Informant Discrepancies in Judgments about Change During Mental Health Treatments

Jessecae K. Marsh[1] Andrew S. Zeveney[1] Andres De Los Reyes[2]

[1] Lehigh University

[2] University of Maryland at College Park

CITATION

Marsh, J.K., Zeveney, A., & De Los Reyes, A. (2020, March). Informant discrepancies in judgments about change during mental health treatments. *Clinical Psychological Science*, *8*, 318-332. doi: 10.1177/2167702619894905

Author Notes

Jessecae K. Marsh and Andrew Zeveney, Department of Psychology, Lehigh University, Bethlehem, PA. Andres De Los Reyes, Comprehensive Assessment and Intervention Program, Department of Psychology, University of Maryland, College Park, MD.

Correspondence regarding this manuscript should be addressed to Jessecae Marsh, Department of Psychology, Lehigh University, 36 University Drive, Bethlehem, PA, 18015; Office: 610-758-2941; E-mail: jessecae.marsh@lehigh.edu.

Abstract

Understanding how mental health treatments benefit those who receive treatment comes with a challenge: Often different people involved in treatment have different impressions of the treatment's ultimate effects. How do people reconcile these different reports to understand the true benefit of treatment? In a series of 4 experiments, we tested people's beliefs about how to integrate information from multiple informants for the treatment improvement of child clients. We found that laypeople (Experiments 1, 2, and 3) and professional mental health clinicians (Experiment 4) trust informants they believe to be insightful about the specific disorder but pessimistic about overall improvement. Our findings suggest important future research avenues to better understand how intuitions about reconciling informants influences the process of weighting information from clients and others involved in their care.

Keywords: Informant Discrepancies; Judgment; Multiple Informants; Treatment

Informant Discrepancies in Judgments about Change During Mental Health Treatments

By age 75, nearly half of the people living in the United States will have met diagnostic criteria for a mental disorder at some point in their lives (Kessler et al., 2005). Worldwide and at any one time, roughly half a billion people experience mental, neurological, and/or behavioral problems that require clinical care (Demyttenaere et al., 2004). The breadth and depth of the public health burdens that stem from mental illness necessitate creating practical, evidence-based techniques for assessing and treating these illnesses (Kazdin, 2017). In fact, many scientific disciplines tasked with developing, testing, and implementing mental health treatments have formed entire initiatives for developing standards to identify evidence-based treatments (APA, 2006; CPA, 2005; Chambless & Ollendick, 2001; Rosen & Proctor, 2003; Southam-Gerow & Prinstein, 2014). These standards largely rely on the outcomes of well-conducted, controlled experiments of manualized mental health treatments (e.g., randomized controlled clinical trials [RCTs]; Kazdin & Blase, 2011). Further, these standards inform treatment work across multiple disciplines including those in the social sciences (e.g., Counseling, Criminology, Education, Psychology, Social Work), Nursing, and the medical sciences (e.g., Family Medicine, Internal Medicine, Pediatrics, Psychiatry, Public Health; De Los Reyes & Kazdin, 2006; Doody et al., 2001; Kazdin & Blase, 2011; Volkmar et al., 2014).

In this paper, we seek to improve interpretations of the evidence supporting evidence-based mental health treatments. At the core of this evidence is behavior. Our current techniques for understanding mental illness, planning treatment for these illnesses, and estimating treatment response all hinge on accurate estimates of behavior and behavior change (e.g., observable symptoms codified in diagnostic criteria; APA, 2013). Consequently, the evidence collected in

the RCTs that inform clinical practice guidelines consists of subjective reports about behavior (e.g., surveys and interviews about mental disorder symptoms) collected from clients and significant others in their lives (Hunsley & Mash, 2007; Weisz, Jensen Doss, & Hawley, 2005). For adults, treatment assessments might include client self-reports and also reports from spouses, or for elderly clients their caregivers (Achenbach, Krukowski, Dumenci, & Ivanova, 2005). For children and adolescents (i.e., hereafter referred to collectively as "children"), informants providing these reports will also often include client self-reports but also reports from adult authority figures such as parents and teachers (De Los Reyes, Thomas, Goodman, & Kundey, 2013). As a general rule, assessments conducted in studies of mental health treatments involve estimating treatment response using multiple informants, who each provide reports about a common assessment target, the client (De Los Reyes & Kazdin, 2008).

Roughly 50 years of assessment research supports the idea that across a host of domains (e.g., aggression, anxiety, attention, conduct problems, depression, hyperactivity, substance use) the multiple informants used in mental health practice and research provide reports that reliably and validly estimate clients' levels of mental health functioning (Hunsley & Mash, 2008). The evidence supporting the veracity of individual informant's reports highlights the value of estimating treatment response using multiple, unique perspectives about clients' mental health functioning (Hunsley & Mash, 2007). However, it makes interpreting the outcomes of treatment response assessments all that more difficult. This is because the most robust finding on the assessments used to estimate treatment response is that reports from the multiple informants who complete these assessments display low levels of between-informant correspondence (Achenbach, 2006; Achenbach et al., 2005; Achenbach, McConaughy, & Howell, 1987; De Los Reyes et al., 2015). These *informant discrepancies* robustly manifest in assessments conducted

all over the world, as evidenced by a recent meta-analysis of over 300 studies of these discrepancies across over 30 countries (De Los Reyes et al., 2019). Importantly, these discrepancies have real implications for mental health care. That is, multiple informants' reports about the same client each tell us different things about clients' treatment experiences, from which aspects of a client's mental health warrant care (e.g., anxiety vs. substance use), to whether a treatment administered to address a client's needs led to improved functioning (Hawley & Weisz, 2003; Ogles, Lambert, Weight, & Payne, 1990; Weisz, Weiss, Alicke, & Klotz, 1987). These discrepancies also occur across many service delivery settings, from community mental health clinics to educational settings (De Los Reyes & Kazdin, 2005; De Los Reyes, Cook, Gresham, Makol, & Wang, 2019).

Ideally, informant discrepancies reflect the very reasons for taking a multi-informant approach to assess clients' functioning. That is, clients may *display* the behaviors that triggered the need for treatment differently depending on the social context, such as home, school or work contexts (Achenbach et al., 1987). Further, informants often systematically vary in the social contexts in which they *observe* these same behaviors, such as a child client's parent primarily observing the child at home and the child's teachers at school (De Los Reyes et al., 2015). As such, if informants' reports yield discrepant conclusions about whether a client benefited from treatment, it may be because the treatment varied in the social contexts in which it enacted meaningful change, such as a treatment improving how a client functions at home to a greater degree than how she or he functions at school (De Los Reyes & Kazdin, 2008).

Across mental health conditions as diverse as autism (Lerner, De Los Reyes, Drabick, Gerber, & Gadow; 2017), social anxiety (De Los Reyes, Bunnell, & Beidel, 2013; Deros et al., 2017), and conduct disorder (De Los Reyes, Henry, Tolan, & Wakschlag, 2009) an emerging

body of empirical work supports the ability of informant discrepancies to point to meaningful variations in clients' mental health. In fact, recent theoretical models point to strategies for constructing assessments that allow for detecting those informant discrepancies that reflect meaningful variations in clients' behavior (De Los Reyes, Thomas, et al., 2013).

However, this recent theoretical and empirical work clashes with long-used strategies for addressing informant discrepancies in both clinical and research settings. In research, analytic strategies such as latent variable modeling and algorithms that integrate informants' reports (e.g., AND/OR rules) essentially assume that informant discrepancies are best explained as measurement error (Holmbeck, Li, Schurman, Friedman, & Coakley, 2002; De Los Reyes, Kundey, & Wang, 2011). In practice circumstances with child clients, a number of uncontrolled studies have explored how clinic staff (e.g., interviewers, therapists) when encountering informant discrepancies make clinical decisions, such as treatment planning or estimating treatment response (De Los Reyes et al., 2015). In these studies, the judgments of clinic staff corresponded to a greater degree with the informant who tends to initiate clinical services in the settings in which staff delivered care (e.g., outpatient settings), namely parents (Brown-Jacobsen, Wallace, & Whiteside, 2011; De Los Reyes, Alfano, & Beidel, 2011; DiBartolo, Albano, Barlow, & Heimberg, 1998; Grills & Ollendick, 2003; Hawley & Weisz, 2003; Kramer et al., 2004; Youngstrom, Findling, & Calabrese, 2004). Crucially, this decision-making practice occurs in the absence of compelling data to support the practice. We know of no psychometric data indicating that the informant who initiates clinical services for a client provides "more valid" reports about their functioning than other informants' reports about the same client.

The data integration strategies that we described above reflect symptoms of a larger problem. Specifically, we previously cited clinical guidelines that inform treatment practices. These

guidelines largely rely on informants' clinical reports to identify proper treatment practices (Weisz et al., 2005). Yet, no clinical guidelines exist on how to properly interpret the multi-informant data used for treatment guidelines (Beidas et al., 2015). Further, all studies to date on the judgments that mental health professionals make when confronted with informant discrepancies are based on correlational field work or controlled observations of clinical assessments. Controlled experimentation can address these issues. For instance, studies that experimentally manipulate participants' exposure to informant discrepancies in reports about behavior change can help us understand whether variation in these discrepancies causally produce variations in clinical decision-making. Like the RCTs that have informed treatment guidelines, controlled experiments of informant discrepancies may improve both the accuracy of decision-making in clinic settings, as well as the validity of clinical practice guidelines.

The purpose of this study was to advance the literature on informant discrepancies in assessments of behavior change. In a series of 4 experiments, we tested how people integrate information from two informants who report differing levels of success for a treatment of a mental health condition. We lead off this exploration with three experiments exploring lay participants' judgments for different informant dyads. Lay participants are an important test group for understanding how discrepant information is assimilated given the gateway role laypeople play in mental health care: untrained laypeople may be the first ones to notice impairment in themselves or a loved one and initiate the beginning and continuation of care for that problem (Hunsley & Lee, 2014; Marsh & Romano, 2016). Once that care is initiated, laypeople also bear the burden of deciding whether to continue care. For example, parents who have started their children on treatment have to make a decision about whether the costs of that treatment outweigh its benefits to symptom improvement. To make such a decision, they may be

talking to their children, talking to their children's teachers, and listening to their own intuitions about improvement, before ever consulting their children's doctors for treatment adjustment. Likewise, a manager may have to assess what is the right course of action for a troublesome employee when two co-workers report differing accounts of the employee's mental health. At a more general level, understanding how laypeople integrate discrepant sources to understand how someone is improving in treatment provides important insight into how consumers of mental health care may be assessing the outcomes and quality of the health care they are engaged in receiving. We follow up this exploration with a fourth experiment that examined how experts integrate discrepant informants' reports. In Experiment 4 we determine whether effects observed with laypeople's decision making generalize to decisions made by providers of care.

Experiment 1

In Experiment 1 we tested how people incorporate information from two informants who often differ in how they are reporting change in mental health treatment. We started with one of the most common dyads who offer conflicting reports, a child experiencing mental health concerns and the child's parent (Achenbach et al., 1987; De Los Reyes et al., 2015). We varied the disorders about which these informants were reporting and which of these two informants reported greater improvement in symptoms to test how participants integrated information across informants.

People may integrate conflicting treatment reports in several different ways. People may believe that certain types of informants (e.g., parents) are always more reliable than other types of informants (e.g., child clients; Loeber, Green, & Lahey, 1990). In this case, people's estimates of clinical factors should track, or at least be more heavily weighted toward, the perceived reliable informant. Alternatively, people may align with the informant they believe has more

insight into the health problem in question. For example, for assessments of internalizing conditions people may see a child client as a "better informant" for change in the internal state symptoms that characterize these conditions, and the parent as a "better informant" for behaviors that characterize assessments of externalizing conditions (De Los Reyes & Kazdin, 2005). In this situation, people's estimates of treatment improvement may reflect different informants *depending* on the condition in question. As another possibility, the importance of continuing treatment until symptoms have completely remitted may cause people to be more hesitant to trust informants who display relatively more optimistic impressions about treatment improvement. As such, people could side with informants who are more pessimistic in their estimates of treatment improvement. Finally, people may have instincts that the effects of certain treatments are more accessible to people who are not the client. All of these different possibilities necessitate experimental tests of how people integrate information when informants provide discrepant estimates of treatment improvement.

Method

Participants. We recruited 100 Amazon Mechanical Turk (MTurk) workers (age range: 19 – 64, $M = 34.5$). The use of MTurk workers provides data comparable to in-person lab studies (Mason & Suri, 2012), while allowing for collection of a broader sample than typically available through data collection focused on college campuses (for an overview of MTurk use in data collection see Buhrmester, Talaifar, & Gosling, 2018). For all MTurk samples, we limited our pool to US-based MTurk workers who have a 95% or better approval rate.[1] Participants predominantly self-identified their gender as male (60%; 39%= female; 1% = prefer not to

[1] A 95% approval rate criterion has been shown to be successful in excluding inattentive / fake respondents (Peer, Vosgerau, & Acquisti, 2014).

respond). The majority of participants reported their ethnicity as non-Hispanic (89%; Hispanic = 8%; prefer not to respond = 3%) and their race as white (79%; Black or African-American = 8%; Asian = 6%; American Indian or Alaska Native = 2%; Native Hawaiian or Other Pacific Islander = 0%; multiple races reported = 3%; prefer not to answer = 2%). Participants most often reported that the highest degree they currently held was a bachelor's degree (42%; High school or equivalent = 34%; Associate's degree = 5%; Master's degree [e.g., M.A., MSW, MBA] = 16%; Ph.D. = 1%; M.D. = 1%; Other not specified degree = 1%). We dropped any participants from analyses if they reported being a licensed mental health professional or for reporting they worked in a facility for mental health care. None of the participants in this experiment met those exclusion criteria.

Materials. We created vignettes that asked participants to consider a child client and the child's parent reporting on the child's treatment progression. We described in the vignette that the child had been receiving treatment from a care provider who was now asking the child and the parent to rate how much the symptoms of the child's mental health concern had improved. We described the child and the parent as separately making these ratings on a scale of 0, no improvement at all, to 100, largest possible improvement.

To allow easy comparison of the treatment improvement estimates, we presented the estimates from the hypothetical child and parent to participants in a bar graph (see Figure 1). The x-axis of the graph was labeled as "Person Providing Rating" and the y-axis was labeled "Amount of Improvement in {Disorder name} symptoms", with the appropriate disorder name filled in by condition. Each informant's rating was presented in a separate, labeled bar. This presentation allowed participants to easily see the relative improvement ratings for each participant and the absolute values the informants reported. The graphs depicted either a rating in

the high 30s for one rater and approximately 80 for the other rater, or one rating in the low 20s for one rater and mid 60s for the other. The absolute difference between the two informants' ratings was the same for all cases (i.e., 42 points). This provided for some variation across graphs but the same difference between raters.

Insert Figure 1 here.

To isolate influences on how discrepant informant reports are integrated, we manipulated several different elements of the vignettes. As we described previously, people may have intuitions about who is a better informant, a child or a parent. If this is the case, then participants should consistently weight one informant (child or parent) more than the other. To be able to detect whether participants' ratings tracked one informant in particular, we manipulated across vignettes whether the child or the parent was reporting relatively less symptom improvement. We refer to this informant who produced a lower estimate as the "pessimistic" informant. To be clear, the pessimistic informant is not saying that the treatment did not work; as can be seen in Figure 1 both raters are reporting improvement. Rather, the pessimistic informant is reporting less improvement than improvements rated by the relatively optimistic informant. Participants' estimates always being closer to one informant regardless of whether the informant is the pessimist would provide evidence toward more heavily considering certain specific informant reports. Alternatively, participants' estimates always being more aligned with the informant who is in the pessimist position would provide evidence toward favoring the informant who is more conservative, regardless of who that was. We also hypothesized that participants may favor different informants who are seen as more insightful about the condition in question. To test for this possibility, we used two internalizing conditions (an anxiety disorder, depression) and two externalizing conditions (attention-deficit/hyperactivity disorder, conduct disorder). We used

these conditions given that they comprise some of the most commonly referred concerns for children's mental health care (Hunsley & Lee, 2014). We can compare ratings for the internalizing and externalizing conditions to see if different informants are preferred for each. Finally, we manipulated the type of treatment the patient was described as experiencing: medication or therapy. It is possible that people see a patient as having more insight into whether a medication makes a difference in the way symptoms are experienced, in line with the hypothesis that people may favor informants seen as more insightful. Across these manipulations, we can determine which informant is favored and how that varies by their relative pessimism, the disorder being assessed, and the type of treatment being used.

Procedure. Participants began the experiment by reading an online consent form and indicating their consent. Each participant then read and made ratings for four client cases presented in a random order, with each case depicting one of our four disorders. Participants were counterbalanced to receive either a child or a parent pessimist for both of the internalizing cases. This design allows us to collapse across the two different internalizing ratings to create a mean treatment improvement score for that disorder type, given that type of pessimist. For the externalizing cases, participants then were assigned to receive the pessimist they *did not* see for the internalizing conditions. That is, a given participant saw one pessimist for both of the internalizing disorders (e.g., child) and the other pessimist for both of the externalizing disorders (e.g., parent). This design choice means that participants did not see every possible combination of disorder type and pessimist and instead saw one of these two combinations: internalizing – child pessimist cases and externalizing – parent pessimist cases, or internalizing – parent pessimist cases and externalizing – child pessimist cases). We made this choice to minimize the repetitiveness of the experiment for online data collection and the obviousness of our

experimental manipulation. Finally, we manipulated treatment condition between subjects, such that all cases participants read described the same treatment (therapy or medication), with participants counterbalanced to each condition.

For each case, we asked participants to rate the following question:

> "From what you have read about this child, how much do you think the child's anxiety symptoms actually improved? Please indicate the level improvement you think occurred on the slider scale below. You can use any number between 0 (no improvement at all) and 100 (largest possible improvement).".

In essence, participants were asked to make the same judgment that the child and parent informants were described as making in the presented materials. Participants answered the question on a slider scale that allowed them to choose any point between 0 and 100.

After rating the four cases, participants completed a series of three post-experiment questions. First, participants were asked to explain in an open-ended text box how they made their estimates. This question helped us screen for any online participants that were randomly or incoherently responding. Participants then rated two questions that assessed how they thought about the specific treatment they rated in the main questions of interest (medication or therapy). First, they rated how likely it would be for different people to notice the effects of the treatment on a child, including the child herself/himself, parent, teacher, therapist, doctor, friend, and classmate. Then, participants rated how likely the treatment would be to help each of the four disorders used in the main experiment. These last two questions were included as piloting for future work and are beyond the scope of this paper. Participants finished the experiment by completing a series of demographics questions (see Participants section for questions). For this and all of the experiments in this paper, informed consent was obtained from all participants and

the first author's Institutional Review Board approved the experimental protocols.

Data-Analytic Plan

The goal of our analyses was to determine if participants reliably defer to one informant or another, and if this varies by the type of disorder being assessed and which informant is more pessimistic about change. If participants are equally weighting both informants, then we could imagine their estimates of symptom improvement should represent an average of the scores provided by the two informants. Taking Figure 1 as an example, completely equally weighting the two informants and perfectly averaging their responses would result in a participant response of 44. However, if participants are favoring one informant over the other, then we would expect responses that differ from that average value, with the direction in which it differs being indicative of which informant is being more heavily weighted. With this logic, subtracting the average value of informants' reports presented in a vignette from a participant's response to that same vignette will provide a metric of how much the participant's response differed from the average value (i.e., absolute value of the difference) and towards which informant the participant skewed (i.e., sign of the difference). This procedure essentially involved subtracting a constant value from all participants' scores for a particular vignette in order to weight each participant's scores relative to the average score depicted in that vignette. Thus, a weighted score of zero would indicate a participant equally weighted the two informants and averaged their reports when making a decision as to treatment improvement; a positively weighted score would reflect a participant favoring the optimist informant more in estimates; and a negatively weighted score would reflect a participant favored the pessimist more in estimates.

Note that this procedure differs from traditional uses of difference scores, in which one individual difference variable is subtracted from another, such as is the case with change scores

in treatment research (e.g., pre-treatment score minus post-treatment score; Cronbach & Furby, 1970) or research examining the differences between reporters in mental health assessments (Laird & De Los Reyes, 2013). Essentially, our procedure involved benchmarking or norming participant ratings relative to the values representing informants' improvement ratings that we created and presented in each of the vignettes. Importantly, to create a realistic set of vignettes that participants found believable, we had to vary the actual values presented in the vignettes representing informants' improvement ratings. At the same time, we also ensured that the informants' ratings we presented in each vignette always yielded the same average value. Thus, our procedure for norming participant ratings allowed us to compare ratings across vignettes which had numerical variations among informants' improvement ratings.[2]

We normed each participant's score for each case as described above. We then averaged the two weighting scores across the same disorder type to create an average informant weighting score for internalizing disorders and an average score for externalizing disorders. We submitted the weighting scores to a multilevel model (MLM). MLMs are well equipped to analyze repeated measure data and their use of maximum likelihood estimation means there is no requirement for complete data; as such they can handle the nature of our data set where participants did not make ratings for every possible cell of our design in a way that repeated-measures ANOVA cannot. We entered our factors of treatment type (medication vs. therapy; between), disorder type (internalizing vs. externalizing; within), and pessimist (child vs. parent; within) as fixed effects into the MLM model, allowing us to test the significance of the main effects and their interactions through *F* tests. Since we are asking ANOVA-style questions in our design, we

[2] Our procedures are consistent with the suggested proper way to interpret difference scores and their relations to criterion variables (Laird & Weems, 2011; Laird & De Los Reyes, 2013).

present the ANOVA-style output of the MLM analyses to match the questions we are asking. Data were analyzed in IBM SPSS (Version 24) using the repeated statement in the MIXED procedure and a compound symmetry covariance matrix. Significant interactions were explored through the multilevel modeling procedure with simple effect analyses and follow-up Sidak-corrected two-tailed *t* tests. Cohen's *d* (*d*) is provided as a measure of effect size for interaction *t* tests, calculated from the means and standard deviations from the estimated marginal means analysis of the MLM analyses. Cohen's *d* for one-sample *t* tests was calculated with the means and standard deviations of the raw data. In the main text we present graphs that depict the group means generated in the estimated marginal means analysis for the interaction comparisons.

Results

Our main question in these analyses was: Do people weight some informants more than others and does that vary by the type of mental disorder in question? We did not find a main effect of treatment type, $p = .505$, or any significant interactions that involved treatment type, ps $> .17$. This suggests that people do not have intuitions as to informants being more or less insightful depending on the used treatment. As such, we collapse across this manipulation and present the mean weighting scores in Figure 2a collapsed across treatment type. As a reminder for interpreting Figure 2a, scores of zero represent an averaging strategy that equally weights both informants, and negative weighting scores indicate weighting the pessimist informant's estimates more heavily. We did not find a main effect of pessimist, $p = .105$, or of disorder type, $p = .144$. However, we found a significant interaction of disorder type and pessimist, $F(1, 96) = 32.68$, $p < .001$, suggesting that which informant receives more weight depends on the disorder about which the informant is reporting. Specifically, a parent pessimistic about the amount of change occurring in their child was weighted more heavily when reporting about an externalizing

condition than an internalizing condition, $p = .005$. $d = .563$. Conversely, a child pessimist was given significantly more weight when reporting about an internalizing condition than an externalizing condition, $p < .001$, $d = .995$. In this way, we found alignment with certain pessimistic informants, depending on the condition being assessed.

Insert Figure 2 here.

Another way to determine if participants are preferentially weighting one informant over another is to compare through one-sample *t* tests the weighting scores to a value of 0, the value representing equal weighting of both informants. We did not find a significant difference for internalizing conditions when the parent informant was more pessimistic, $p = .906$, $d = .017$, or when the child informant was more pessimistic for externalizing conditions, $p = .832$, $d = .041$. However, weighting scores were significantly different from zero for externalizing conditions when the parent was more pessimistic, $t(47) = 3.99$, $p < .001$, $d = .574$, and for internalizing conditions when the child was more pessimistic, $t(47) = 8.75$, $p < .001$, $d = 1.03$.

Discussion

Our findings suggest that people do not always favor one informant (e.g., the parent) or always are conservative and trust the pessimist. Rather, who people believe is a good informant for behavior change depends on the type of disorder being reported and on whether that informant is relatively pessimistic about change. Of note, participants never weighted the relatively optimistic informant more heavily in any comparison, which would have been indicated by positive weighting scores. Participants either equally weighted the informants or showed deference to the more pessimistic informant. This pessimist weighting was specific to informants who should have had the most insight into the described condition. Namely, participants weighted more heavily a pessimistic child informant only for an internalizing

condition and a pessimistic parent informant only for an externalizing condition. Finally, this effect held regardless of the type of treatment being administered, suggesting that the insight an informant can gather is just as strong when medication is used as therapy.

Our results suggest that people trust who they perceive as insightful pessimists when making estimates of behavior change. To test the robustness of this phenomenon, in Experiment 2 we expand to testing intuitions about another dyad that often reports discrepant information about a child client's clinical improvement: a child and the child's teacher.

Experiment 2

In Experiment 2 we tested whether the findings of Experiment 1 are generalizable to other dyads, namely when a different external informant reports on behavior. If the phenomenon of Experiment 1 generalizes to this dyad, then we would expect a child client informant to be seen again as the "insightful informant" for internalizing conditions and have his/her reports weighted more heavily. For externalizing conditions, the teacher should function in the same role as the parent and be weighted more heavily when a pessimist about improvement in these conditions.

Method

Participants. We recruited a new sample of 100 US-based Amazon Mechanical Turk workers. Two participants were dropped from data analysis for reporting they were licensed mental health professionals or worked in a mental health care setting, resulting in a final sample of 98 (age range: 21 – 62, $M = 34.1$). Participants self-identified their gender equally often as female (50%) and male (50%). The majority of participants reported their ethnicity as non-Hispanic (92%; Hispanic = 6%; prefer not to respond = 2%) and their race as white (82%; Black or African American = 7%; Asian = 8%; multiple races reported = 2%; prefer not to answer =

1%). Participants most often reported that the highest degree they currently held was a bachelor's degree (46%; High school or equivalent = 44%; Associate's degree = 2%; Master's degree= 8%).

Procedure. The procedure for Experiment 2 was identical to Experiment 1 with the following exceptions. We substituted teacher for parent in all of the materials. Additionally, we dropped the medication versus therapy manipulation, using therapy as the stated treatment in all materials. Participants made all the same ratings as in Experiment 1.

Data-Analytic Plan

The design is the same as Experiment 1, with the exception of dropping the treatment type variable. We entered our factors of disorder type (internalizing vs. externalizing; within), and pessimist (child vs. parent; within) as fixed effects into an MLM model. All other elements of the data analytic approach were the same as in Experiment 1.

Results and Discussion

Figure 2b depicts the mean weighting scores across conditions. As in Experiment 1, we found a significant interaction of disorder type and pessimist, $F(1, 96) = 16.41$, $p < .001$, and no significant main effects, ps > .35. A teacher pessimist was weighted more heavily when reporting about an externalizing condition than an internalizing condition, $p = .009$, $d = .532$. A child pessimist was weighted significantly more when reporting for an internalizing condition than an externalizing condition, $p = .003$, $d = .607$. One-sample t tests comparing the weighting scores to 0 did not find a significant difference for internalizing conditions when the teacher informant was more pessimistic, $p = .299$, $d = .136$, or when the child informant was pessimistic for externalizing conditions, $p = .812$, $d = .033$. Weighting scores were significantly different from zero for externalizing conditions when the teacher was more pessimistic, $t(49) = 4.71$, $p < .001$, $d = .670$, and for internalizing conditions when the child was more pessimistic, $t(49) = 3.67$, $p =$

.001, $d = .574$.

The findings of Experiment 2 provide more evidence that client informants are trusted more for internalizing problems and external informants (e.g., parents and teachers) are trusted more for externalizing conditions when they are pessimistic about behavior change. We have discussed this weighting as about the informant who is thought to have more insight into the symptoms of the condition (the actual client for internalizing conditions, others for externalizing conditions). Is this insight about who has the *absolute* most insight into a problem, or about who in the pair of informants has *relatively* more insight into a problem? We test this question in Experiment 3 by testing a new dyad: parent and teacher informants.

Experiment 3

In Experiment 3 we tested how people weight the reports of a parent and a teacher informant who differ on reported amounts of behavior change for a child. If the results of Experiment 1 and 2 reflect beliefs about absolute insight, that is only clients can be trusted for internalizing conditions and external reporters for externalizing conditions, then we would expect that parents and teachers should be seen as equally good reporters for externalizing conditions and equally poor reporters for internalizing conditions. Looking at the conditions we used in our previous experiments, this could take the form of participants averaging the reports of both informants in all conditions. Alternatively, our previous results may reflect a belief in relative insight among informants. That is, in a given dyad people may look for which of the two informants seems to have *relatively* more insight for the given disorder and trust their pessimistic responses.

Method

Participants. A total of 102 US-based Amazon Mechanical Turk workers who did not

participate in the previous experiments completed the experiment. One participant was dropped from data analysis for reporting working in a mental health care setting, which resulted in a final sample of 101 (age range: 19 – 66, $M = 34.3$). Participants' self-identified gender was evenly balanced between female (49%) and male (50%; 1% = prefer not to respond). The majority of participants reported their ethnicity as non-Hispanic (95%; Hispanic = 3%; prefer not to respond = 2%) and their race as white (84%; Black or African American = 5%; American Indian or Alaska Native = 1%; Asian = 7%; multiple races reported = 2%; prefer not to answer = 1%). Participants most often reported that the highest degree they currently held was a bachelor's degree (45%; High school or equivalent = 41%; Associate's degree = 2%; Master's degree = 10%; Ph.D. = 1%; M.D. = 1%).

Procedure. The procedure was identical to Experiment 2 with the exception that the dyad used was parent and teacher. All ratings were identical to Experiment 2.

Results and Discussion

The same method of constructing weighting scores and the same multilevel model was used as in Experiment 2. Figure 2c depicts the mean weighting scores across conditions. We found a significant interaction of disorder type and pessimist, $F(1, 99) = 10.66$, $p = .002$, and no significant main effects, ps > .16. A teacher pessimist was weighted more heavily when reporting about an externalizing condition than an internalizing condition, $p = .044$, $d = .404$. A parent pessimist was weighted significantly more when reporting for an internalizing condition than an externalizing condition, $p = .007$, $d = .543$. One-sample t tests comparing the weighting scores to 0 did not find a significant difference for internalizing conditions when the teacher informant was more pessimistic, $p = .728$, $d = .043$, or when the parent informant was pessimistic for externalizing conditions, $p = .164$, $d = .215$. Weighting scores were significantly different from

zero for externalizing conditions when the teacher was more pessimistic, $t(49) = 3.01$, $p = .004$, $d = .447$, and for internalizing conditions when the parent was more pessimistic, $t(49) = 2.38$, $p = .021$, $d = .328$. Overall, the pattern for teachers is the same as in Experiment 2, with their pessimistic ratings being given more weight in externalizing conditions. Parents took on the role previously seen for child clients, with their pessimistic ratings being given more weight in internalizing conditions.

To this point we have tested the beliefs of lay participants. An open question is whether professional mental health clinicians would show the same weighting in their decisions as laypeople. We test this question in Experiment 4.

Experiment 4

Experiment 4 uses the base dyad of a child and a parent informant to test how professional mental health clinicians integrate conflicting reports. If professional experience teaches clinicians different ways to integrate informant reports, then we would expect a different pattern of responding in Experiment 4 compared to our previous three experiments, such as clinicians always trusting a specific informant. However, if our previous results reflect a fundamental way of thinking about incorporating multiple respondents, then we may see a similar pattern as previous experiments with these professional clinicians.

Method

Participants. To find a potential participant base, we conducted online searches for licensed mental health professionals. We contacted clinicians through email addresses they posted in their online web presence. Our recruitment email included a link to our study so that clinicians could directly access the experiment. We sent emails to 417 posted email addresses representing clinicians from across geographical regions of the United States. Some of these addresses were

individual email addresses for a clinician, and some were more general emails for the practice at which a mental health professional worked. Upon completion of the study, participants were sent to a separate survey that collected their email addresses for payment.

A total of 32 licensed mental health professionals who currently work or have worked in a mental health care setting completed the experiment (age range: 31 – 73, *M* = 45.9). Participants predominantly self-identified their gender as female (84%; 16% = male). The majority of participants reported their ethnicity as non-Hispanic (97%; Hispanic = 3%) and their race as white (81%; Black or African American = 9%; Asian = 6%; prefer not to answer = 3%).[3] The most common degree held by participants was a Ph.D. (66%; Psy.D. = 9%; MSW = 16%; M.S./M.A. = 9%) with the majority of participants reporting being licensed as a psychologist (69%; social worker = 16%; counselor = 9%; LMFT = 6%). Participants reported having seen clients in a clinical setting for on average 16.8 years (range: 2 – 45). The majority of participants reported having worked in private practice (91%), with a large number of participants having worked in other settings (community clinic = 44%, hospital = 40%, university counseling center = 19%, department clinic = 13%, other settings = 6%). An additional three participants completed the experiment but did not have their data analyzed because they reported either not being a licensed mental health professional or having never worked in a mental health setting.[4]

[3] Published statistics for licensed psychologists and social workers suggest our sample approximates the representation of clinicians more broadly. The most recent American Psychological Association reported demographics for active psychologists are as follows: mean age of 49, 67% female, 85% white, 3% Asian, 4% Black or African-American, and 6% Hispanic (Hispanic was included as a race category, APA, 2018). The most recent Council on Social Work Education reported demographics for active social workers are as follows: 85% female, 73% White, 19% Black or African-American, 3% Asian, 0.5% American Indian, and 9.5% Hispanic (Hispanic was a separate ethnicity category from race, as in our data collection; Salsberg et al., 2017).

[4] Assuming that each email address represented one clinician, we had an overall completion rate of 8.4%. However, since some of these emails went to practice emails, we may be underestimating the number of clinicians we had

All participants who completed the experiment were compensated with a $10 gift certificate to a major online retail store for their participation.

Procedure

We used the same basic procedure as Experiment 2 and 3 with the following changes. The dyad described a child and a parent informant, the same dyad as Experiment 1. Instead of specifying therapy as the treatment, we used the generic phrasing of "Imagine a child client who is receiving a treatment for ...", keeping treatment as the term used in the materials. This change helped us avoid any biases our expert participants may have had about the use of any specific treatment we could have listed for the given conditions. Participants answered a shorter set of post-experiment questions. First, they rated how likely different informants would be to notice the effects of the treatment on a child, with the informant list being the child herself/himself, parent, teacher, therapist, friend, and classmate. Participants then indicated what type of general treatment form they were thinking of when rating the four vignettes: therapy (47%), medication (0%), a different type of treatment for each vignette (16%), no specific type of treatment (31%), or other (6%).

Results and Discussion

The same method of constructing weighting scores and the same multilevel model was used as in Experiment 3. Figure 3 depicts the mean weighting scores across conditions for our clinician participants. Overall, we find the same pattern in professional clinicians as we have in our lay samples. We found a significant interaction of disorder type and pessimist, $F(1, 30) =$

---

potential to contact. Likewise, our email invitation came from the first author's lab Google account, which could have resulted in some number of our emails being filtered as spam, causing an overestimating of the potentially contacted clinicians.

8.99, $p = .005$, and no significant main effects, ps > .57. A parent pessimist was weighted more heavily when reporting about an externalizing condition than an internalizing condition, $p = .022$, $d = .838$. A child pessimist was weighted significantly more when reporting for an internalizing condition than an externalizing condition, $p = .008$, $d = .986$. One-sample t tests comparing the weighting scores to 0 did not find a significant difference for internalizing conditions when the parent informant was more pessimistic, $p = .678$, $d = .108$, or when the child informant was pessimistic for externalizing conditions, $p = .823$, $d = .080$. Weighting scores were significantly different from zero for externalizing conditions when the parent was more pessimistic, $t(16) = 3.31$, $p = .004$, $d = .730$, and for internalizing conditions when the child was more pessimistic, $t(16) = 5.52$, $p < .001$, $d = .906$.

Insert Figure 3.

General Discussion

Across four studies, we showed that people have specific beliefs about how to weight impressions of behavior change provided by discrepant informants. Specifically, who lay and mental health provider participants trusted depended on the type of mental health condition described: clients or informants who had relatively more experience with the client were trusted more for internalizing conditions, whereas informants who were the most removed from the client were trusted more for externalizing conditions. Importantly, we observed this finding only when those trusted informants provided more pessimistic ratings about overall improvement, relative to the other informants' ratings presented in the vignettes. Never was an optimistic informant trusted more than a pessimistic informant.

Our results suggest that people are sensitive to the nature of different disorders and have different expectations as to the informants who may have insight into those disorders.

Participants differentiated internalizing and externalizing conditions, suggesting that this distinction used in clinical practice is recognizable to non-clinically trained participants. There is an intuitive nature to trusting clients for reports of internalizing symptoms. For example, it seems sensical that the client may have better insight into whether they are still feeling the low mood that comes with depression. However, why is it that clients are not seen as equally insightful into externalizing symptoms? Engaging in the actions of a behavioral symptom (e.g., destroying property for conduct disorder) should be equally noticeable to an observer and to the client. It is possible that when people think of externalizing symptoms they may think of symptoms they conceptualize as more opaque to the actor. For example, people may think of the conduct disorder symptom of bullying as an action a client does not notice he is exhibiting in the moment. Consistent with these notions, prior work indicates that under some circumstances, child clients may provide psychometrically suspect self-reports of externalizing concerns (e.g., attention and hyperactivity, conduct problems; McMahon & Frick, 2005; Pelham, Fabiano, & Massetti, 2005). While this opaqueness could explain our findings when the client was part of the dyad (Experiment 1, 2, and 4), it does not help explain the findings of Experiment 3 where the client was not in the dyad. In Experiment 3 we found the same pattern of the previous experiments of one informant being trusted for externalizing conditions and the other trusted for internalizing conditions. One explanation for the persistence of this pattern is that people do not weight one informant more than another because of a belief in some form of absolute insight (e.g., clients know more about internalizing conditions). Rather, people may view informants in relation to each other and make a decision about who should know *relatively* more than the other (e.g., in this pair, this informant should know more about internalizing conditions than the other). Such a relative comparison process would explain why in Experiment 3 parents are treated like

the child client of previous experiments. Future research can investigate how relative versus absolute beliefs in informants' accuracy could influence integrating information from discrepant informants.

Why do our participants trust the insightful pessimists and not the insightful optimists? Judgments that overweight negative relative to positive evidence have been seen in a range of other domains, such as processing negative information more and weighting negative information more in the formation of impressions about other people (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Ito, Larsen, Smith, & Cacioppo, 1998; Peeters & Czapinski, 1990; Skowronski & Carlston, 1989). The negativity bias is often explained from an evolutionary perspective: it is often more costly to miss negative information than positive information (Baumeister et al., 2001). In the case of mental health treatment, focusing on the relatively more negative informant means avoiding the cost of leaving treatment too early, before treatment has yet to exact beneficial effects on client functioning (while the cost of staying in treatment longer than necessary might be less impactful).

Notably, our participants were not overweighting all negative reports, just negative reports coming from informants seen as more insightful. It is possible that in our task that provided relatively minimal information about how to integrate the information received from the two informants, our participants needed some type of justification to feel validated in weighting one informant more than the other. A similar idea comes from the social psychology literature where people have been shown to resist using stereotypes to judge people until they feel there is some minimal amount of justification to allow for the stereotype's use (Yzerbyt, Schadron, Leyens, & Rocher, 1994). Perceiving one informant as more insightful about a condition may have provided a type of minimal justification and allowed participants to express their negativity bias. Without

this justification, participants did the seemingly more neutral action of basically averaging the two informants.

Research and Theoretical Implications

Our findings have several implications for how we can think about improving clinical assessment and research on informant discrepancy more generally. Our studies show who both laypeople and professionals are likely to trust when information comes in that is conflicting about a child client, not who they should actually trust. Clinical research should be undertaken to test whether the different constituents in the process of referring clients for care are actually more or less insightful than each other. If research evidence suggests that the intuitions of our participants are accurate, then we have identified an important place where the training of mental health professionals can be updated to reflect who should be trusted more in situations in which informants' reports vary as to treatment improvement. However, even if research does not support our participants' intuitions, it is important to recognize that this is how people *think* they should be reconciling discrepancies among informants' ratings of treatment improvement. It is crucial to train mental health professionals on the ways they may naturally favor different informants so that if this favoritism does not have a basis in reality, it can be corrected. In future work, controlled experiments can decipher whether favoring one informant over another can be reduced with modifications to the process of collecting clinic data, such as objective records of behavior change (or lack thereof).

Limitations

A few of the design constraints of our experiments suggest directions for future research. One, we only used reports about child clients. Would people have similar intuitions about who has insight into adult clients' disorders? We suggest that if people are willing to endorse children

as having insight into internalizing disorders, that it is likely they would be willing to endorse adult clients having the same insight. An interesting question is whether these adult clients would still be seen as having relatively less insight into externalizing conditions. Future research should use adult pairs (e.g., adult client and spouse) with internalizing and externalizing conditions common in adults (e.g., depression, substance-use disorder) to see if lay and professional samples similarly see one member of the pair as having more insight into symptom improvement than the other.

A second limitation of our research is that we used a set of four disorders, rather than a more encompassing list of disorders. There may exist a set of disorders that people do not believe any age client has insight into. For example, people may not believe people experiencing thought disorders such as schizophrenia have insight into what they are experiencing. In such disorders, would people ever trust the client to have insight into improvement, or would people always defer to pessimistic others? This is an important area of research to further explore how people's views of different types of disorders influences how they reconcile discrepant informant reports.

A third limitation is that our data collection for Experiments 1 - 3 came from Amazon Mechanical Turk. While MTurk is a sound source for collecting online data (Burhmester et al., 2018; Mason & Suri, 2012; Peer et al., 2014), we used it to sample from a broad range of people rather than actual parents or patients who were trying to reconcile discrepant information.[5] Parents reconciling discrepant reports of their children and their children's teachers have much more knowledge about patterns of symptoms, accuracy of reports over time, and other associated

[5] MTurk workers in comparison to epidemiological samples have been shown to report higher than average levels of symptoms from a range of disorders, including, importantly for our study, social anxiety and depression (Arditte, Çek, Shaw, & Timpano, 2016). As such, it is possible that our sample actually has more experience than the average layperson thinking about our disorders of interest and how to integrate symptom information.

information that could help integrate informant reports.

A fourth limitation involves our sole focus on treatment *improvement*. That is, treatments might vary in whether they ultimately yield beneficial effects and some may yield harmful effects (see Lilienfeld, 2007). Future research seeking to replicate and extend our findings might expand use of our experimental paradigm to situations in which informants are tasked to provide ratings of worsening following treatment. Similarly, we also see value in testing this paradigm to evaluate people's impressions of informant discrepancies in impressions of treatment change for mental health domains beyond those examined in this study (e.g., autism and substance use).

Concluding Comments

Overall, our study provides evidence of how laypeople and mental health professionals integrate information from discrepant informants. We have provided evidence that for pairs of informants, people believe informants when they are pessimistic about change for a condition they are seen to have relatively more insight into compared to the other informant. Future research can explore how accurate these perceptions are, and regardless of that accuracy, how they guide the availability and maintenance of mental health care.

Authorship Statement

All authors developed the study concept. All authors contributed to the study design. Testing, data collection, and data analysis were performed by J. K. Marsh and A. S. Zeveney. All authors provided data interpretation, drafted the manuscript, and provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A180032 to University of Maryland at College Park. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors have no competing interests that influenced this work.

References

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science, 15,* 94-98. doi:10.1111/j.0963-7214.2006.00414.x

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin, 131,* 361-382. doi:10.1037/0033-2909.131.3.361

Achenbach, T. M., McConaughy, S. H., & Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232. doi:10.1037/0033-2909.101.2.213

American Psychiatric Association. (2006). *American Psychiatric Association Practice Guidelines for the treatment of psychiatric disorders: compendium 2006*. American Psychiatric Pub. doi:10.1176/appi.ajp.2015.1720501

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author. Canadian Psychiatric Association. (2005). Clinical practice guidelines: treatment of schizophrenia. *Canadian Journal of Psychiatry*, *50*, 7S-57S. doi:10.1176/appi.books.9780890425596

American Psychological Association. (2018). Demographics of the U.S. psychology workforce [Interactive data tool]. Retrieved from http://www.apa.org/workforce/data-tools/demographics.aspx

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, *28*, 684-691.

doi: 10.1037/pas0000217

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323-370. doi:10.1037/1089-2680.5.4.323

Beidas, R.S., Stewart, R.E., Walsh, L., Lucas, S., Downey, M.M., Jackson, K., . . . Mandell, D.S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice, 22,* 5-19. doi:10.1016/j.cbpra.2014.02.002

Brown-Jacobsen, A.M., Wallace, D.P., & Whiteside, S.P.H. (2011). Multimethod, multi-informant agreement, and positive predictive value in the identification of child anxiety disorders using the SCAS and ADIS-C. *Assessment, 18,* 382-392. doi:10.1177/1073191110375792

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*, 149–154. doi:10.1177/1745691617706516

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52*, 685-716. doi:10.1146/annurev.psych.52.1.685

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we?. *Psychological Bulletin*, *74*, 68-80. doi: 10.1037/h0029382

De Los Reyes, A., Alfano, C.A., & Beidel, D.C. (2011). Are clinicians' assessments of improvements in children's functioning "global"? *Journal of Clinical Child and Adolescent Psychology, 40,* 281-294. doi:10.1080/15374416.2011.546043

De Los Reyes, A., Augenstein, T.M., Wang, M., Thomas, S.A., Drabick, D.A.G., Burgers, D., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and

adolescent mental health. *Psychological Bulletin, 141*, 858-900. doi:10.1037/a0038498

De Los Reyes, A., Bunnell, B.E., & Beidel, D.C. (2013). Informant discrepancies in adult social anxiety disorder assessments: Links with contextual variations in observed behavior. *Journal of Abnormal Psychology, 122,* 376-386. doi:10.1037/a0031150

De Los Reyes, A., Cook, C.R., Gresham, F.M., Makol, B.A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology*, *74*, 74-89. doi: 10.1016/j.jsp.2019.05.005

De Los Reyes, A., Henry, D.B., Tolan, P.H., & Wakschlag, L.S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology, 37,* 637-652. doi:10.1007/s10802-009-9307-3

De Los Reyes, A., & Kazdin, A.E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483-509. doi:10.1037/0033-2909.131.4.483

De Los Reyes, A., & Kazdin, A.E. (2006). Conceptualizing changes in behavior in intervention research: the range of possible changes model. *Psychological review*, *113*, 554-583. doi:10.1037/0033-295X.113.3.554

De Los Reyes, A., & Kazdin, A.E. (2008). When the evidence says, "Yes, no, and maybe so": Attending to and interpreting inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science, 17,* 47-51. doi:10.1111/j.1467-8721.2008.00546.x

De Los Reyes, A., Kundey, S. M., & Wang, M. (2011). The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology*

Review, *31*, 829-838. doi: 10.1016/j.cpr.2011.03.011

De Los Reyes, A., Lerner, M.D., Keeley, L.M., Weber, R., Drabick, D.A.G., Rabinowitz, J., & Goodman, K.L. (2019). Improving interpretability of subjective assessments about psychological phenomena: A review and cross-cultural meta-analysis. *Review of General Psychology*. Advance online publication. doi: 10.1177/1089268019837645

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology, 9*, 123-149. doi:10.1146/annurev-clinpsy-050212-185617

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J., ... & Polidori, G. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Jama, 291*, 2581-2590. doi:10.1001/jama.291.21.2581

Deros, D. E., Racz, S. J., Lipton, M. F., Augenstein, T. M., Karp, J. N., Keeley, L. M., . . . De Los Reyes (2018). Multi-informant assessments of adolescent social anxiety: Adding clarity by leveraging reports from unfamiliar peer confederates. *Behavior Therapy, 49*, 84-98. doi:10.1016/j.beth.2017.05.001

DiBartolo, P. M., Albano, A. M., Barlow, D. H., & Heimberg, R. G. (1998). Cross-informant agreement in the assessment of social phobia in youth. *Journal of Abnormal Child Psychology, 26,* 213-220. doi:10.1023/A:1022624318795

Doody, R. S., Stevens, J. C., Beck, C., Dubinsky, R. M., Kaye, J. A., Gwyther, L. M. S. W., ... & Cummings, J. L. (2001). Practice parameter: Management of dementia (an evidence-based review) Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, *56*, 1154-1166. doi:10.1212/WNL.56.9.1154

Grills, A. E., & Ollendick, T. H. (2003). Multiple informant agreement and the Anxiety Disorders Interview Schedule for parents and children. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 30-40. doi:10.1089/cap.2011.0114

Hawley, K. M., & Weisz, J. R. (2003). Child, parent, and therapist (dis)agreement on target problems in outpatient therapy: The therapist's dilemma and its implications. *Journal of Consulting and Clinical Psychology, 71,* 62-70. doi:10.1037//0022-006X.71.1.62

Holmbeck, G.N., Li, S.T., Schurman, J.V., Friedman, D., & Coakley, R.M. (2002). Collecting and managing multisource and multimethod data in studies of pediatric populations. *Journal of Pediatric Psychology, 27*, 5-18. doi:10.1093/jpepsy/27.1.5

Hunsley, J., & Lee, C.M. (2014). *Introduction to clinical psychology* (2nd ed.). Hoboken: Wiley & Sons.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29-51. doi:10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Mash, E.J. (Eds.). (2008). *A guide to assessments that work*. Oxford University Press. doi:10.1093/med:psych/9780195310641.003.0001

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, *75*, 887-900. doi:10.1037/0022-3514.75.4.887

Kazdin, A.E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy, 88*, 7-18. doi:10.1016/j.brat.2016.06.004

Kazdin, A.E., & Blase, S.L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on psychological science*, *6*, 21-37.

doi:10.1177/1745691610393527

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*, 593-602. doi:10.1001/archpsyc.62.6.593

Kramer, T.L., Phillips, S.D., Hargis, M.B., Miller, T.L., Burns, B.J., & Robbins, J.M. (2004). Disagreement between parent and adolescent reports of functional impairment. *Journal of Child Psychology and Psychiatry, 45,* 248-259. doi:10.1111/j.1469-7610.2004.00217.x

Laird, R. D., & De Los Reyes, A. (2013). Testing informant discrepancies as predictors of adolescent psychopathology: Why difference scores cannot tell you what you want to know and how polynomial regression may. *Journal of Abnormal Child Psychology*, *41*, 1–14. doi:10.1007/s10802-012-9659-y

Laird, R. D., & Weems, C. F. (2011). The equivalence of regression models using difference scores and models using separate scores for each informant: implications for the study of informant discrepancies. *Psychological Assessment*, *23*, 388–397. doi:10.1037/a0021926

Lerner, M.D., De Los Reyes, A., Drabick, D.A.G., Gerber, A.H., & Gadow, K.D. (2017). Informant discrepancies define discrete, clinically useful autism spectrum disorder subgroups. *Journal of Child Psychology and Psychiatry, 58*, 829-839. doi:10.1111/jcpp.12730

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53-70. doi: 10.1111/j.1745-6916.2007.00029.x

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood

psychopathology. *Journal of Clinical Child Psychology*, *19*(2), 136-143. doi:10.1207/s15374424jccp1902_5

Marsh, J. K., & Romano, A. L. (2016). Lay judgments of mental health treatment options: The mind vs. body problem. *Medical Decision Making: Policy & Practice, 1,* 1-12. doi:10.1177/2381468316669361

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*, 1–23. doi:10.3758/s13428-011-0124-6

McMahon, R. J., & Frick, P. J. (2005). Evidence-based assessment of conduct problems in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 477–505. doi: 10.1207/s15374424jccp3403_6

Ogles, B. M., Lambert, M. J., Weight, D. G., & Payne, I. R. (1990). Agoraphobia outcome measurement: A review and meta-analysis. *Psychological Assessment, 2,* 317-325. doi:10.1037/1040-3590.2.3.317

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023-1031. doi:10.3758/s13428-013-0434-y

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*, 33-60. doi:10.3389/fpsyg.2018.00213

Pelham, W. E., Jr., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 449–476. doi: 10.1207/s15374424jccp3403_5

Rosen, A., & Proctor, E. K. (2003). *Developing practice guidelines for social work intervention:*

Issues, methods, and research agenda. Columbia University Press.

Salsberg, E., Quigley, L., Mehfoud, N., Acquaviva, K., Wyche, K., & Sliwa, S. (2017). *Profile of the social work workforce*. Retrieved from https://www.cswe.org/Centers-Initiatives/Initiatives/National-Workforce-Initiative/SW-Workforce-Book-FINAL-11-08-2017.aspx.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*, 131-142. doi:10.1037/0033-2909.105.1.131

Southam-Gerow, M. A., & Prinstein, M. J. (2014). Evidence base updates: The evolution of the evaluation of psychological treatments for children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 43*, 1-6. doi:10.1080/15374416.2013.855128

Volkmar, F., Siegel, M., Woodbury-Smith, M., King, B., McCracken, J., & State, M. (2014). Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*, 237-257. doi:10.1016/j.jaac.2013.10.013

Weisz, J. R., Jensen Doss, A., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology, 56*, 337-363. doi:10.1146/annurev.psych.55.090902.141449

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55,* 542-549. doi:10.1037/0022-006X.55.4.542

Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2004). Effects of adolescent manic symptoms on agreement between youth, parent, and teacher ratings of behavior problems.

Journal of Affective Disorders, 82, S5–S16. doi:10.1016/j.jad.2004.05.016

Yzerbyt, V. Y., Schadron, G., Leyens, J. P., & Rocher, S. (1994). Social judgeability: The impact of meta-informational cues on the use of stereotypes. *Journal of Personality and Social Psychology*, *66*, 48-55. doi:10.1037/0022-3514.66.1.48

Imagine a child patient who is receiving therapy for ADHD. The patient's doctor has asked the child and the child's mother to rate how the child's ADHD symptoms have improved. The first bar shows how much the child rated the ADHD symptoms to have improved, and the second bar shows how much the child's mother rated the ADHD symptoms to have improved (on a scale from 0, no improvement at all, to 100, largest possible improvement).

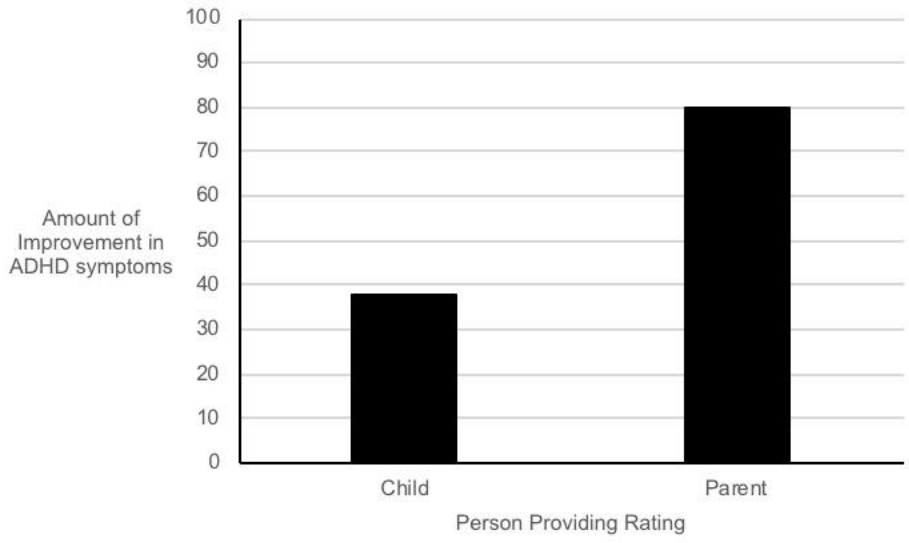
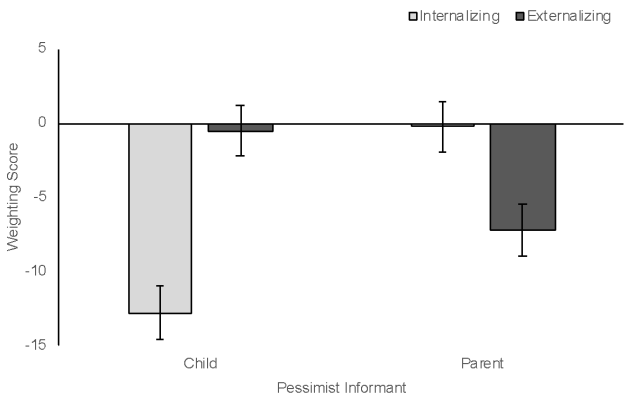
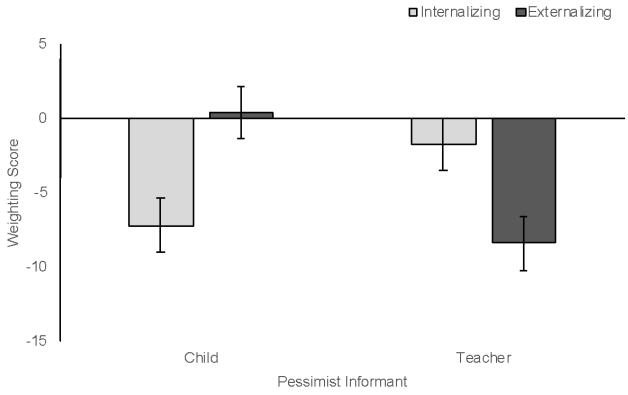


Figure 1. Example material shown to participants.

a)



b)



c)

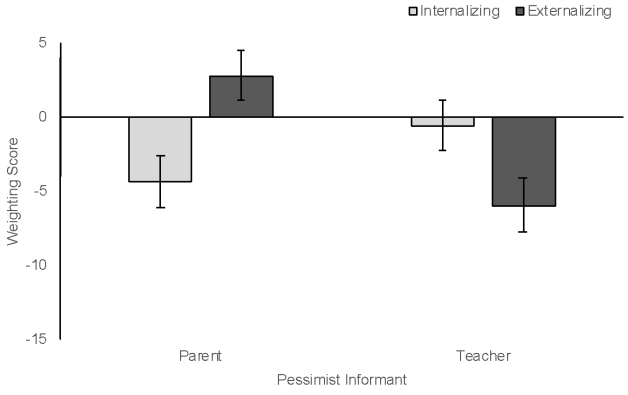


Figure 2 a-c. Lay weighting scores for Experiment 1 – 3 dyads. Error bars represent standard error. Graphs show ratings for a) child-parent dyads, b) child-teacher dyads, and c) parent-teacher dyads.

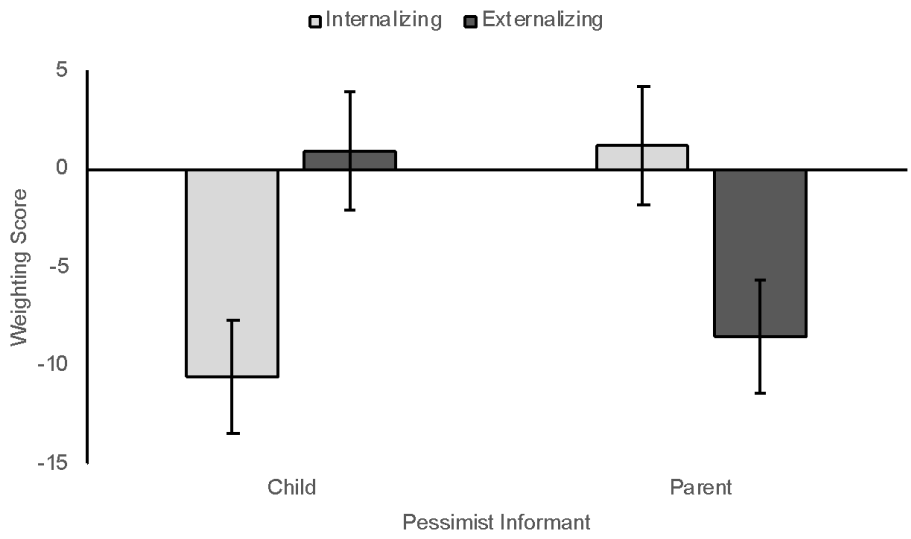


Figure 3. Experiment 4 weighting scores. Mental health professional ratings for child-parent dyads. Error bars represent standard error.