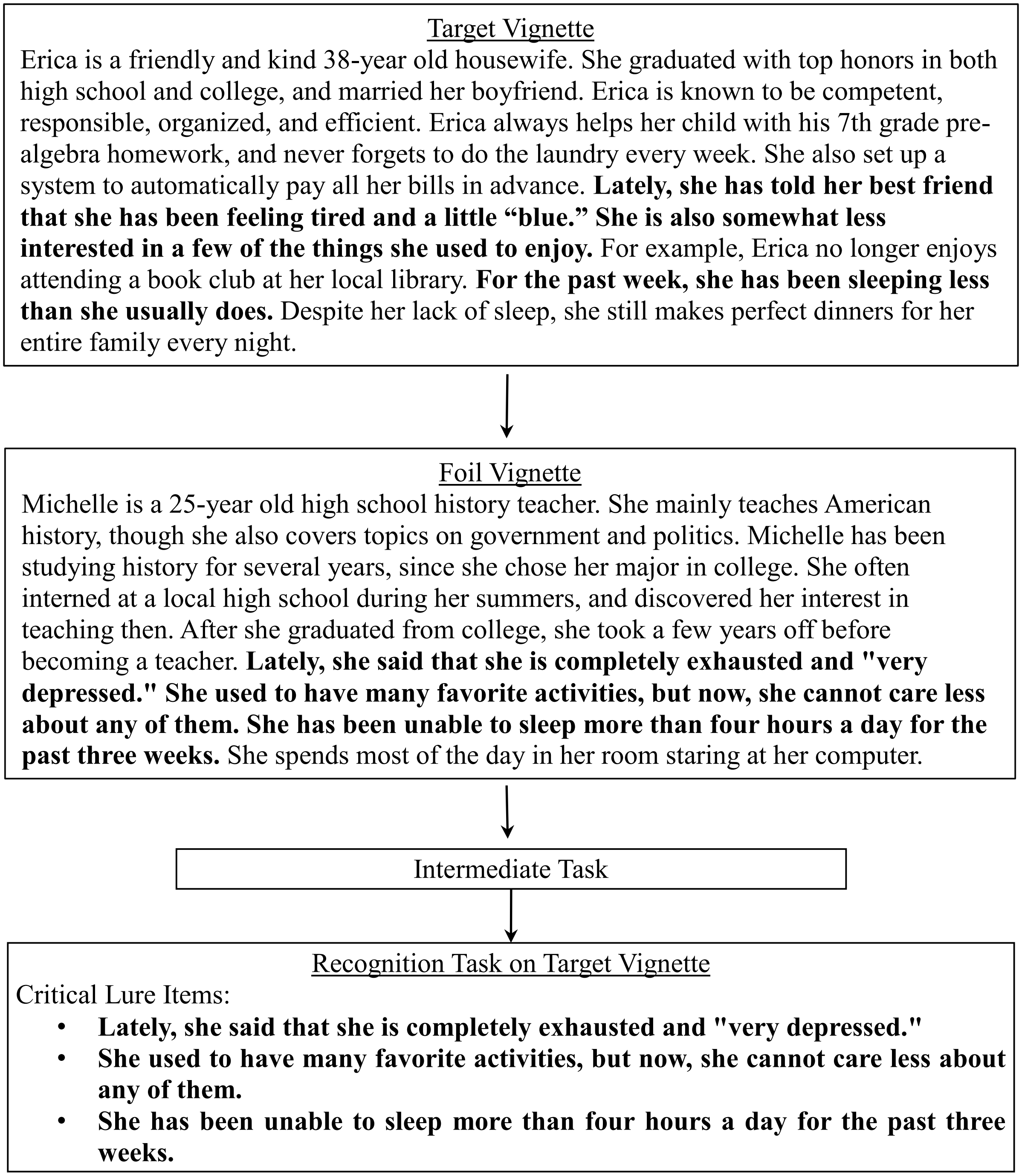
Experiment 1a showed that laypeople were more likely to misremember highly competent people's symptoms as having been recovered than average or incompetent people's symptoms. We designed Experiment 2 to show that highly competent people's symptoms could also be *less* likely to be misremembered as *more* severe. Participants again read a target vignette and a foil vignette, but unlike in Experiment 1, after reading about a target character's less severe depression, participants in Experiment 2 read about a foil character with more severe depression. If participants believed that competent people are less likely to suffer from severe depression, then participants would be less likely to confuse the more severe foil symptoms as the target character's symptoms in the competent condition than in the less competent conditions. Experiment 2a tested lay participants recruited from Mechanical Turk, while Experiment 2b tested clinicians.

Experiment 2a

Methods. Out of 302 participants recruited from Mechanical Turk only for this study, 258 participants remained after exclusions (Mean Age = 36.79, 42% Female; 64% White, 19% Asian, 8% Black, 9% Other). The methods of Experiment 2a were the same as in Experiment 1a except for the following changes. The depressive symptoms in the target vignette were less severe versions of the three depressive symptoms used in Experiment 1a (see Sentences 6, 7, and 9 of Tables 3 and 4 for the actual symptoms used for the female and male versions, respectively). The foil vignette had three depressive symptoms that are more severe than the symptoms in the target vignette. Figure 4 shows these symptoms for the female version. The male version of the foil vignette was also developed using the same background information from the male foil vignette in Experiment 1a. The depressive symptoms used were the same severe depressive symptoms as those of the female foil vignette (Figure 4).

One-hundred thirty-four participants were randomly assigned to either competent (N = 46), average (N = 44), or incompetent (N = 44) condition for the female version, and a separate group of 124 participants were randomly assigned to either competent (N = 40), average (N = 41), or incompetent (N = 43) condition for the male version.

Figure 4. *An example of what a participant in the competent condition saw in progressing through Experiment 2a. The bolded font is used here to indicate depressive symptoms, and was not presented to participants.*



Target Vignette

Erica is a friendly and kind 38-year old housewife. She graduated with top honors in both high school and college, and married her boyfriend. Erica is known to be competent, responsible, organized, and efficient. Erica always helps her child with his 7th grade pre-algebra homework, and never forgets to do the laundry every week. She also set up a system to automatically pay all her bills in advance. **Lately, she has told her best friend that she has been feeling tired and a little "blue." She is also somewhat less interested in a few of the things she used to enjoy.** For example, Erica no longer enjoys attending a book club at her local library. **For the past week, she has been sleeping less than she usually does.** Despite her lack of sleep, she still makes perfect dinners for her entire family every night.

Foil Vignette

Michelle is a 25-year old high school history teacher. She mainly teaches American history, though she also covers topics on government and politics. Michelle has been studying history for several years, since she chose her major in college. She often interned at a local high school during her summers, and discovered her interest in teaching then. After she graduated from college, she took a few years off before becoming a teacher. **Lately, she said that she is completely exhausted and "very depressed." She used to have many favorite activities, but now, she cannot care less about any of them. She has been unable to sleep more than four hours a day for the past three weeks.** She spends most of the day in her room staring at her computer.

Intermediate Task

Recognition Task on Target Vignette

Critical Lure Items:

- **Lately, she said that she is completely exhausted and "very depressed."**
- **She used to have many favorite activities, but now, she cannot care less about any of them.**
- **She has been unable to sleep more than four hours a day for the past three weeks.**

Table 3. *The three female versions of target vignettes used in Experiment 2a, varying in competency across columns. The bolded font is used here only to indicate depressive symptoms, and was not presented to the participants.*

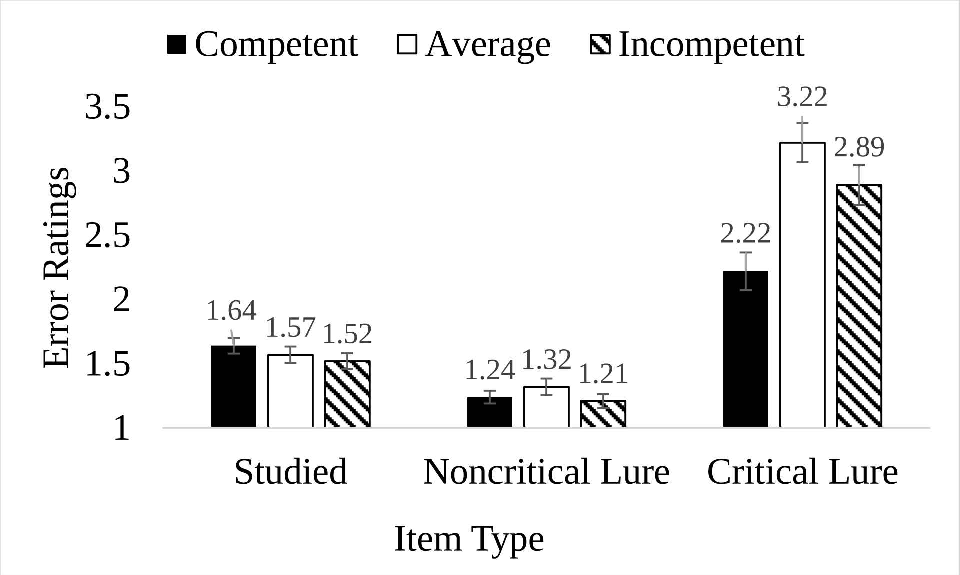
Sentence Number	Competent	Average	Incompetent
1	Erica is a friendly and kind 38-year old housewife.	Erica is a friendly and kind 38-year old housewife.	Erica is a friendly and kind 38-year old housewife.
2	She graduated with top honors in both high school and college, and married her boyfriend.	She attended a public high school and graduated from a state university before marrying her boyfriend.	She barely graduated high school, and attended the local community college before marrying her boyfriend.
3	Erica is known to be competent, responsible, organized, and efficient.	Erica is of average intelligence, and reasonably organized, though not perfect.	Erica is known to be incompetent, irresponsible, disorganized, and inefficient.
4	She always helps her child with his 7th grade pre-algebra homework, and never forgets to do the laundry every week.	She helps her child with his 3rd grade math homework, and tries not to forget to do the laundry every week.	She doesn't know how to help her child with his 4th grade math homework, and frequently forgets to do the laundry.
5	She also set up a system to automatically pay all her bills in advance.	She usually pays her bills on time, though she has forgotten once or twice.	She always forgets to pay her bills on time, and has accumulated interest and fines.
6	**Lately, she has told her best friend that she has been feeling tired and a little "blue."**		
7	**She is also somewhat less interested in a few of the things she used to enjoy.**		
8	For example, Erica no longer enjoys attending a book club at her local library.	For example, Erica does not care about watching movies anymore.	For example, Erica no longer watches the home shopping network on TV.
9	**For the past week, she has been sleeping less than she usually does.**		
10	Despite her lack of sleep, she still makes perfect dinners for her entire family every night.	Despite her lack of sleep, she still tries to make dinner for her family.	Because of her lack of sleep, she just orders takeout for her family every night.

Table 4. *The three male versions of target vignettes used in Experiment 2a, varying in competency across columns. The bolded font is used here only to indicate depressive symptoms, and was not presented to the participants.*

Sentence Number	Competent	Average	Incompetent
1	Eric is a friendly and kind 32-year old man.	Eric is a friendly and kind 32-year old man.	Eric is a friendly and kind 32-year old man.
2	He graduated with top honors in both high school and college, and married his girlfriend.	He attended a public high school and graduated from a state university before marrying his girlfriend.	He barely graduated high school, and attended the local community college before marrying his girlfriend.
3	Eric is known to be competent, responsible, organized, and efficient.	Eric is of ordinary intelligence, and reasonably organized, though not perfect.	Eric is known to be incompetent, irresponsible, disorganized, and inefficient.
4	Due to his exceptional performance, Eric has had a successful career for the past several years in upper-level management.	Eric has had the same job for the past several years in mid-level management and is considered a mostly unremarkable manager.	Despite frequent mistakes, Eric has managed to stay in a low-level management position for the local supermarket.
5	At home, he set up a system to automatically pay all his bills in advance.	At home, he usually pays his bills on time, though he has forgotten once or twice.	At home, he always forgets to pay his bills on time, and has accumulated interest and fines.
6	**Lately, he has told his best friend that he has been feeling tired and a little "blue."**		
7	**He is also somewhat less interested a few of the things he used to enjoy.**		
8	For example, Eric does not care about reading his favorite classic novels anymore.	For example, Eric does not care about watching football games on TV anymore.	For example, Eric does not care about channel-surfing while sitting on his favorite couch anymore.
9	**For the past week, he has been sleeping less than he usually does.**		
10	He still maintains his high performance in all of his work as before.	He still maintains his average performance in all of his work as before.	He continues to be a poor performer in all of his work as before.

Results and Discussion. The error ratings were analyzed using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (condition; competent, average, incompetent) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. There was no main effect of item type, $F(1.35, 343.05) = 2.71$, $p = .09$, $\eta_p^2 = .01$, and a significant main effect of condition, $F(2, 254) = 6.96$, $p < .001$, $\eta_p^2 = .05$. This significant main effect is qualified by a significant interaction effect between item type and condition, $F(2.70, 343.05) = 16.75$, $p < .001$, $\eta_p^2 = .12$ (see Figure 5 for the error ratings for each item type and for each condition).

Figure 5. *Mean error rating in the recognition task of Experiment 2a. Error bars are ±1 standard error from the mean.*



To understand the pattern of this interaction effect, one-way ANOVAs testing the effect of condition were performed for each item type. There was a significant effect of condition for the critical lure item type, $F(2, 254) = 15.26$, $p < .001$, $\eta_p^2 = .11$. Post hoc comparisons using Bonferroni corrections for the critical lures showed that as predicted, the error ratings for the competent condition were significantly lower than the error ratings for the average condition, $p < .001$, as well as those for the incompetent condition, $p < .001$. The error ratings for the critical lures in the average condition were not significantly different from those in the incompetent condition, *p* = .99. The lower error ratings for the competent condition are unlikely due to overall better memory for the competent character's vignette. It is because although there was a significant effect of condition for the studied items, *F*(2, 254) = 4.11, *p* = .02, η_p^2= .03, post hoc comparisons using Bonferroni corrections for the studied items showed that the error ratings for the competent condition were actually significantly higher than those for the incompetent condition, *p* = .01. The error ratings for the average condition did not differ from the competent condition, *p* = .73, nor the incompetent condition, *p* = .18. There was no significant effect of condition for the noncritical lures, *F*(2, 254) = 1.14, *p* = .32, η_p^2= .009.

These results therefore supported our hypotheses, showing that laypeople participants were less likely to confuse the more severe symptoms in the foil vignette with the target vignette in the competent condition than in the less competent conditions. Moreover, the effect appeared to be unique to the competent condition, showing no difference between the average and the incompetent condition.[4]

[4] One possible alternative explanation that one may argue for the results is that given that the error ratings for critical lures in the competent condition is similar between Experiments 1a and 2a and that the error ratings for the average/incompetent conditions more significantly differ between the studies, those with the competent targets may perhaps be unaffected by the directionality of the foil's symptoms. Rather, perhaps participants with the average/incompetent targets showed change, such that they ignored the foil when it showed recovery but were affected by the foil when it was more severe. However, there are several problems with this interpretation. First, comparing the results across Experiments 1a and 2a involves comparing memory results from different items, and such comparison is not warranted if the difficulty of remembering these items is not equated. Indeed, it appears that rejecting critical lures in Experiment 1a appears easier than rejecting critical lures in Experiment 2a. Recovering from depression is a categorical change (abnormal to normal) whereas more severe depression is differences in a continuum, which is much more easily confusable. Therefore, it was much easier for participants to reject the recovery from depression (Experiment 1a) than it was to reject more severe depression (Experiment 2a). Given this difference in difficulty between the two tasks, it seems more appropriate to compare between the conditions within the same task, rather than comparing the performances across the tasks. Second, there is a theoretical difficulty of

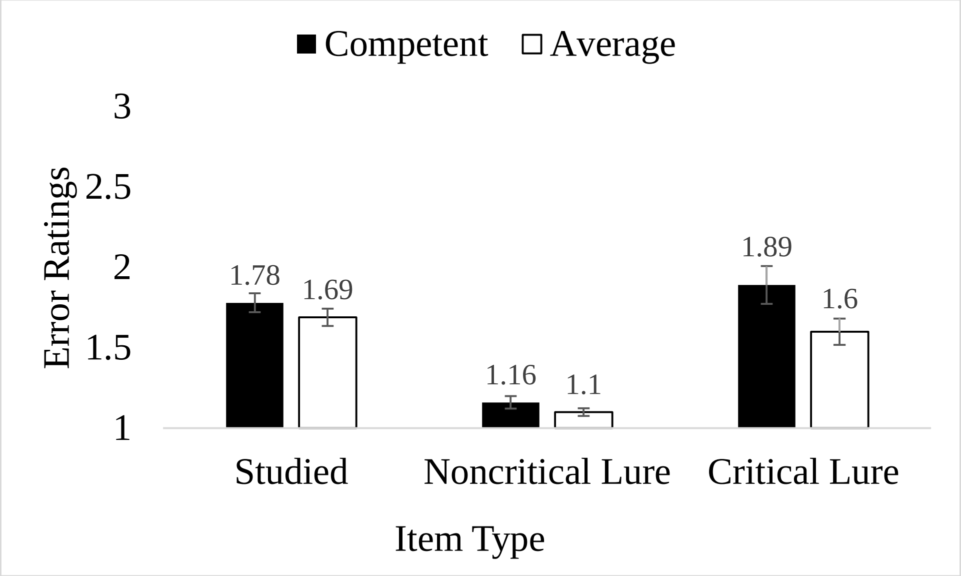
Experiment 2b

Methods. Out of 248 clinician participants recruited from Psychlist only for this study, 228 participants remained after exclusions (83% Female; 82% White, 6% Asian, 5% Black, 7% Other). The methods were the same as those in Experiment 2a, except that participants were currently practicing, licensed clinicians and that only the competent and average conditions were used. Participants were randomly assigned to either the female competent (N = 55), female average (N = 65), male competent (N = 57), or male average conditions (N = 51).

Results and Discussion. To check the difference between laypeople and clinicians, we analyzed error ratings using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (conditions; competent, average) x 2 (participant type; laypeople from Experiment 2a, clinicians from Experiment 2b) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. There was a highly significant 3-way interaction effect, $F(1.46, 573.43) = 21.35$, $p < .001$, $\eta_p^2 = .05$. As suggested by Figure 6, which shows the clinicians' error ratings for each item type for each condition, this significant 3-way interaction effect appears to be obtained because the clinician participants' error ratings on different item types did not vary across the two conditions.

concluding that these results are because of laypeople's beliefs about average and competent people's depression rather than about competent people's depression, as it is not clear why laypeople would inflate the severity of not only incompetent people's but also average people's depression. That is, similar to our theory that laypeople believe that competent people are better at managing depression, perhaps they believe that incompetent people are more likely to suffer from depression, but it is unclear why they would believe that even average people should suffer from severe depression.

Figure 6. *Mean error rating in the recognition task of Experiment 2b. Error bars are ±1 standard error from the mean.*



Thus, we examined the clinicians' error ratings using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (conditions; competent, average) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. (See the Supplemental Materials, S9 for the corresponding analyses for laypeople.) Indeed, there was no significant interaction effect, *F*(1.54, 347.49) = 1.74, *p* = .19, η_p^2= .008. The main effect of the item type was not significant, *F*(1.54, 347.49) = 2.91, *p* = .07, η_p^2= .01. The main effect of condition was significant, *F*(1, 225), = 6.20, *p* = .01, η_p^2= .03, because error ratings in the competent condition were higher than error ratings in the average condition, which is, if anything, in the opposite direction to the laypeople's results where the critical lure's error ratings were lower in the competent condition. Thus, this difference only serves to emphasize that clinicians were not affected by the vignettes across conditions in the same way that laypersons were.

These results thus suggest that once again, clinicians do not appear to assume that competent people are less severely depressed. Unlike with laypersons, clinicians' error ratings on the critical lures were not significantly lower in the competent condition compared to the average condition.

Experiment 3

Finally, Experiment 3 tested whether the effects of competence would extend to illnesses other than depression. If this memory bias is a result of the halo effect (e.g., Nisbett & Wilson, 1977), for instance, with competent people's positive characteristics influencing laypersons' judgments of all their qualities, then we might expect similar effects with other illnesses. Alternatively, this memory bias may be unique to depression, since other disorders may not have this type of association with levels of competence in reality. Anxiety, for example, might actually be more severe among competent people due to high expectations and pressure (Kiamanesh, Dyregrov, Haavind, & Dieserud, 2014). Other mental disorders like schizophrenia are believed to be largely biologically based (Read, Mosher, & Bentall, 2004) and more treatable by medications than psychotherapy (Kuppin & Carpiano, 2008), and thus laypeople may believe that one is unlikely to overcome these disorders by being competent. Similarly, it is probably unreasonable to believe that someone can overcome physical illness, such as fever, by being competent. In order to test the scope of the competency effect, we conducted the same procedure as in Experiment 2a, only replacing the depressive symptoms with symptoms of either anxiety, schizophrenia, or physical illness.

Table 5. *Less severe version of symptoms used in the target vignette, alongside more severe symptoms used in the foil vignette of Experiment 3.*

Symptom Number and Name	Symptoms used in the Target vignette	Symptoms used in the Foil Vignette and as Critical Lures
	Anxiety	
S1. Restlessness	Lately, she has told her best friend that she has been feeling a little restless, often shaking her leg whenever she is sitting down.	Recently, she has told her husband that she has been feeling very "on edge," constantly fidgeting and moving around.
S2. Excessive worry	She also noticed that she has been worrying about her work and family more than usual.	She has also been constantly worrying about everything, including her job, friends, and children.
S3. Irritability	She has been somewhat irritated by her colleagues at work as well.	She has been highly irritated by her colleagues at work as well.
	Schizophrenia	
S1. Hallucinations	Lately, she has told her best friend that she has been hearing voices in her head about once a week.	Recently, she has told her husband that she has been hearing voices in her head about every other day.
S2. Disorganized speech	Also, her speech has changed, and she occasionally blurts out disorganized and somewhat incoherent sentences.	She frequently also speaks incomprehensibly, babbling made-up words and not completing her sentences.
S3. Delusions	She also started suspecting that her boss is bugging her phone calls and secretly listening to her conversations.	She also started thinking that the director of the CIA is bugging her phone calls and secretly spying on her.
	Physical Illness	
S1. Joint pain	Lately, she has told her best friend that she has been having minor joint pain in her elbows.	Recently, she told her husband that she has been having severe joint pain in her elbows and knees.
S2. Digestive problems	She has also been experiencing some digestive issues about once a week.	She has also been experiencing major digestive issues nearly every day.
S3. Fever	For last 3 days, she also has had a low-grade fever.	For the past 5 days, she also has been running a fairly high fever.

Methods

Out of 926 participants recruited from Mechanical Turk only for this study, 823 participants remained after exclusions (Mean Age = 35; 43% Female; 56% White, 30% Asian, 5% Black, 9% Other). The methods were the same as in Experiment 2a except for the following changes. Three new sets of materials were created, using symptoms of anxiety (restlessness, excessive worry, irritability; American Psychiatric Association, 2013), schizophrenia (hallucinations, disorganized speech, delusions; American Psychiatric Association, 2013), or physical illness (joint pain, digestive problems, a fever). A less severe version was used for target vignettes, and a more severe version was used for foil vignettes and the critical lures of the recognition test (see Table 5).

In addition, the target vignettes' sentence 8 (Table 1) was replaced with a sentence that corresponded to the hobby and the symptom of the character depending on the condition. For instance, the female versions were; "For example, while [discussing her favorite classic novel with her friend / discussing a recent movie with her friend / talking about the home shopping network with her friend], she suddenly [started worrying about her family's health / began to speak in a confused and disjointed way / had minor stomach cramps and pains]" where the first bracketed phrase shows differences among the competent, average, and incompetent condition, respectively, and the second bracketed phrase shows differences among anxiety, schizophrenia, and physical illness, respectively. This new sentence 8 was also used as one of the studied items in the recognition test. Participants were randomly assigned to one of the three disorders, and then randomly assigned again to one of the 3 competency conditions (N = 88, 90, and 89 for anxiety, N = 101, 78, and 94 for schizophrenia, and N = 91, 103, and 89 for physical illness in the competent, average, and incompetent conditions, respectively).

Results and Discussion

Table 6 shows means broken down by the item type and conditions for each of the disorder condition. For each of the disorder conditions, we analyzed the error ratings using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (conditions; competent, average, incompetent) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. As summarized in Table 7, there were no significant interaction effects between the item type and the conditions for anxiety or schizophrenia. There was a significant interaction effect for physical illness. However, one-way ANOVAs revealed that this significant interaction effect for physical illness was obtained because there was a significant effect of condition on the noncritical lures[5], $F(2, 279) = 4.24$, $p = .02$, $\eta_p^2 = .03$, while there was no significant effect of condition on the studied items, $F(2, 279) = 2.37$, $p = .10$, $\eta_p^2 = .02$, or most importantly, on the critical lures, $F(2, 279) = 1.17$, $p = .31$, $\eta_p^2 = .008$.

Thus, the current study using symptoms of anxiety, schizophrenia, and physical illness found no evidence suggesting that participants were less likely to false alarm on severe symptoms in the competent condition than the less competent conditions. Nonetheless, these null effects may be due to biases in symptom selection. That is, while the false alarm rates on critical lures may not have been significantly different across the conditions when these items were averaged together, some of these items may have been. Thus, we additionally ran one-way ANOVAs on each of the individual critical lures, as reported in Supplemental Materials, and there was no significant effect of condition on any of the critical lures. These results therefore present fairly solid evidence that the effect of competence is unlikely to occur with Anxiety,

[5] Given the identical noncritical lures were not significantly different across the conditions for any other experiment, this variation is likely due to chance error.

Table 6. *Means and standard deviations for each item type across conditions in Experiment 3.*

Item Type	Competent	Average	Incompetent	Overall
		Anxiety		
Studied	1.69 (.71)	1.61 (.60)	1.46 (.54)	1.59 (.62)
Noncritical Lures	1.58 (.93)	1.53 (.91)	1.45 (.67)	1.52 (.84)
Critical Lures	3.72 (1.43)	3.93 (1.41)	3.96 (1.42)	3.87 (1.42)
		Schizophrenia		
Studied	1.52 (.60)	1.61 (.66)	1.63 (.59)	1.58 (.62)
Noncritical Lures	1.58 (.80)	1.52 (.73)	1.59 (.75)	1.57 (.76)
Critical Lures	3.04 (1.42)	2.95 (1.52)	3.27 (1.42)	3.09 (1.45)
		Physical Illness		
Studied	1.54 (.62)	1.56 (.61)	1.56 (.62)	1.55 (.61)
Noncritical Lures	1.69 (.72)	1.52 (.68)	1.43 (.64)	1.55 (.69)
Critical Lures	2.49 (1.36)	2.63 (1.50)	2.75 (1.48)	2.62 (1.45)

Table 7. *Analyses of variance results, with item type as the within-subject variable and condition as the between subject variable, in Experiment 3.*

Effect	F-statistics	p	η_p^2
	Anxiety		
Item Type	*F*(1.48, 389.89) = 17.63	<.001	.06
Condition	*F*(2, 263) = 2.44	.09	.02
Item Type X Condition	*F*(2.97, 389.89) = 1.03	.38	.008
	Schizophrenia		
Item Type	*F*(1.56, 420.34) = 16.93	<.001	.06
Condition	*F*(2, 269) = .58	.56	.004
Item Type X Condition	*F*(3.13, 420.34) = .42	.75	.003
	Physical Illness		
Item Type	*F*(1.42, 396.58) = .69	.46	.002
Condition	*F*(2, 279) = .11	.90	.001
Item Type X Condition	F(2.84, 396.58) = 3.18	.03	.02

Schizophrenia, and other physical symptoms.

General Discussion

We hypothesized that people believe that highly competent people are less likely to be depressed, perhaps due to real-world associations between competence and depression (e.g., Cole, et al., 1997; Sheldon, et al., 1996). These lay theories may lead to bias in remembering people's depressive symptoms even when they are explicitly and concretely stated. We presented participants with two vignettes, whose characters displayed identical symptoms of depression but varied in their levels of perceived competence. In Experiment 1-a, participants were more likely to false alarm highly competent people as having recovered from their depression compared to less competent people. These results are unlikely to be obtained simply because people have worse memory for competent vignettes in general, because the effect was obtained only with depressive symptoms, and not with other features that were present or absent in the vignette.

Experiment 2-a provided converging evidence of these memory biases regarding competent people with a converse pattern of results. After reading a vignette about a character with depressive symptoms, participants who read about a competent person were *less* likely to confuse the more severe symptoms as belonging to the character compared to participants who read about an average or an incompetent person.

Notably, in both experiments all of the differences in memory errors were limited to high levels of perceived competence in the vignette characters. That is, lay participants did not believe that less competent people would be more depressed than an average person. Rather, lay theories linking depression and competence appear to specifically posit that *high* levels of competency equal less depression.

We also examined whether these biases are due to laypeople simply generalizing competence in one domain to many others across a person’s entire life. If this type of halo effect were driving these differences, then it would persist regardless of what disorder was in question. However, Experiment 3 demonstrated that the effect of perceived competency did not extend to physical illness, or to two other mental disorders, anxiety and schizophrenia. While there are many potential reasons for why this effect did not occur with other disorders (e.g., anxiety’s potentially increased severity among competent people, as discussed above) that have yet to be explored in detail, the experiment showed that this effect appeared to be unique to depression.

The current findings that laypeople tend to underestimate depression of highly competent people have several real-life implications. First, this lay theory can cause someone to miss symptoms of depression in competent people, and thereby preclude them from being able to offer help or support to these depressed people. Second, people holding this lay theory may expect competent people to be less depressed (e.g., Fiske, 2002), and these expectations could contribute to an overall societal pressure for competent people to hide their depressive symptoms. This, in turn, can make the detection of depression among competent people even more difficult.

How can these potentially detrimental effects be prevented? Our results from the clinician participants suggest a possible remedy. Experiments 1-b and 2-b showed that compared to laypeople, clinicians were much less likely to discount depression among competent people. These findings are particularly surprising given past studies (e.g., Christiansen & Jacobsen, 1994; Ebling & Levenson, 2003) showing that clinicians are no more accurate than laypeople in their clinical judgments. What appears to be unique about the current study compared to those past studies is that the phenomenon uncovered may be directly related to the mere encounters with the clients. Clinicians might not necessarily believe that competent people are less depressed in the first place, because competent people are more open about their depressive symptoms with their clinicians than with their acquaintances. Furthermore, as detailed in the Supplemental Materials, nearly half of the clients that our clinician participants see have depression, and thus, compared to lay people, clinicians must be much more likely to be exposed to those who are competent but depressed. This increased exposure might have prevented them from overgeneralizing the belief that competent people are less likely to suffer from depression (see Blair, Ma, & Lenton, 2001; Dasgupta & Asgari, 2004; Dasgupta & Greenwald, 2001 for similar effects of exposure on counteracting stereotypes). Although it would not be feasible to raise laypeople's exposure to competent, depressed people, increasing awareness that seemingly "perfect" people can also suffer from severe depression might make it less likely for laypeople to ignore signs of depression among competent others.

The current study has several limitations that future studies can address. First, we failed to find that clinicians assume competent people are less likely to be depressed. However, this null effect might be the results of clinicians' memory for vignettes being too good, possibly because clinicians are generally better at remembering mental disorder symptoms, and/or because clinicians put more cognitive effort into the task. Thus, we cannot yet conclude that clinicians do not have any biases about competent people's depression, and future studies with tasks that are more challenging to clinicians may be able to address this question.

Another potential limitation involves the use of the foil vignette in inducing confusion regarding the depressive symptoms. As with most false memory or false recognition studies, this methodology was employed to illuminate a distinct *disposition* that laypersons have in their memory or recognition processes. This recognition test paradigm shows how memories of a depressed person can be contaminated by subsequent information in a variety of ways, depending on the depressed person's perceived competence. When laypeople encounter depressed persons in real-world interactions, however, it is unlikely that they will immediately encounter another person with similar depressive symptoms that cause confusion. More generally speaking, it is yet unclear how robust this effect of lay theories about competent people's depression is in real-life settings.

Additionally, in the current study competence-based details were always presented first, followed by the information about depression. We used this order to simulate the way most people are exposed to information when meeting a new person (i.e., typically people do not first find out about a person's depression before learning about other basic characteristics). Nonetheless, whether learning about a person's competence after learning the depressive symptoms can retrospectively affect people's memory remains an interesting research question for future studies.

Finally, although we have assumed that the memory biases found in the current study are due to lay theories on competent people's depression, the current study did not provide any direct evidence for the operation of such lay theories. At this point, it is difficult to conjecture how else the memory biases might have occurred if not for the lay theories. Yet in case there are other factors resulting in the memory biases, future research can obtain measures of the perceived severity of depression in competent versus less competent people, and examine whether the extent to which one expects differences in severity predicts the amount of memory biases.

In conclusion, the present study provided evidence of laypeople's bias in perceiving depressive emotions among competent people by presenting differences in memory for competent versus less competent people's identical depressive symptoms. Given these results, it is possible that at least some of the aforementioned shock in response to a model student's suicide stems from an inaccurate interpretation and recollection of what this student was truly experiencing. Being mindful of these potential distortions, and recommending such vulnerable people to professionals who may be less prone to these biases, appears to be key to prevention and treatment.

References

Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews*, *1*(1), 21-62.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological science*, *27*(8), 1069-1077.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of personality and social psychology*, *81*(5), 828.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: the impact of task complexity. *Journal of personality and social psychology*, *52*(5), 871.

Borchard, T. (2009). Depression Happens to Successful People. *Psych Central*. Retrieved on June 11, 2017, from https://psychcentral.com/blog/archives/2009/07/24/depression-happens-to-successful-people/

Burr, J. (2002). Cultural stereotypes of women from South Asian communities: mental health care professionals' explanations for patterns of suicide and depression. *Social Science & Medicine*, *55*(5), 835-845.

Christensen, A., & Jacobson, N. S. (1994). Who (or what) can do psychotherapy: The status and challenge of nonprofessional therapies. *Psychological science*, *5*(1), 8-14.

Cohen, S., & Italiano, L. (2017, February 2). Suicide Wave Grips Columbia. *New York Post*.

Retrieved from https://nypost.com/2017/02/02/suicide-wave-grips-columbia/

Cole, D. A., Martin, J. M., & Powers, B. (1997). A competency-based model of child depression: A longitudinal study of peer, parent, teacher, and self-evaluations. *Journal of Child Psychology and Psychiatry*, *38*(5), 505-514.

Cooper-Patrick, L., Powe, N. R., Jenckes, M. W., Gonzales, J. J., Levine, D. M., & Ford, D. E. (1997). Identification of patient attitudes and preferences regarding treatment of depression. *Journal of general internal medicine*, *12*(7), 431-438.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92(4), 631.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642-658.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of personality and social psychology*, *81*(5), 800.

Ebling, R., & Levenson, R. W. (2003). Who are the marital experts?. *Journal of Marriage and Family*, *65*(1), 130-142.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44(3), 227-240.

Fagan, K. (2015, May 7). Split Image. *ESPN*. Retrieved January 5, 2018, from http://www.espn.com/espn/feature/story/_/id/12833146/instagram-account-university-pennsylvania-runner-showed-only-part-story

Fiske, S. T. (2002). What we know now about bias and intergroup conflict, the problem of the century. *Current Directions in Psychological Science*, *11*(4), 123-128.

Fiske, S. T. (2012). Warmth and competence: Stereotype content issues for clinicians and researchers. *Canadian Psychology/Psychologie Canadienne*, *53*(1), 14.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, *82*(6), 878.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis) respecting versus (dis) liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, *55*(3), 473-489.

Gulliver, A., Griffiths, K. M., & Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC psychiatry*, *10*(1), 113.

Hollon, S. D., Thase, M. E., & Markowitz, J. C. (2002). Treatment and prevention of depression. *Psychological Science in the public interest*, *3*(2), 39-77.

Jordan, A. H., Monin, B., Dweck, C. S., Lovett, B. J., John, O. P., & Gross, J. J. (2011). Misery has more company than people think: Underestimating the prevalence of others' negative emotions. *Personality and Social Psychology Bulletin*, *37*(1), 120-135.

Kiamanesh, P., Dyregrov, K., Haavind, H., & Dieserud, G. (2014). Suicide and perfectionism: A psychological autopsy study of non-clinical suicides. *OMEGA-Journal of Death and*

Dying, *69*(4), 381-399.

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2009). Facial emotion perception in schizophrenia: a meta-analytic review. *Schizophrenia bulletin*, *36*(5), 1009-1019.

Kuppin, S., & Carpiano, R. M. (2008). Public conceptions of serious mental illness and substance abuse, their causes and treatments: Findings from the 1996 General Social Survey. *American Journal of Public Health*, *98*(Supplement_1), S120-S125.

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American journal of epidemiology*, *157*(2), 98-112.

Merkin, D. (2018, June 7). Kate Spade and the Illness Hidden With a Smile. *New York Times*. Retrieved from https://www.nytimes.com/2018/06/07/opinion/kate-spade-depression.html

National Institute of Mental Health (2016). *Major Depression*. Retrieved December 12, 2017, from https://www.nimh.nih.gov/health/statistics/major-depression.shtml#part_155033.

Potts, M. K., Burnam, M. A., & Wells, K. B. (1991). Gender differences in depression detection: A comparison of clinician diagnosis and standardized assessment. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(4), 609.

Read, J., Mosher, L. R., & Bentall, R. P. (2004). *Models of madness: Psychological, social and biological approaches to schizophrenia*. Psychology Press.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Sheldon, K. M., Ryan, R., & Reis, H. T. (1996). What makes for a good day? Competence and autonomy in the day and in the person. *Personality and social psychology bulletin*, *22*(12), 1270-1279.

Schacter, D. L. (1995). Memory distortion: History and current status. *Memory distortion: How minds, brains, and societies reconstruct the past*, 1-43.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American psychologist*, *54*(3), 182.

Scelfo, J. (2015, July 27). Suicide on Campus and the Pressure of Perfection. *The New York Times*. Retrieved from https://www.nytimes.com/2015/08/02/education/edlife/stress-social-media-and-suicide-on-campus.html

Stoppe, G., Sandholzer, H., Huppertz, C., Duwe, H., & Staedt, J. (1999). Gender differences in the recognition of depression in old age. *Maturitas*, *32*(3), 205-212.

Young, A. S., Klap, R., Sherbourne, C. D., & Wells, K. B. (2001). The quality of care for depressive and anxiety disorders in the United States. *Archives of general psychiatry*, *58*(1), 55-61.

Supplemental Materials

S1. Participant Information

Table S1 shows the basic demographic information of all participants. Table S2 shows additional information collected from clinician participants.

In all of the experiments, we used several exclusion criteria to eliminate data that clearly suggest failure to follow the instructions or a failed manipulation check. First, participants were excluded if they failed the manipulation check, by providing too low competence rating for the competent condition, too high competence rating for the incompetent condition, or too high or too low rating for the average condition. More specifically, we excluded participants if their averaged competence rating was below 2 SDs[1] of the mean competence rating for the competent condition, above 2 SDs of the mean competence rating for the incompetent condition, or above or below for the average condition. (See the third column of Table S1 for the number and the percentages of those excluded as a result). Second, participants were excluded if they incorrectly rated five or more of the studied or noncritical lures out of the total of 10 (see the fourth column of Table S1), suggesting that they did not carefully read the target vignette. Third, in order to ensure similar levels of cognitive disengagement from the vignettes that they read before taking the memory test, participants were excluded if they failed to answer half or more of the intermediate task's questions (See the fifth column of Table S1).

In Experiments 1b and 2b, clinician participants were excluded if they indicated through their answers to the demographic questions that they might not be currently licensed to practice (e.g., writing "n/a" or a similar answer in response to the licensure year question, or selecting only the "other" option for the question asking about their highest degree earned). Though Psychlist Marketing assures quality checks for their address lists, these exclusions were enacted to eliminate participants that exhibited any possibility of not being licensed clinicians.

[1] The scale used for the competence ratings was 5-point, ranging from 1 (Not at all) to 5 (Extremely). That is, if the rating was below 3, it means that participants did not believe the person was competent. Using 2 SDs as our cutoff was a conservative decision for the following reasons. (1) We predicted that there would be differences between the competent and the average conditions of Experiments 1-a and 2-a. The higher cutoffs for the competence exclusion criterion in the average condition using +2 SD were in the high 4's (e.g., 4.99 out of 5). Despite including the participants who already believed that the "average" person in the vignette was highly competent, we still found the significant differences between the competent and the average conditions. (2) We predicted no differences between the average and the incompetent conditions of Experiments 1-a and 2-a. The lower cutoffs for the competence exclusion criterion in the average condition were not at the low 1's, and instead ranged between 1.90 and 2.1. Thus, we excluded participants who believed the average person in the vignette was highly incompetent, making the average and the incompetent conditions more distinguishable. Despite that, we found no differences between these two conditions in Experiments 1-a and 2-a.

Table S1. *Participant sample size, mean age, and demographics per experiment. Percentages were calculated after exclusions.*

Experiment[2]	N before exclusions	Reasons for Exclusion[3]			N Unlicensed	N after exclusions	Mean Age (SD)	% of Women	% of Race[4]			
		Manipulation Check	Recognition test	Intermediate task					White	Asian	Black	Other
1a (Female)	153	10	8	3	n/a	132	35.49 (11.74)	42%	56%	34%	3%	5%
1a (Male)	151	7	17	1	n/a	126	35.73 (12.01)	41%	67%	21%	5%	6%
1b (Female)	114	5	1	2	5	101	n/a	84%	84%	2%	8%	7%
1b (Male)	129	4	6	2	3	114	n/a	82%	84%	0%	9%	8%
2a (Female)	152	8	8	2	n/a	134	37.38 (12.74)	46%	72%	10%	9%	9%
2a (Male)	150	9	14	3	n/a	124	36.15 (12.66)	39%	56%	27%	6%	10%
2b (Female)	128	5	1	1	1	120	n/a	86%	80%	8%	6%	8%
2b (Male)	120	5	1	2	4	108	n/a	81%	84%	5%	6%	6%
3a (Female)	450	19	19	11	n/a	401	35.19 (11.23)	47%	67%	19%	6%	8%
3b (Male)	476	22	22	10	n/a	422	34.81 (11.63)	39%	45%	41%	4%	9%

[2] The “male” and “female” stand for male and female versions of vignettes.

[3] See text for the details.

[4] All demographic questions were optional; some rows do not total 100% because some participants did not select a response, or selected more than one response.

Table S2. *Additional clinician demographic information.*

Demographic	Experiment 1b		Experiment 2b	
	Competent Condition	Average Condition	Competent Condition	Average Condition
Years since licensure (range; SD)	13.98 (0-42, 10.96)	10.77 (0-43, 9.17)	11.35 (0-37, 8.41)	11.50 (0-40, 9.45)
% of Clients with Depression	45.63	52.11	42.20	40.39
Highest degree earned, %[5]				
Ph.D	9	10	9	15
Psy.D	3	8	8	10
Ed.D	1	0	1	2
MD	0	0	1	1
LCSW/MSW	55	51	46	46
Other Masters Degrees	31	30	33	24
Other	2	2	2	2

[5] Participants were allowed to select more than one response.

S2. Analyses Involving Gender of the Character in Vignette

As described in the main text, two versions of stimuli were used, one involving a male character and the other involving a female character. None of the critical effects reported in the main text interacted with the gender of the character in the vignettes as described below.

Experiment 1a. A 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (conditions; competent, average, incompetent) x 2 (gender of vignette; male, female) mixed ANOVA on error ratings with the item type as a within-subject variable and ratings on warmth as a covariate found no significant 3-way interaction effect, $F(3.06, 384.85) = 1.98$, $p = .12$; $\eta_p^2=.016$.

Experiment 1b. A 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (conditions; competent, average) x 2 (gender of vignette; male, female) x 2 (participant type; laypeople, clinicians) mixed ANOVA on error ratings with the item type as a within-subject variable and ratings on warmth as a covariate showed no significant 4-way interaction effect, $F(1.54, 570.50) = .71$, $p = .46$; $\eta_p^2= .001$.

Experiment 2a. A 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (conditions; competent, average, incompetent) x 2 (gender of vignette; male, female) mixed ANOVA on error ratings with the item type as a within-subject variable and ratings on warmth as a covariate found no significant 3-way interaction effect, $F(2.67, 335.42) = .03$, $p = .99$; $\eta_p^2 < .001$.

Experiment 2b. A 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (conditions; competent, average) x 2 (gender of vignette; male, female) x 2 (participant type; laypeople, clinicians) mixed ANOVA on error ratings with the item type as a within-subject variable and ratings on warmth as a covariate found no significant 4-way interaction effect, $F(1.44, 562.08) = .53$, $p = .53$; $\eta_p^2= .001$.

Experiment 3. For each of the disorder conditions, we analyzed the error ratings using a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (conditions; competent, average, incompetent) x 2 (gender of vignette; male, female) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate. The 3-way interaction effects were not significant in any of the disorder conditions; $F(2.96, 384.70) = .38$, $p = .77$; $\eta_p^2= .003$ for anxiety, $F(3.10, 412.53) = .75$, $p = .53$; $\eta_p^2= .006$ for schizophrenia, and $F(2.82, 388.43) = .61$, $p = .60$; $\eta_p^2= .004$ for physical illness.

S3. Manipulation Check Results

In all of the experiments throughout this paper, we measured participants' perception of the competence of the character in the target vignette to confirm that competence was successfully varied between conditions. Participants went through a list of adjectives used to describe competence according to the stereotype content model (e.g., efficient, organized, competent, capable, intelligent, skillful; Fiske, Cuddy, & Xu, 2002), and rated how much they thought the character fit each description, from 1 (not at all) to 5 (extremely). These ratings for each adjective were averaged over the total number of adjectives to obtain one overall perceived competence rating for each participant. In each experiment, we split participants across the gender of the vignette's character, as perceptions of competence and warmth might vary depending on their held stereotypes about gender. We then ran one-way ANOVAs testing the effects of condition for separately for male and female version of the vignette on average competence ratings of each participant. As reported in Table S3, main effects of condition were significant in all versions in all experiments. Post-hoc Tukey HSD tests revealed that across all experiments, the characters in the competent condition were perceived as significantly more competent than characters in the average and incompetent conditions, all *p'*s < .001. Characters in the average condition were also perceived as significantly more competent than characters in the incompetent condition, all *p*'s < .001. See Table S3 for the means and standard deviations broken by condition, vignette version, and experiment.

Table S3. *Results of one-way ANOVAs testing effects of condition on competence ratings, and means and standard deviations of competence ratings broken down by condition, vignette version, and experiment.*

Experiment	Competent		Average		Incompetent		F value	*p*	*df*
	M	SD	M	SD	M	SD			Error
1a (Female)	4.45	.46	3.63	.56	1.84	.62	262.62	<.001	129
1a (Male)	4.46	.53	3.22	.52	2.07	.62	189.16	<.001	123
1b (Female)	4.50	.43	3.48	.45	--	--	136.12	<.001	99
1b (Male)	4.62	.41	3.42	.44	--	--	232.69	<.001	112
2a (Female)	4.51	.41	3.64	.61	1.66	.44	395.92	<.001	131
2a (Male)	4.48	.46	3.32	.56	1.84	.59	248.39	<.001	121
2b (Female)	4.37	.47	3.58	.41	--	--	97.82	<.001	118
2b (Male)	4.39	.44	3.42	.40	--	--	140.87	<.001	106
3 (anxiety, Female)	4.50	.55	3.67	.52	1.99	.75	183.57	<.001	126
3 (anxiety, Male)	4.31	.66	3.67	.62	2.03	.69	148.29	<.001	135
3 (schizophrenia, Female)	4.26	.63	3.53	.49	1.67	.55	267.37	<.001	133
3 (schizophrenia, Male)	4.29	.61	3.33	.54	2.58	.57	103.04	<.001	134
3 (physical illness, Female)	4.40	.60	3.79	.60	1.83	.63	211.94	<.001	133
3 (physical illness, Male)	4.71	.35	3.62	.65	1.83	.61	321.25	<.001	144

We also measured participants' perception of warmth of the target character because it is the other important component of the stereotype content model, and its status as a possible confound was examined. This was done using adjectives used to describe warmth according to the stereotype content model (e.g., friendly, well-intentioned, trustworthy, warm, good-natured, sincere; Fiske, et al., 2002) from 1 (not at all) to 5 (extremely). We again split participants across the gender of the vignette's character, and ran one-way ANOVAs to test for condition's effect on warmth ratings. Warmth ratings were significantly different across conditions for several experiments (see Table S4 for the descriptive and inferential statistics). As summarized in Table S5, however, post-hoc tests show that most of the competent and average vignettes do not significantly differ in their warmth ratings -- thus making it unlikely that these differences in warmth might be driving the memory distortion effects observed throughout this paper. The decreased warmth is apparent in the incompetent condition; however, given the lack of significant differences in error ratings between the incompetent and average conditions throughout the experiments, it appears improbable that this low warmth is causing any significant effects in memory errors. To confirm this, analyses in the main text for all experiments were run controlling for warmth as a covariate.

Table S4. *Results of one-way ANOVAs testing effects of condition on warmth ratings, and means and standard deviations of warmth ratings broken down by condition, vignette version, and experiment.*

Experiment	Competent		Average		Incompetent		F value	*p*	*df*
	M	SD	M	SD	M	SD			
1a (Female)	4.37	.48	4.20	.51	3.13	.78	54.83	<.001	129
1a (Male)	4.10	.63	3.85	.57	3.60	.69	6.69	.002	123
1b (Female)	4.13	.63	3.90	.59	--	--	3.52	.06	99
1b (Male)	3.87	.60	3.70	.42	--	--	2.95	.09	112
2a (Female)	4.38	.49	4.22	.61	2.99	1.04	46.18	<.001	131
2a (Male)	4.03	.49	3.78	.59	3.49	.61	9.27	<.001	121
2b (Female)	4.09	.57	4.01	.46	--	--	.82	.36	118
2b (Male)	3.76	.68	3.73	.46	--	--	.08	.78	106
3 (anxiety, Female)	4.43	.56	4.25	.42	3.15	.86	50.42	<.001	126
3 (anxiety, Male)	4.07	.69	3.99	.72	3.17	.70	23.54	<.001	135
3 (schizophrenia, Female)	4.25	.58	3.99	.64	3.12	.72	40.97	<.001	133
3 (schizophrenia, Male)	4.13	.57	3.73	.57	3.20	.59	30.36	<.001	134
3 (physical illness, Female)	4.38	.59	4.49	.53	3.32	.75	48.28	<.001	133
3 (physical illness, Male)	4.37	.51	4.15	.47	3.38	.88	31.02	<.001	144

Table S5. *P-values from Tukey's HSD comparing means for warmth ratings. Experiments 1b and 2b were excluded here, as they only had two conditions.*

Experiment	Compared Conditions		
	Competent-Average	Average-Incompetent	Competent-Incompetent
1a (Female)	.40	<.001	<.001
1a (Male)	.17	.17	.001
2a (Female)	.58	<.001	<.001
2a (Male)	.13	.06	<.001
3 (anxiety, Female)	.41	<.001	<.001
3 (anxiety, Male)	.86	<.001	<.001
3 (schizophrenia, Female)	.16	<.001	<.001
3 (schizophrenia, Male)	.004	<.001	<.001
3 (physical illness, Female)	.66	<.001	<.001
3 (physical illness, Male)	.19	.17	.001

S7. Analyses of Results of Experiment 3 broken down by Individual Symptoms

As explained in the main text, the lack of significant interaction effects between condition and item type in Experiment 3 could be because the analyses were done averaged over three symptoms in each disorder, which might not have been a coherent set in laypeople's conceptualization. For each symptom in each disorder, a one-way ANOVA was conducted testing the effect of condition to test whether the null results are not due to idiosyncratic symptoms selected for the study. As summarized in Table S6, there was no significant effect of condition on any of the symptoms.

Table S6. *Results of one-way ANOVAs testing the effect of condition on error ratings for each individual symptom of each disorder.*

	Estimated Marginal Means (SE) of Each Condition			Results of One-way ANOVA		
	Competent	Average	Incompetent	F-value	*p*	df
			Anxiety			
S1. Restlessness	2.84 (.24)	3.53 (.23)	3.16 (.26)	2.44	.09	263
S2. Excessive worry	3.93 (.23)	3.93 (.22)	3.59 (.25)	.57	.57	263
S3. Irritability	4.61 (.20)	4.50 (.20)	4.74 (.22)	.30	.74	263
			Schizophrenia			
S1. Hallucinations	3.14 (.21)	3.03 (.23)	3.08 (.23)	.06	.95	269
S2. Disorganized speech	4.13 (.21)	3.51 (.22)	3.94 (.22)	2.24	.11	269
S3. Delusions	2.15 (.20)	2.37 (.21)	2.41 (.21)	.45	.64	269
			Physical Illness			
S1. Joint pain	2.41 (.22)	2.59 (.20)	3.06 (.24)	1.76	.17	279
S2. Digestive problems	2.91 (.23)	2.92 (.21)	3.37 (.26)	.95	.39	279
S3. Fever	2.00 (.19)	2.28 (.18)	2.08 (.21)	.68	.51	279

S8. Estimated Marginal Means, Standard Error, and Confidence Intervals Across Experiments

As mentioned in the main text, warmth was significantly different across conditions in some experiments, and was therefore included as a covariate in all analyses. The estimated marginal means (summarized in Table S7, along with the 95% confidence intervals), adjusted for warmth, show that the pattern of results remains the same even with this covariate included.

Table S7. *Estimated marginal means and 95% confidence intervals of each experiment, broken down by condition and item type.*

Item Type	Conditions			Overall
	Competent	Average	Incompetent	
	M, *SE* [95% CI]	*M*, *SE* [95% CI]	*M*, *SE* [95% CI]	
		Experiment 1a		
Studied	1.50, .06 [1.38, 1.61]	1.53, .06 [1.42, 1.64]	1.45, .06 [1.34, 1.57]	1.49, .03
Noncritical Lures	1.46, .08 [1.31, 1.61]	1.28, .07 [1.14, 1.42]	1.35, .07 [1.21, 1.50]	1.36, .04
Critical Lures	2.28, .14 [2.01, 2.55]	1.53, .13 [1.27, 1.79]	1.55, .13 [1.28, 1.81]	1.79, .07
		Experiment 1b		
Studied	1.66, .06 [1.55, 1.77]	1.59, .05 [1.48, 1.69]	--	1.62, .04
Noncritical Lures	1.31, .05 [1.22, 1.40]	1.19, .04 [1.11, 1.28]	--	1.25, .03
Critical Lures	1.48, .10 [1.28, 1.68]	1.50, .10 [1.31, 1.70]	--	1.49, .07
		Experiment 2a		
Studied	1.70, .06 [1.58, 1.83]	1.60, .06 [1.48, 1.72]	1.42, .07 [1.29, 1.55]	1.58, .04
Noncritical Lures	1.25, .06 [1.13, 1.36]	1.33, .06 [1.21, 1.44]	1.20, .06 [1.08, 1.32]	1.26, .03

Critical Lures	2.07, .16 [1.76, 2.38]	3.14, .15 [2.84, 3.44]	3.12, .16 [2.79, 3.44]	2.78, .09
		Experiment 2b		
Studied	1.78, .06 [1.67, 1.89]	1.69, .06 [1.58, 1.80]	--	1.73, .04
Noncritical Lures	1.16, .03 [1.10, 1.23]	1.10, .03 [1.03, 1.16]	--	1.13, .02
Critical Lures	1.89, .10 [1.69, 2.09]	1.60, .10 [1.40, 1.80]	--	1.75, .07
		Experiment 3: Anxiety		
Studied	1.76, .07 [1.62, 1.89]	1.65, .07 [1.52, 1.79]	1.35, .08 [1.20, 1.50]	1.59, .04
Noncritical Lures	1.64, .10 [1.45, 1.83]	1.57, .09 [1.39, 1.75]	1.34, .10 [1.14, 1.54]	1.52, .05
Critical Lures	3.80, .16 [3.48, 4.11]	3.98, .15 [3.68, 4.29]	3.83, .17 [3.49, 4.17]	3.87, .09
		Experiment 3: Schizophrenia		
Studied	1.60, .07 [1.47, 1.73]	1.63, .07 [1.50, 1.77]	1.52, .07 [1.38, 1.66]	1.59, .04
Noncritical Lures	1.63, .08 [1.47, 1.79]	1.53, .09 [1.36, 1.70]	1.53, .09 [1.36, 1.71]	1.56, .05
Critical Lures	3.21, .16 [2.90, 3.52]	3.00, .16 [2.67, 3.32]	3.05, .17 [2.72, 3.38]	3.08, .09
		Experiment 3: Physical Illness		
Studied	1.62, .07 [1.49, 1.75]	1.62, .06 [1.50, 1.74]	1.41, .07 [1.27, 1.56]	1.55, .04
Noncritical Lures	1.72, .07 [1.57, 1.87]	1.54, .07 [1.40, 1.68]	1.38, .08 [1.21, 1.54]	1.55, .04
Critical Lures	2.44, .16 [2.13, 2.76]	2.60, .15 [2.31, 2.89]	2.84, .18 [2.48, 3.19]	2.63, .09

S9: Analyses Prior to Exclusions or Using One of Exclusion Criteria

As explained in S1, three exclusion criteria were used; manipulation check, performance in non-critical and studied items during the recognition test, and attention check during the intermediate task. Although all three exclusion criteria are crucial in validating the experimental manipulations and participants' attention, this section reports analyses using none or only one of the three exclusion criteria for each experiment for full transparency. Tables S8-S14 show the estimated means, standard errors, and ANOVA results for each experiment when various exclusion criteria were used. The ANOVs that were used were a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 3 (conditions; competent, average, incompetent) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate for Experiments 1a, 2a, and 3, and a 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (condition; competent, average) mixed ANOVA with the item type as a within-subject variable and ratings on warmth as a covariate for Experiments 1b and 2b.

Most results prior to exclusions were nonsignificant, as would be expected due to the data including participants who failed manipulation and attention checks, but the numbers generally remained consistent with the directions found in the results.

Table S8. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 1a*

Exclusion Used	Item Type	Conditions			Overall	ANOVA results
		Competent	Average	Incompetent		
None	Studied	1.77 (.08)	1.65 (.08)	1.63 (.08)	1.68	Item Type: $F(1.57, 469.93) = .22$, $p = .75$, $\eta_p^2=.001$ Condition: $F(2, 300) = 3.24$, $p = .04$, $\eta_p^2= .02$ Interaction: $F(3.13, 469.93) = .93$, $p = .43$, $\eta_p^2=.006$
	Noncritical Lures	1.79 (.10)	1.54 (.10)	1.50 (.10)	1.61	
	Critical Lures	2.20 (.14)	1.93 (.14)	1.72 (.14)	1.95	
Manipulation Check	Studied	1.79 (.08)	1.68 (.08)	1.48 (.08)	1.65	Item Type: $F(1.56, 441.77) = .32$, $p = .67$, $\eta_p^2= .001$ Condition: $F(2, 283) = 8.61$, $p < .001$, $\eta_p^2= .057$ Interaction: $F(3.12, 441.77) = 1.37$, $p = .25$, $\eta_p^2= .009$
	Noncritical Lures	1.83 (.10)	1.54 (.10)	1.35 (.10)	1.57	
	Critical Lures	2.27 (.14)	1.81 (.14)	1.57 (.14)	1.88	
Recognition Test	Studied	1.50 (.06)	1.53 (.06)	1.49 (.06)	1.51	Item Type: $F(1.47, 397.98) = 1.37$, $p = .25$, $\eta_p^2= .005$ Condition: $F(2, 270) = 3.72$, $p = .03$, $\eta_p^2= .03$ Interaction: $F(2.95, 397.98) = 4.98$, $p = .002$, $\eta_p^2= .036$
	Noncritical Lures	1.46 (.08)	1.30 (.07)	1.40 (.07)	1.39	
	Critical Lures	2.28 (.14)	1.71 (.14)	1.64 (.14)	1.87	
Intermediate Task	Studied	1.75 (.08)	1.60 (.08)	1.64 (.08)	1.66	Item Type: $F(1.60, 470.80) = .20$, $p = .77$, $\eta_p^2 < .001$ Condition: $F(2, 294) = 3.51$, $p = .03$, $\eta_p^2 = .02$ Interaction: $F(3.20, 470.80) = 1.09$, $p =.36$, $\eta_p^2 = .007$
	Noncritical Lures	1.77 (.10)	1.46 (.10)	1.50 (.10)	1.57	
	Critical Lures	2.16 (.14)	1.86 (.14)	1.69 (.14)	1.90	

Table S9. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 1b*

Exclusion Used	Item Type	Conditions		Overall	ANOVA results
		Competent	Average		
None*	Studied	1.72 (.07)	1.67 (.07)	1.70	Item Type: $F(1.68, 389.06) = 1.16$, $p = .31$, $\eta_p^2 = .004$ Condition: $F(1, 232) = .01$, $p = .92$, $\eta_p^2 < .001$ Interaction: $F(1.68, 396.06) = 2.08$, $p = .14$, $\eta_p^2 = .009$
	Noncritical Lures	1.34 (.05)	1.21 (.05)	1.28	
	Critical Lures	1.46 (.11)	1.61 (.11)	1.54	
Manipulation Check	Studied	1.73 (.06)	1.63 (.06)	1.68	Item Type: $F(1.63, 363.34) = 2.78$, $p = .07$, $\eta_p^2 = .01$ Condition: $F(1, 223) = .51$, $p = .48$, $\eta_p^2 = .002$ Interaction: $F(1.63, 363.34) = 1.64$, $p = .20$, $\eta_p^2 = .007$
	Noncritical Lures	1.35 (.05)	1.20 (.05)	1.28	
	Critical Lures	1.46 (.11)	1.56 (.11)	1.51	
Recognition Test	Studied	1.66 (.05)	1.58 (.05)	1.62	Item Type: $F(1.62, 361.92) = 3.04$, $p = .06$, $\eta_p^2 = .01$ Condition: $F(1, 224) = .39$, $p = .53$, $\eta_p^2 = .002$ Interaction: $F(1.62, 361.92) = 1.27$, $p = .28$, $\eta_p^2 = .006$
	Noncritical Lures	1.30 (.04)	1.19 (.04)	1.24	
	Critical Lures	1.43 (.10)	1.51 (.10)	1.47	
Intermediate Task	Studied	1.72 (.07)	1.67 (.07)	1.69	Item Type: $F(1.67, 381.77) = .83$, $p = .42$, $\eta_p^2 = .004$ Condition: $F(1, 228) = .03$, $p = .87$, $\eta_p^2 < .001$ Interaction: $F(1.67, 381.77) = 1.78$, $p = .18$, $\eta_p^2 = .008$
	Noncritical Lures	1.34 (.05)	1.21 (.05)	1.28	
	Critical Lures	1.46 (.11)	1.60 (.11)	1.53	

Note: *Participants who are not licensed are excluded from all of these analyses.

Table S10. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 2a*

Exclusion Used	Item Type	Conditions			Overall	ANOVA results
		Competent	Average	Incompetent		
None	Studied	1.88 (.08)	1.76 (.08)	1.62 (.09)	1.75	Item Type: *F*(1.37, 408.62) = 6.81, *p* = .004, η_p^2=.02 Condition: *F*(2, 298) = 1.62, *p* = .20, η_p^2= .01 Interaction: *F*(2.74, 408.62) = 4.30, *p* = .007, η_p^2=.03
	Noncritical Lures	1.47 (.08)	1.47 (.08)	1.36 (.09)	1.43	
	Critical Lures	2.53 (.15)	3.12 (.15)	2.82 (.16)	2.82	
Manipulation Check	Studied	1.88 (.08)	1.79 (.08)	1.51 (.09)	1.73	Item Type: *F*(1.35, 379.22) = 7.53, *p* = .003, η_p^2= .026 Condition: *F*(2, 281) = 5.42, *p* = .005, η_p^2= .037 Interaction: *F*(2.70, 379.22) = 5.96, *p* = .001, η_p^2= .04
	Noncritical Lures	1.49 (.08)	1.50 (.07)	1.17 (.08)	1.39	
	Critical Lures	2.49 (.15)	3.20 (.15)	2.75 (.16)	2.81	
Recognition Test	Studied	1.69 (.07)	1.56 (.06)	1.51 (.07)	1.59	Item Type: *F*(1.38, 374.00) = 13.55, *p* < .001, η_p^2= .05 Condition: *F*(2, 272) = 2.40, *p* = .09, η_p^2= .02 Interaction: *F*(2.75, 374.00) = 6.99, *p* < .001, η_p^2= .05
	Noncritical Lures	1.28 (.07)	1.31 (.07)	1.33 (.07)	1.31	
	Critical Lures	2.40 (.16)	3.17 (.15)	2.87 (.17)	2.81	
Intermediate Task	Studied	1.86 (.08)	1.76 (.08)	1.61 (.09)	1.74	Item Type: *F*(1.36, 398.24) = 6.21, *p* = .007, η_p^2= .02 Condition: *F*(2, 293) = 2.16, *p* = .12, η_p^2= .01 Interaction: *F*(2.72, 398.24) = 4.46, *p* = .006, η_p^2= .03
	Noncritical Lures	1.43 (.08)	1.47 (.08)	1.35 (.09)	1.42	
	Critical Lures	2.49 (.16)	3.12 (.15)	2.80 (.16)	2.80	

Table S11. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 2b*

Exclusion Used	Item Type	Conditions		Overall	ANOVA results
		Competent	Average		
None	Studied	1.85 (.06)	1.71 (.06)	1.78	Item Type: *F*(1.59, 382.29) = 2.96, *p* = .07, η_p^2= .01 Condition: *F*(1, 240) = 7.44, *p* = .007, η_p^2 = .03 Interaction: *F*(1.59, 382.29) = 1.72, *p* = .19, η_p^2= .007
	Noncritical Lures	1.19 (.03)	1.11 (.03)	1.15	
	Critical Lures	1.96 (.10)	1.64 (.10)	1.80	
Manipulation Check	Studied	1.83 (.06)	1.69 (.06)	1.76	Item Type: *F*(1.61, 370.24) = 3.35, *p* = .05, η_p^2= .01 Condition: *F*(1, 230) = 8.14, *p* = .005, η_p^2= .03 Interaction: *F*(1.61, 370.24) = 2.01, *p* = .15, η_p^2= .009
	Noncritical Lures	1.17 (.03)	1.10 (.03)	1.14	
	Critical Lures	1.94 (.10)	1.60 (.10)	1.77	
Recognition Test	Studied	1.81 (.06)	1.71 (.06)	1.76	Item Type: *F*(1.53, 363.95) = 2.52, *p* = .10, η_p^2= .01 Condition: *F*(1, 238) = 5.84, *p* = .016, η_p^2= .02 Interaction: *F*(1.53, 363.95) = 1.57, *p* = .21, η_p^2= .007
	Noncritical Lures	1.18 (.03)	1.11 (.03)	1.15	
	Critical Lures	1.93 (.10)	1.64 (.10)	1.79	
Intermediate Task	Studied	1.85 (.06)	1.71 (.06)	1.78	Item Type: *F*(1.60, 378.97) = 3.05, *p* = .06, η_p^2= .01 Condition: *F*(1, 237) = 7.33, *p* = .007, η_p^2= .03 Interaction: *F*(1.60, 378.97) = 1.65, *p* = .20, η_p^2= .007
	Noncritical Lures	1.19 (.03)	1.11 (.03)	1.15	
	Critical Lures	1.96 (.11)	1.64 (.10)	1.80	

Table S12. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 3, Anxiety*

Exclusion Used	Item Type	Conditions			Overall	ANOVA results
		Competent	Average	Incompetent		
None	Studied	1.99 (.09)	1.80 (.09)	1.55 (.09)	1.78	Item Type: *F*(1.48, 447.90) = 16.95, *p*< .001, η_p^2=.05 Condition: *F*(2, 303) = 1.48, *p* = .23, η_p^2= .01 Interaction: *F*(2.96, 447.90) = .85, *p* = .46, η_p^2= .006
	Noncritical Lures	1.75 (.10)	1.67 (.10)	1.51 (.11)	1.64	
	Critical Lures	3.84 (.14)	3.94 (.14)	3.89 (.15)	3.89	
Manipulation Check	Studied	1.89 (.09)	1.83 (.09)	1.43 (.09)	1.72	Item Type: *F*(1.45, 421.52) = 15.21, *p* < .001, η_p^2= .05 Condition: *F*(2, 291) = 3.44, *p* = .03, η_p^2= .02 Interaction: *F*(2.90, 421.52) = 1.13, *p* = .34, η_p^2= .008
	Noncritical Lures	1.73 (.10)	1.70 (.10)	1.36 (.11)	1.60	
	Critical Lures	3.84 (.15)	3.93 (.15)	3.87 (.16)	3.88	
Recognition Test	Studied	1.80 (.07)	1.66 (.07)	1.40 (.07)	1.62	Item Type: *F*(1.51, 434.41) = 20.79, *p* < .001, η_p^2= .07 Condition: *F*(2, 287) = 1.48, *p* = .23, η_p^2= .01 Interaction: *F*(3.03, 434.41) = 1.29, *p* = .28, η_p^2= .009
	Noncritical Lures	1.68 (.10)	1.54 (.10)	1.44 (.10)	1.55	
	Critical Lures	3.87 (.15)	4.01 (.15)	3.94 (.16)	3.94	
Intermediate Task	Studied	1.90 (.09)	1.80 (.09)	1.54 (.10)	1.75	Item Type: F(1.48, 428.37) = 13.92, *p* < .001, η_p^2= .05 Condition: *F*(2, 289) = 1.24, *p* = .29, η_p^2= .009 Interaction: *F*(2.96, 428.37) = 1.04, *p* = .38, η_p^2= .007
	Noncritical Lures	1.75 (.10)	1.67 (.10)	1.54 (.11)	1.65	
	Critical Lures	3.76 (.15)	3.93 (.14)	3.85 (.16)	3.84	

Table S13. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 3, Schizophrenia*

Exclusion Used	Item Type	Conditions			Overall	ANOVA results
		Competent	Average	Incompetent		
None	Studied	1.77 (.08)	1.75 (.08)	1.63 (.09)	1.72	Item Type: *F*(1.62, 489.28) = 10.69, *p*< .001, η_p^2=.03 Condition: *F*(2, 303) = 1.09, *p* = .34, η_p^2= .007 Interaction: *F*(3.23, 489.28) = .34, *p* = .81, η_p^2= .002
	Noncritical Lures	1.84 (.09)	1.72 (.10)	1.55 (.10)	1.70	
	Critical Lures	3.24 (.15)	3.10 (.15)	3.12 (.16)	3.15	
Manipulation Check	Studied	1.73 (.08)	1.77 (.08)	1.56 (.09)	1.69	Item Type: *F*(1.63, 468.64) = 13.00, *p* < .001, η_p^2= .04 Condition: *F*(2, 288) = 1.16, *p* = .32, η_p^2= .008 Interaction: *F*(3.25, 468.64) = .76, *p* = .53, η_p^2= .005
	Noncritical Lures	1.76 (.10)	1.73 (.10)	1.54 (.10)	1.68	
	Critical Lures	3.28 (.15)	3.03 (.16)	3.07 (.17)	3.12	
Recognition Test	Studied	1.62 (.06)	1.63 (.07)	1.53 (.07)	1.59	Item Type: *F*(1.60, 456.44) = 13.76, *p* < .001, η_p^2= .05 Condition: *F*(2, 286) = .66, *p* = .52, η_p^2= .005 Interaction: *F*(3.19, 456.44) = .29, *p* = .85, η_p^2= .002
	Noncritical Lures	1.70 (.09)	1.56 (.09)	1.55 (.10)	1.60	
	Critical Lures	3.24 (.15)	3.07 (.16)	3.11 (.17)	3.14	
Intermediate Task	Studied	1.76 (.08)	1.75 (.09)	1.63 (.09)	1.71	Item Type: *F*(1.59, 471.23) = 11.91, *p* < .001, η_p^2= .04 Condition: *F*(2, 297) = .73, *p* = .48, η_p^2= .005 Interaction: *F*(3.17, 471.23) = .22, *p* = .89, η_p^2= .001
	Noncritical Lures	1.80 (.09)	1.69 (.09)	1.56 (.10)	1.68	
	Critical Lures	3.19 (.15)	3.10 (.15)	3.11 (.16)	3.13	

Table S14. *Estimated means, standard errors (in parentheses), and ANOVA results from analyses using various exclusion criteria in Experiment 3, Physical Illness*

Exclusion Used	Item Type	Conditions			Overall	ANOVA results
		Competent	Average	Incompetent		
None	Studied	1.75 (.08)	1.75 (.07)	1.51 (.09)	1.67	Item Type: F(1.53, 471.87) = 1.26, p = .28, η_p^2=.004 Condition: F(2, 308) = .08, p = .92, η_p^2= .001 Interaction: F(3.06, 471.87) = 3.40, p = .02, η_p^2= .02
	Noncritical Lures	1.82 (.09)	1.69 (.09)	1.59 (.10)	1.70	
	Critical Lures	2.51 (.15)	2.71 (.14)	2.92 (.17)	2.71	
Manipulation Check	Studied	1.73 (.07)	1.75 (.07)	1.40 (.08)	1.63	Item Type: F(1.49, 436.51) = 1.04, p = 34, η_p^2= .003 Condition: F(2, 294) = .82, p = .44, η_p^2= .006 Interaction: F(2.97, 436.51) = 3.98, p = .008, η_p^2= .03
	Noncritical Lures	1.81 (.09)	1.73 (.09)	1.42 (.10)	1.65	
	Critical Lures	2.50 (.16)	2.71 (.14)	2.89 (.18)	2.70	
Recognition Test	Studied	1.61 (.06)	1.62 (.06)	1.43 (.07)	1.55	Item Type: F(1.47, 424.55) = 1.66, p = .20, η_p^2= .006 Condition: F(2, 289) = .02, p = .98, $\eta_p^2 < .001$ Interaction: F(2.94, 424.55) = 3.16, p = .03, η_p^2= .02
	Noncritical Lures	1.70 (.08)	1.50 (.07)	1.51 (.09)	1.57	
	Critical Lures	2.44 (.16)	2.64 (.15)	2.88 (.17)	2.65	
Intermediate Task	Studied	1.75 (.08)	1.76 (.07)	1.51 (.09)	1.67	Item Type: F(1.54, 473.17) = 1.25, p = .28, η_p^2= .004 Condition: F(2, 307) = .06, p = .94, η_p^2= .006 Interaction: F(3.08, 473.17) = 3.37, p = .02, η_p^2= .02
	Noncritical Lures	1.82 (.09)	1.69 (.09)	1.59 (.10)	1.70	
	Critical Lures	2.51 (.15)	2.69 (.14)	2.92 (.16)	2.71	

S10. Corresponding 3 x 2 ANOVAs for Lay Participants

Given that we conducted 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (condition; competent, average) ANOVAs for the clinicians in order to compare them to lay participants, we also conducted 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (condition; competent, average) mixed ANOVAs for the lay participants, excluding the incompetent condition to provide a proper comparison.

Experiment 1a. The 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (condition; competent, average) ANOVAs revealed no main effect of the item type, $F(1.40, 226.13) = .50$, $p = .54$, $\eta_p^2 = .003$, and a significant main effect of condition, $F(1, 162) = 11.87$, $p = .001$, $\eta_p^2 = .068$. As was predicted, a significant interaction effect qualified these results, $F(1.40, 226.13) = 11.78$, $p < .001$, $\eta_p^2 = .07$.

To understand the pattern of this interaction effect, one-way ANOVAs testing the effect of condition were performed for each item type. There was no significant effect of condition for the studied items, $F(1, 162) = .002$, $p = .96$, $\eta_p^2 < .001$. There was a significant effect of condition for the noncritical lures, $F(1, 162) = 4.28$, $p = .04$, $\eta_p^2 = .026$, because error ratings for the competent condition ($M = 1.40$, $SD = .65$) were higher than those for the average condition ($M = 1.26$, $SD = .49$). However, given this difference was relatively small, and did not occur with laypeople in Experiment 2, this effect was likely due to chance. Most importantly, there was a significant effect of condition for the critical lures, $F(1, 162) = 15.10$, $p < .001$, $\eta_p^2 = .09$, because the error ratings for the competent condition ($M = 2.23$, $SD = 1.48$) were significantly higher than those for the average condition ($M = 1.51$, $SD = .89$).

Experiment 2a. The 3 (item type; Studied, Noncritical Lure, Critical Lure) x 2 (condition; competent, average) ANOVAs revealed no significant main effect of the item type, $F(1.35, 227.46) = 2.94$, $p = .08$, $\eta_p^2 = .017$. There was a significant main effect of condition, $F(1, 168) = 9.69$, $p = .002$, $\eta_p^2 = .055$. As was predicted, there was a significant interaction effect, $F(1.35, 227.46) = 21.35$, $p < .001$, $\eta_p^2 = .11$.

To understand the pattern of this interaction effect, one-way ANOVAs testing the effect of condition were performed for each item type. There was no significant effect of condition for the studied items, $F(1, 168) = 1.79$, $p = .18$, $\eta_p^2 = .01$, or for noncritical lures, $F(1, 168) = .34$, $p = .56$, $\eta_p^2 = .002$. However, there was a significant effect of condition for the critical lures, $F(1, 168) = 20.59$, $p < .001$, $\eta_p^2 = .11$, because the error ratings for the competent condition ($M = 2.22$, $SD = 1.36$) were significantly lower than those for the average condition ($M = 3.22$, $SD = 1.39$).