Accepted for publication in *Psychological Bulletin* on 29 May, 2019.

A meta-analysis of the emotional victim effect for female adult rape complainants: Does complainant distress influence credibility?

Faye T. Nitschke, Blake M. McKimmie and Eric J. Vanman

The University of Queensland

Author Note

Faye T. Nitschke, The School of Psychology, The University of Queensland; Blake M. McKimmie, The School of Psychology, The University of Queensland; Eric J. Vanman, The School of Psychology, The University of Queensland.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

The data and analysis code for this work are publicly available on the Open Science Framework at: https://osf.io/9gwmn/

Abstract

Rape cases have a disproportionately high attrition rate and low conviction rate compared to other criminal offenses. Evaluations of a rape complainant's credibility often determine whether a case progresses through the criminal justice system. Even though emotional demeanor is not related to witness honesty or accuracy, distressed rape complainants are perceived to be more credible than complainants who present with controlled affect. To understand the extent and robustness of the influence of emotional demeanor on credibility judgments of female adult rape complainants, we conducted a systematic review, meta-analysis and *p*-curve analysis of the experimental simulated decision-making literature on the influence of complainant emotional demeanor on complainant credibility. The meta-analysis included 20 studies with participants who were criminal justice professionals (e.g., police officers and judges), community members, and mock jurors ($N = 3128$). Results suggest that distressed demeanor significantly increased perceptions of complainant credibility, with a small to moderate effect size estimate. Importantly, the results of *p*-curve analysis suggest that reporting bias is not a likely explanation for the effect of emotional demeanor on rape complainant credibility. Sample type (whether perceivers were criminal justice professionals or prospective jurors) and stimulus modality (whether perceivers read about or watched the complainant recount the alleged rape) were not found to moderate the effect size estimate. These results suggest that effective methods of reducing reliance on emotional demeanor to make credibility judgments about rape complainants should be investigated to make credibility assessments fairer and more accurate.

Keywords: rape, credibility, emotional victim effect, jury, meta-analysis

Public significance statement: This meta-analysis indicates that rape complainants with distressed emotional demeanor are perceived as more credible than their emotionally controlled counterparts. As emotional demeanor is not a reliable indicator of honesty, and poor complainant credibility is associated with case attrition in the criminal justice system, preventing complainant distress from influencing credibility judgments should be a research and policy priority.

A meta-analysis of the emotional victim effect for female adult rape complainants: Does distress influence credibility?

On average, 9% of rape allegations made to police in the United States, Europe, and Australia proceed to trial (Alderden & Ullman, 2012; Daly & Balhours, 2010; Jehle, 2012). In up to 88% of rape cases, the defendant and complainant know each other (acquaintance rape; Australian Bureau of Statistics [ABS], 2017; Flatley, 2018; Smith et al., 2017) and the complainant's testimony about consent is critical (Quadara, Fileborn & Parkinson, 2013). If the complainant is not perceived to be credible, these cases do not progress through the criminal justice system (Brown, Hamilton & O'Neill, 2007; O'Neal, 2017). Distressed complainants are perceived to be more credible than complainants who show other emotions (called the emotional victim effect or EVE; Ask & Landström, 2010). In this review and meta-analysis of the emotional victim effect, we focus on judgments about female adult rape complainants' credibility, as the majority of reported rape offenses are committed by men against women (e.g., Hohl & Stanko, 2015).

Complainant Credibility and Case Progression

For a rape case to proceed to trial, there are several phases in the criminal justice system the case must pass through. Attrition occurs at all stages (e.g., Cox, 2015; Jehle, 2012). First, the complainant must report the alleged offense to police. The police must investigate the case and choose to charge the alleged offender. The prosecutor's office then must decide to proceed with the case. Finally, a jury and judge hear the trial and make a judgment about the defendant's guilt. If the complainant is deemed credible, the case is more likely to proceed at any of these steps. Complainants who think they will not be believed do not report to authorities (e.g., Filipas & Ullman, 2001; Stern Review, 2010). Police officers are more likely to recommend investigating and prosecuting cases (Ask, 2010; Alderden & Ullman, 2012; Brown et al., 2007; Kerstetter, 1990; Morabito, Pattavina & Williams, 2016;

Tasca, Rodriguez, Spohn & Koss, 2013), prosecutors are more likely to proceed to trial (Frohmann, 1991; 1997; Lievore, 2005; Spohn & Tellis, 2012), and jurors are more likely to convict the defendant (e.g., Ellison & Munro, 2009) when the complainant is deemed credible. This means it is critical that credibility judgments are made without prejudice.

Judging Complainant Credibility

Perceivers, whether they be police officers, judges or jurors, must form an impression of a rape complainant to judge credibility. The Heuristic-Systematic Model (HSM; Chaiken, 1980; Chaiken & Ledgerwood, 2012; Chaiken, Liberman & Eagly, 1989) is a dual process explanation of persuasion that is often used to describe the processing of social information (e.g., Reinhard & Sporer, 2010; Todorov, Chaiken & Henderson, 2002). The model suggests that observers may use two information processing styles to judge the credibility of a rape complainant. If observers use systematic processing, then a careful consideration of the available information is undertaken to determine the relevant information to judge credibility (Chaiken & Maheswaran, 1994). Perceivers must be more motivated and cognitively capable to engage in systematic processing. In heuristic processing, a pre-existing cognitive structure (e.g., stereotype or schema) is used to guide information processing. Heuristic processing is less effortful, and often used by less motivated perceivers, but still requires some cognitive capacity (Chen & Chaiken, 1999).

Unlike other dual process persuasion models (for example, the Elaboration Likelihood Model; Petty & Cacioppo, 1986) which assume that perceivers use one style of information processing to make a judgment, the HSM proposes that perceivers can use both heuristic and systematic processing to evaluate information. The level of information processing used by the perceiver to make a judgment depends on the level of confidence the perceiver needs to make this judgment (called the sufficiency threshold). Perceivers will continue to engage in information processing until they reach their sufficiency threshold—if heuristic processing does not produce a judgment the perceiver is confident in then systematic processing will be used until a judgment the perceiver is confident in is made. In rape cases, perceivers are either criminal justice professionals who are motivated to perform their role competently and to assist the community (e.g., Raganella & White, 2004; White, Cooper, Saunders, & Raganella, 2010), or community members acting as jurors who aim to make an accurate decision (e.g., Thomas, 2010).

For motivated perceivers, two predicted ways that heuristic and systematic processing can interact to influence the credibility judgment made are the attenuation and bias hypotheses (Chaiken & Maheswaran, 1994). Under the attenuation hypothesis, if the judgment arising from initial heuristic processing is at odds with the judgment arising from subsequent systematic processing, then the decision made through systematic processing will override the conclusion from heuristic processing (Maheswaran, Mackie & Chaiken, 1992). If attenuation occurs, systematic processing is curative against the biasing effect of heuristics in judgments made by highly motivated perceivers. In contrast, under the bias hypothesis, when the evidence on which the judgment must be made is ambiguous, then the Heuristic-Systematic Model suggests that heuristics can influence subsequent information processing even for highly motivated perceivers (called biased systematic information processing). If bias occurs, then systematic processing does not overcome the influence of heuristics on judgments made by highly motivated perceivers (Chen & Chaiken, 1999). When bias occurs, the influence of the heuristic can only be reduced if the heuristic is challenged by other evidence considered in systematic processing.

Studies of credibility judgments made of people in everyday social situations (i.e., when subletting an apartment or cancelling a date) show support for the operation of the attenuation hypothesis for motivated perceivers (e.g., Reinhard, 2010; Reinhard & Sporer, 2008; 2010). Highly motivated perceivers rely on the content of messages provided by the person rather than characteristics about the person to judge that individual's credibility compared to perceivers with low motivation (Reinhard & Sporer, 2008; Reinhard & Sporer, 2010). Perceivers who engage in systematic processing provide more reasons for their credibility judgment which focus on the content of the person's statement rather than heuristic cues about the person (Reinhard & Sporer, 2008). Together, this suggests that subsequent systematic processing, in which the content of statements made by the person are considered carefully, is overcoming the effect of heuristic cues about the person in credibility judgments made about the person (i.e., the attenuation hypothesis). There is some evidence that the attenuation hypothesis may explain information processing for perceivers in rape cases. Ask and Landström (2010) found that a heuristic cue about the complainant's credibility (i.e., emotional demeanor) influenced judgments about credibility when perceivers were under cognitive load (and likely to engage in heuristic processing), whereas perceivers not under cognitive load (and likely to use systematic processing) were not influenced by the heuristic cue in their credibility judgments.

However, unlike the scenarios used in studies of everyday credibility judgments, rape cases are frequently highly ambiguous, as they commonly involve the complainant's testimony without strong corroborative evidence (e.g., conclusive medical evidence or third-party eye-witness testimony; Quadara et al., 2013). This message ambiguity means that, from a Heuristic-Systematic Model perspective, the bias hypothesis likely explains how motivated perceivers process information to make credibility judgments about rape complainants. If biased information processing occurs for perceivers to make credibility judgments about rape complainants, heuristic cues about the complainant will shape systematic processing of evidence to judge complainant credibility, regardless of the perceiver's motivation (Chaiken & Ledgerwood, 2012; Todorov et al., 2002). Chaiken and Maheswaran (1994) found that highly motivated perceivers who were exposed to an ambiguous message engaged in systematic processing that was biased by the heuristic cue about the person delivering the message. When perceivers viewed a positive heuristic cue about the person delivering the message, subsequent systematic processing focused on the positive persuasive arguments made in the message (and vice versa). Although Ask and Landström (2010) found some evidence in support of the attenuation hypotheses in judgments made in rape cases, they did not measure case ambiguity. It is therefore possible that the rape case presented in that study was not viewed as ambiguous by perceivers and thus the conditions to trigger biased systematic information processing were not present. The bias hypothesis suggests that heuristic cues about rape complainants could have an impact on credibility judgments, regardless of perceiver motivation.

Heuristics About Rape Complainants

There are several aspects of complainant behavior that influence the judgments perceivers make about rape complainants and may operate as heuristic cues within information processing. Endorsement of mistaken beliefs about rape events (called rape myths; Hockett, Smith, Klausing & Saucier, 2016; Lonsway & Fitzgerald, 1994), the extent to which the assault events match with schemas for rape (McKimmie, Masser, Nitschke, Schuller & Goodman-Delahunty, 2018a; Smith, 1991; 1993), and the extent to which the victim conforms with stereotypes about rape complainants (e.g., Angelone, Mitchell & Grossi, 2015; Schuller & Hastings, 2002) all affect how rape complainants are judged. Perceiver attributes, like gender, also influence evaluations of rape complainants.

Men and women tend to evaluate rape complainants differently (Anderson, Cooper & Okamura, 1997). In general, men tend to have more permissive attitudes towards rape and evaluate rape complainants more negatively than do women (e.g., Kleinke & Meyer, 1990; Newcombe, van der Eynde, Hafner & Jolly, 2008). Two meta-analytic reviews of rape myth endorsement suggest that men have greater endorsement of rape myths than women

(Anderson et al., 1997; Suarez & Gadalla, 2010). In addition, several narrative reviews suggest that men perceive rape complainants to be more to blame for their assault than women (e.g., Grubb & Harrower, 2008; Grubb & Turner, 2012; van der Bruggen & Grubb, 2014). In contrast, a recent review of acquaintance rape literature suggests that the effect of perceiver gender on judgments of victim blame is mixed (Gravelin, Biernat & Bucher, 2019). However, as perceiver gender is a natural confound, there is no way to tell if the differences observed are due to perceiver gender or differences between men and women in attitudes or other cognitive structures (Kite & Whitley, 2018). For example, there is evidence that attitudes towards sexual behavior in romantic relationships mediates the relationship between perceiver gender and judgments that the complainant of an acquaintance sexual assault was to blame for the assault (Lynch, Jewell, Wasarhaley, Golding, & Renzetti, 2017).

Perceivers might rely on several heuristic cues to assess the complainant's credibility. These include rape myths, rape schemas or scripts, and the rape victim stereotype. Rape myths are factually inaccurate and include beliefs like "false accusations of rape are common" or "women say no to sex when they really mean yes" (i.e., women exhibit 'token' resistance). As perceivers' endorsement of rape myths increases, rape victims are evaluated more negatively (Süssenbach, Bohner & Eyssel, 2012; Süssenbach, Eyssel & Bohner, 2013). The more an alleged assault conforms to rape myths, the more likely it is perceived as a legitimate assault and the victim is evaluated positively (Süssenbach, Albrecht & Bohner, 2017; Süssenbach, Eyssel, Rees & Bohner, 2017). Perceivers who endorse rape myths also tend to see complainants as less credible (Bohner & Schapansky, 2018; Nitschke, Masser, McKimmie & Riachi, 2018). Some authors suggest that this means that rape myths are a schema used to make decisions about the legitimacy of a rape allegation (Eyssel & Bohner, 2011; Krahé, 2016).

Perceivers’ expectations for how rape typically occurs also form schemas that are used to evaluate complainants (e.g., Carroll & Clark, 2006; Littleton & Axsom, 2003, Littleton, Tabernik, Canales & Backstrom, 2009). Perceivers typically expect the complainant and perpetrator to be strangers and for the perpetrator to use force to subdue the complainant (e.g., Littleton & Axsom, 2003). Complainants who are assaulted by a stranger tend to be evaluated more positively by perceivers (e.g., Abrams, Viki, Masser & Bohner, 2003; Bieneck & Krahé, 2011; Bridges & McGrail, 1989; Krahé, Temkin & Bieneck, 2007).

Perceivers also use the rape victim stereotype to judge the complainant. Perceivers typically expect complainants to vigorously physically and verbally resist the perpetrator, be sober, and report immediately to authorities in a distressed state (Carroll & Clark, 2006; Littleton & Axsom, 2003). When complainants do not behave as expected, they tend to be blamed for the assault (e.g., Davies, Rogers & Whitelegg, 2009; Masser, Lee & McKimmie, 2010) and judged less credible (e.g., Angelone, Mitchell & Grossi, 2015, McKimmie, Masser & Bongiorno, 2014; Schuller & Hastings, 2002).

The complainant’s emotional demeanor strongly influences credibility judgments (Kaufmann, Drevland, Wessel, Overskeid & Magnussen, 2003). Perceivers tend to expect rape complainants to experience negative emotions that are much stronger than those experienced by other victims of crime (Wrede & Ask, 2015). Emotional demeanor also influences perceptions of the complainant’s typicality as a rape victim (Schuller, McKimmie, Masser & Klippenstine, 2010), perhaps indicating demeanor is part of the rape victim stereotype or a schema for rape events. Regardless of whether complainant emotional demeanor is a unique heuristic cue, or part of a rape victim stereotype or rape schema, the bias hypothesis (Chaiken, 1980) suggests that perceivers may use complainant emotional demeanor to form an impression of the complainant and judge credibility.

Distressed Demeanor and Credibility

A typical study investigating the effect of complainant emotional demeanor on credibility employs an experimental design and a simulated decision-making paradigm. In this paradigm, participants are presented with complainant evidence about an alleged rape event. The complainant (an actress) portrays an emotional state, most frequently distress or controlled affect (e.g., Kaufmann et al., 2003; Klippenstine & Schuller, 2012) via a video or written evidence synopsis. Participants make credibility judgments using a questionnaire (e.g., Ask & Landström, 2010; Hackett, Day & Mohr, 2008). Studies have reported that complainants who appear distressed (i.e., upset, crying) are perceived to be more credible than their emotionally ‘flat’ or ‘controlled’ counterparts (Schuller et al., 2010). Mock jurors’ (Peace & Valois, 2014), community members’ (Calhoun, Cann, Selby & Magee, 1981) and police officers’ (Ask & Landström, 2010; Baldry, 1996) judgments of complainant credibility all show this effect.

Several moderators, including perceiver gender and the rape victim stereotype, have been investigated in studies on the effect of complainant emotional demeanor (e.g., Peace & Valois, 2014; Schuller et al., 2010). Most studies suggest that perceiver gender has no unique effect on complainant credibility judgments and does not interact with the effect of complainant emotional demeanor on credibility (Bollingmo et al., 2007; 2009; Calhoun et al., 1981; Kaufmann et al., 2003; Klippenstine, 2010; Schuller et al., 2010; Wessel et al., 2006). One study (Bohner & Schapansky, 2018) found perceiver gender interacted with complainant emotional demeanor to influence credibility judgments. Women high in rape myth endorsement judged distressed complainants as more credible. Women low in rape myth endorsement found the complainant credible irrespective of her emotional demeanor. Men showed no difference in credibility judgments as a function of rape myth endorsement or complainant emotional demeanor. Collectively, perceiver gender does not seem to influence the impact of complainant emotional demeanor on credibility.

Stereotypes about rape victims have also been investigated in studies of complainant emotional demeanor. Kaufmann et al. (2003) found no effect of strong or weak testimony (manipulated by presence or absence of complainant resistance and perpetrator force during the assault) on judgments of complainant credibility. In contrast, Schuller et al. (2010) found that when the complainant physically resisted the perpetrator, she was evaluated as more credible than when she only verbally resisted. The complainant's stereotypicality did not interact with the effect of complainant emotional demeanor on credibility in either study. However, Hackett et al. (2008) found perceivers who expected rape complainants to become distressed evaluated the distressed complainant as more credible than the neutral complainant. Perceivers who had no expectations about the emotions that complainants should show were not influenced by the complainant's emotional state. Complainant emotional demeanor may operate as one of several heuristic cues available for judging credibility. But should perceivers use distress as a cue to judge complainant credibility accurately?

Emotional Demeanor as a Credibility Cue

Credibility judgments should be about whether the witness is honest and reliable (e.g., in Australia and Canada; *Reymond v Township of Bosanquet*, 1919; *White v R*, 1947). However, emotions (including distress) do not accurately indicate that witnesses are telling the truth (Bond & DePaulo, 2006; DePaulo et al., 2003). Perceivers are no more accurate at detecting deception when witnesses display strong emotion compared to when no strong emotion is shown (Hartwig & Bond, 2014). Rape complainants are also equally likely to appear distressed or with controlled affect (Burgess & Carretta, 2016; Burgess & Holmstrom, 1974; Caretta & Burgess, 2013). Complainants experience post-traumatic stress disorder

(PTSD) and depression at high rates (Norris & Krysztof, 1994; Snipes, Calton, Green, Perrin & Benotsch, 2017; Ullman & Filipas, 2001) and often suppress emotion to manage their trauma (Burgess & Holmstrom, 1986; Maddox, Lee & Barker, 2011; Maddox, Lee & Barker, 2012). This means many rape complainants may be unfairly judged as not credible due to their emotional demeanor when giving evidence.

Given that complainant emotional demeanor can adversely affect credibility judgments by criminal justice professionals and jurors, it is critical that we understand the extent and robustness of this effect. We undertook a meta-analytic review of the simulated decision-making literature to investigate this. In our review, we focused on studies comparing distressed and neutral expressions of emotional demeanor by complainants, as these two conditions were used in the majority of EVE studies and these emotions are commonly experienced by rape complainants while giving evidence (e.g., Konradi, 1999; 2007). We explored two potential moderators that may influence the extent to which complainant emotional demeanor influences credibility judgments – stimulus modality and sample type – which also assess the ecological and external validity of the effect. We used *p*-curve analysis to determine whether the significant effects in this literature can be explained by selective reporting (e.g., publication bias or *p*-hacking).

Evidence Stimulus Format

Video and written complainant evidence have been used to investigate the effect of complainant emotional demeanor on credibility judgments (e.g., Kaufmann et al., 2003; Peace & Valois, 2014). If perceivers undertake biased systematic information processing to judge complainant credibility, then motivation to engage in information processing will not substantially influence whether complainant emotional demeanor is used to judge credibility (Chen & Chaiken, 1999). However, stimuli modality may still influence the extent to which complainant emotion influences credibility judgments by increasing the salience and accessibility of complainant emotional demeanor as a heuristic cue. Perceivers evaluate accessible heuristic cues to be more reliable, which increases how relevant the heuristic cue appears to make the judgment (Chen & Chaiken, 1999). If a heuristic cue like complainant emotional demeanor is viewed as more relevant to judge credibility, then it is unlikely to be challenged in biased systematic processing and will have an effect on credibility judgments. This means complainant emotional demeanor may have a larger effect on credibility judgments when the modality of complainant evidence makes complainant emotional demeanor more salient. Chaiken and Eagly (1983) argued that characteristics of the speaker are more salient in video messages, so it is possible that complainant emotional demeanor has a larger impact on credibility judgments when video complainant evidence is presented.

Perceivers typically struggle to differentiate negatively valenced emotions, but video stimulus improves emotion recognition. Sadness is accurately identified around half the time (Gitter, Kozel & Mostofsky, 1972) and has moderate cross-cultural recognition (Elfenbein & Ambady, 2002). Police trainees rated distressed rape complainants as experiencing significantly more sadness but also more anger, fear, and disgust (Ask & Landström, 2010). Sadness is most accurately recognized when video of the target is provided (Gitter et al., 1972). Given that angry complainants are sometimes viewed as less credible than distressed complainants (e.g., Bohner & Schapansky, 2018 cf. Vrij & Fischer, 1997), it is possible that the influence of distress on credibility judgments would be larger when distress is readily recognized (i.e., in studies using video rather than written complainant evidence).

However, complainant emotional demeanor may be more salient in written complainant evidence. In written and spoken communication, perceivers expect others to provide as much information as needed to convey their point and for information to be truthful and relevant (conversational norms or maxims of quantity, quality, and relation; Grice, 1975). Studies on person perception suggest that perceivers may infer information provided about a target is relevant to form an impression of the target, not because of pre-existing heuristics, but because relevance is inferred from conversational norms (Kahneman & Tversky, 1973; Schwarz, Strack, Hilton & Naderer, 1991). It is possible complainant emotional demeanor is used to make credibility judgments because the presence of the information in the written experimental synopsis implies, under conversation norms, that this information is relevant and reliable (Bless, Strack, & Schwarz, 1993). Obvious experimental manipulations can make participants more susceptible to demand characteristics – where participants guess at the study hypotheses and respond accordingly (Goodwin & Goodwin, 2013; Haslam & McGarty, 2008; Kite & Whitley, 2018). If the relevance of complainant emotional demeanor to judge credibility is increased in written complainant evidence, and perceivers use biased systematic information processing, the effect of complainant emotional demeanor on credibility could be larger than when video complainant evidence is presented.

However, perceivers who read complainant evidence have the opportunity to review the evidence as needed, which may facilitate critical systematic information processing and lessen the effect of complainant emotional demeanor as a heuristic cue by providing perceivers an opportunity to challenge emotion as a heuristic cue in systematic processing. This may lessen the effect of complainant emotional demeanor on credibility judgments when biased systematic information processing takes place. Perceivers have higher comprehension of written messages compared to video messages (Chaiken & Eagly, 1976) and also focus more on message content in their information processing (Chaiken & Eagly, 1983). Perceivers who review notes taken during video messages rely less on heuristic cues to make judgments than perceivers who did not review or take notes during the video message (Strub & McKimmie, 2012).

Given the variety of explanations about the effect of stimulus modality on the salience of complainant emotional demeanor, it is unclear whether stimulus modality influences the emotional victim effect. To our knowledge, no experimental study has directly examined this. In rape cases, police officers, lawyers, judges, and jurors experience the complainant's emotional demeanor in person during the investigation and trial proceedings (Maddox et al., 2012; Temkin & Krahé, 2008) over hours or even days (e.g., Konradi, 2007). This means written complainant evidence used in experiments lacks ecological validity (Kerr, 2017), so examining the effect of stimuli modality also permits an investigation of whether more ecologically valid experimental stimuli (video complainant evidence) influences the emotional victim effect. We therefore investigated the effect of stimulus modality, whether observers read about or saw the complainant, as a moderator in this meta-analysis.

Sample Type

Another potential moderator of the emotional victim effect is the type of perceiver making the credibility judgment about the complainant. Criminal justice professionals (e.g., judges, police officers) and trainees (e.g., police trainees, law students) and community members have participated in studies exploring the effect of complainant emotional demeanor on credibility judgments (Hackett et al., 2008; Wessel et al., 2006). Decisions about complainant credibility made by police officers and prosecutors are associated with rape case attrition (e.g., Spohn & Tellis, 2012; Tasca et al., 2013), so understanding whether criminal justice professionals are influenced by complainant emotional demeanor in these decisions is critical to understanding how emotional demeanor may be contributing to rape case attrition in the criminal justice system. As such, many researchers have called for further investigation of how criminal justice professionals may differ to lay perceivers in their decision-making about rape complainant credibility (e.g., Ask, 2010; Wessel et al., 2006).

Some criminal justice professionals and scholars assume that experience and training which judges, lawyers, and police officers receive or training which police trainees or law students receive equips them to make more accurate decisions in cases than jurors or community members (e.g., Kahan, 2015 cf. Lidén, Gräns & Juslin, 2018). Judges often write about how they believe jurors are easily influenced by bias and their emotions (e.g., Edwards, 1984; Hans & Vidmar, 1986) and prosecutors often raise jurors' beliefs in victim stereotypes as an impediment to prosecution in rape cases but not necessarily an issue in their own decision-making (Bluett-Boyd & Fileborn, 2014; Temkin, 2000; Temkin & Krahé, 2008). Similarly, police officers raise jurors' doubts about the complainant's credibility as barriers to convictions in rape trials but maintain that many aspects of the victim stereotype do not influence their own decision-making (e.g., Venema, 2013).

If perceivers engage in biased systematic information processing to decide complainant credibility, then professional training or experience may have a limited impact on whether complainant emotion influences credibility judgments. Factors that may modify criminal justice professionals' motivation to engage in information processing relative to lay perceivers, like feeling more accountable for decisions made (e.g., Lerner & Tetlock, 1999; Ashton, 1992), will have a limited effect on how complainant emotional demeanor influences information processing to judge credibility. However, training or experience may encourage criminal justice professionals to challenge the heuristic cue of complainant emotional demeanor with other information during information processing, which would lessen the effect of complainant demeanor on judgments of complainant credibility through biased systematic information processing.

Knowledge and expertise have been identified as cognitive resource factors which modify the nature of information processing within the Heuristic-Systematic Model (Todorov et al., 2002). Research suggests that familiarity (or knowledge) of the decision context leads perceivers to focus on more complex information presented in the content of messages to judge the credibility of people in every day social situations (Reinhard, Scharmach, & Sporer,

2012) and to more accurately discriminate between deceptive and honest accounts of an event (Reinhard, Sporer, Scharmach, & Marksteiner, 2011).

Expert decision-makers are able to integrate complex and conflicting information, and attend to relevant information to make judgments more ably than lay perceivers (e.g., Bédard & Chi, 1992; Kahneman & Klein, 2009; Nee & Ward, 2015; Shanteau, 1992), which may allow them to challenge heuristic cues that are influencing systematic processing. Some evidence suggests that criminal justice professionals are more open to dealing with conflicting or complex information. For example, experienced police officers create a greater number of alternative hypotheses and engage in a wider range of investigative actions than novice police officers (Fahsing & Ask, 2016). However, other studies have found that experienced police officers make fewer alternative hypotheses than lay perceivers in simulated criminal cases (Ask & Granhag, 2005). Specialized training programs have been found to reduce the extent to which the victim stereotype influences rape case judgments by police officers (Ask, 2010; Darwinkel, Power & Tidmarsh, 2013). It is possible that professional training or experience may lessen the effect that complainant emotional demeanor has on credibility judgments.

However, if a heuristic cue is seen as relevant to deciding complainant credibility then it is unlikely it will be challenged within biased systematic information processing. Aspects of the victim stereotype influence criminal justice professionals' judgments of rape complainants. Advanced law students and probationary lawyers evaluate rape complainants who do not match the rape victim stereotype more negatively (e.g., Krahé, Temkin, Bieneck & Berger, 2008; Temkin & Krahé, 2008). Barristers and judges are influenced by the rape victim stereotype and rape myths in their court practice (e.g., Feldman-Summers & Palmer, 1980; Gray & Horvath, 2018; Sleath & Bull, 2015; Temkin, 2000; Temkin, Gray & Barrett, 2018). Time spent as a police officer, or investigating sexual assault, does not consistently improve judgments of rape complainants including credibility (Goodman-Delahunty & Graham, 2011; Sleath & Bull, 2012; Wentz & Archbold, 2012 cf. Campbell, 1995; Page, 2007). Critically, nearly three-quarters of prosecutors and police officers surveyed agreed (in part or more strongly) that a rape complainant's emotional demeanor was a valid indicator of whether the complainant was being honest (Ask, 2010). If criminal justice professionals see complainant emotional demeanor as relevant to judge credibility, it is likely it could be relied on and bias systematic information processing. This means it is possible that professional training or expertise may not lessen the effect that complainant emotional demeanor has on credibility judgments.

Given that it is unclear what the effect of professional training or expertise may be on the use of complainant emotional demeanor in judging credibility, we explored sample type as a moderator of the effect of complainant emotion on credibility judgments. The inclusion of sample type as a moderator also allowed us to examine how well the emotional victim effect generalizes across samples (Krauss & Lieberman, 2017). This addresses an important practical question about whether complainant emotional demeanor, by biasing complainant credibility judgments, is one explanation for rape case attrition across the criminal justice system.

Validity and Reliability of Meta-Analytic Estimates

In a meta-analysis, it is critical to consider whether study validity may cause an effect to be over- or underestimated (Higgins & Altman, 2008; Lakens, Hilgard & Staaks, 2016). By examining whether the stimulus format moderates the effect-size estimate, we can also assess how accurately the experimental paradigm captures the real-life context of the effect of complainant emotional demeanor on credibility (sometimes referred to as ecological validity; Kerr, 2017). Similarly, we can examine a specific aspect of the external validity — whether the results will generalize to other populations — by exploring sample type as a moderator (Krauss & Lieberman, 2017).

It is also important to carefully scrutinize internal validity of the studies included in a meta-analysis. An internally valid study uses an adequate method to answer the research question proposed (Higgins & Altman, 2008), but biased experimental methodology can threaten this (Haslam & McGarty, 2008). For example, if study participants systematically differ across conditions (selection bias) this may be a viable alternative explanation for the results (Higgins & Altman, 2008; Kite & Whitley, 2018). Performance bias (whether the independent variable was effectively manipulated), detection bias (whether the dependent variable was appropriately sensitive and reliable), and reporting bias (whether the results were selectively reported for publication) also threaten study internal validity (Bornstein, 2017; Kite & Whitley, 2018; Reeves, Deeks, Higgins & Wells, 2008).

The effect of reporting and publication bias on the effect size estimate is of particular concern in meta-analyses, as such analyses rely substantially on published literature to estimate the overall effect size. Even when meta-analysts make efforts to access unpublished literature, the total number of unpublished studies available to include in the analysis can remain small (e.g., Bornstein et al., 2017; Paluck, Green & Green, 2017). The published literature tends to over-represent statistically significant findings (e.g., Rotton, Foos, Van Meek & Levitt, 1995), which in turn can inflate the effect size estimates calculated through meta-analysis (Kotiaho & Tomkins, 2002).

Some common research practices (John, Lowenstein & Prelec, 2012), particularly flexibility in data analysis strategies (*p*-hacking or unconscious analytic tuning; Chambers, 2017; Simmons, Nelson & Simonsohn, 2011) and hypothesizing after results are known (HARKing; Chambers, 2017; van Assen, van Aert, Nuijten & Wicherts, 2014), might be inflating statistical error rates in the published literature. There may be substantial numbers of false positives in the published literature in psychology and other scientific disciplines (Simonsohn, Nelson & Simmons, 2014a; Simonsohn, Nelson & Simmons, 2014b). A meta-analytic effect size estimate based on literature which reports inflated or false positive effects will yield an artificially inflated effect size estimate (Lakens et al., 2016).

To assess whether publication bias or problematic research practices might influence the findings of this meta-analysis, we conducted a *p*-curve analysis of the included studies. *P*-curve analysis is a diagnostic test of the evidential value of a set of studies (Simonsohn et al., 2014a) that suggests whether selective reporting (via publication bias or *p*-hacking) is a viable explanation for statistically significant results. *P*-curve analysis more comprehensively assesses reporting bias than other publication bias assessment methods for meta-analysis.

Aims and Hypotheses

To our knowledge, this is the first meta-analytic review of the effect of female adult rape complainant emotional demeanor on credibility judgments. Our first aim for this meta-analysis, *p*-curve analysis, and review was to investigate the direction, size, and robustness of the effect. Our second aim was to investigate stimulus modality and sample type as moderators of this effect. We had three hypotheses:

> *Hypothesis 1 (H1):* We expected there would be a significant and positive aggregate effect size for the emotional victim effect, such that a distressed female adult complainant of rape would be perceived as more credible than her emotionally controlled counterpart.

We hypothesized this based on the Heuristic-Systematic Model (Chaiken, 1980), which suggests that when evidence is ambiguous possible heuristic cues, like complainant emotional demeanor, influence systematic information processing regardless of the perceiver's motivation to engage in information processing (i.e., the bias hypothesis or biased systematic information processing; Chaiken & Maheswaran, 1994; Chen & Chaiken, 1999).

Hypothesis 2 (H2): We expected the effect of complainant emotional demeanor on credibility judgments to be greater when perceivers had viewed video complainant evidence compared to when they had read an evidence synopsis.

We hypothesized this on evidence which suggests that visual displays of emotion are more salient (Gitter et al., 1972), which may make complainant emotional demeanor more relevant as a heuristic cue to judge complainant credibility (Chen & Chaiken, 1999). If complainant emotional demeanor is seen as relevant to judging credibility, then it may have larger effect within biased systematic information processing and on credibility judgments (Chaiken & Maheswaran, 1994).

Hypothesis 3 (H3): We expected the effect of complainant emotional demeanor on credibility judgments would be less when perceivers had relevant training or experience in the criminal justice system (i.e. police officers, police trainees, law students, judges) compared to when perceivers had no relevant training or experience judging credibility (i.e. mock jurors and community members).

We hypothesized this based on research that suggests that familiarity with the decision context, which professional training or experience may provide, facilitates engaging with more complex and conflicting information within systematic information processing (Reinhard et al., 2011; 2012). Challenging the relevance of complainant emotional demeanor to judge credibility within systematic processing may lessen the effect of complainant emotion within biased systematic information processing.

Unregistered and Exploratory Analyses

While executing this review, we decided to conduct some additional tests of publication bias, a study quality assessment, and exploratory content analyses to examine the internal validity of included studies in the meta-analysis.

Additional analyses of publication bias. All methods used to detect publication bias have limitations (e.g., Bruns & Ioannidis, 2016; Stanley, 2017). In addition to the *p*-curve analysis, which was pre-registered as part of our analytic strategy, we decided to conduct some additional tests of publication bias. This allowed us to assess the cumulative evidence of the impact of publication bias in our review across multiple analytic methods.

Study quality assessment. To assess the internal validity of included studies we executed an adapted form of the Cochrane Risk of Bias assessment (Higgins et al., 2011; Higgins & Altman, 2008). The Cochrane tool was developed to assess internal validity of randomized controlled trials, so we adapted it for cross-sectional psychology studies. This assessment enabled us to comprehensively evaluate the internal validity of studies included in the review.

Content analysis. Content analysis is a systematic approach to categorizing qualitative data (Krippendorff, 2013; Weber, 1980). In the content analyses, we investigated consistency and face validity in manipulations of complainant emotional demeanor and credibility measures in the included studies. This permitted us to make comprehensive suggestions for improvements to simulation methodology based on a systematic evaluation of the methods reported within the literature.

Method

This meta-analysis, including the hypotheses, search protocol, and analysis plan were pre-registered at the Open Science Framework: https://osf.io/9x58r

Eligibility Criteria

To be included in the review, studies met all of the following inclusion criteria:

(1) Reported in English

(2) Empirical (e.g., no opinion pieces or legal commentary)

(3) Sampled adults (18 years of age and older)

(4) Manipulated the emotional demeanor of a female adult complainant of sexual assault or rape (sexual intercourse without consent)

(5) Measured perceptions of a female adult complainant of rape (i.e., perceived credibility of the witness)

Search Strategy

We executed a scoping review to determine whether any reviews or meta-analyses were registered on this topic. A search of the Centre for Reviews and Dissemination (DARE), the Campbell Collaboration, and the International Prospective Register of Systematic Reviews (PROSPERO) was conducted using the key words victim, rape, sexual assault, and jury. We found no registrations that addressed the aims and hypotheses of this review.

To identify eligible articles, systematic searches of the PsycINFO, Scopus, Web of Science, and ProQuest Social Sciences databases were undertaken. To improve the reliability of the effect size estimate, unpublished studies were included in this meta-analysis (Lefebvre, Manheimer & Glanville, 2008). To identify unpublished literature, electronic searches of the ProQuest Dissertations and Theses and Open Grey databases were completed. All databases were searched using the following keywords in document title, abstract, and subject or index terms: rape OR sexual assault AND credibility OR emotion. The database searches included studies published and indexed before 8 November 2017.

We engaged in additional search techniques to ensure we identified as many eligible reports as possible. We manually searched the reference lists of eligible articles and conducted ‘cited by’ searches in Web of Science and Google Scholar to find additional eligible articles. We conducted author searches using Google Scholar to search for further eligible publications produced by authors of eligible articles. To access unpublished data, we circulated requests for unpublished data to relevant psychology email listservs, including the Society for Personality and Social Psychology and the Society of Australasian Social

Psychologists. We emailed authors of eligible studies, when we were able to find current contact details, and asked them to share any unpublished data or reports that met our eligibility criteria.

Study Selection

In total, 838 potentially eligible records were identified through database searches and additional search techniques. After duplicate search returns were removed using Endnote, the title and abstract of 516 references were screened for potential eligibility. A total of 449 records were excluded through the title and abstract screening process for failure to meet our eligibility criteria. The full text of 65 articles, reports and research dissertations considered potentially eligible were assessed for eligibility against our five criteria. After full-text screening, a final sample of 20 studies were included in the meta-analysis (see Figure 1).

Meta-Analytic Data Extraction

For the meta-analysis and meta-regression, the following information was extracted for each eligible study: sample size, sample type, study design, and stimulus modality. Means, standard deviations, and number of participants in each condition (cell sizes) were also extracted to calculate effect sizes for the meta-analysis. We extracted means, standard deviations, and cell sizes associated with main effects of complainant emotional demeanor focusing on conditions where the complainant's emotion was neutral or distressed. Several included studies (Ask & Landström, 2010; Bollingmo et al., 2009; Hackett et al., 2008; Klippenstine & Schuller, 2012; Schuller et al., 2010) used factorial designs and found interactions between complainant emotional demeanor and other factors on complainant credibility. As there was no parsimony in the additional factors in these studies (e.g., cognitive load, instructions about the use of complainant emotional demeanor, perpetrator speech style), we decided to extract data associated with the main effect of complainant emotional demeanor only (i.e., we did not extract any significant simple effects).

We scrutinized the sample sizes and outcome measures for studies conducted by the same authors to ensure no duplicate reports were included in the meta-analysis. One potentially eligible study was excluded on the basis that it was duplicate reporting of a study already included.

Transformations. For some studies, cell size and or standard deviation was not reported. If cell size was not reported, cell sizes were calculated based on total sample size, assuming equal distribution of participants to all conditions. If standard deviations were not reported, then they were calculated from *F* statistics using the transformation procedure from the Cochrane Handbook of Systematic Reviews of Interventions (Higgins & Deeks, 2008).

Moderator coding. To assess stimulus modality and sample type as moderators of the EVE through meta-regression, two categorical variables were coded. Stimulus modality was coded as a dichotomous variable. Experimental stimuli were coded as video format if participants saw a video of the complainant describing the alleged assault (70.0% or 14 studies). If participants read a description of the complainant's story about the alleged assault events, stimulus modality was coded as text (30.0% or 6 studies). Sample type was also coded as a dichotomous variable. Participants were classed as having professional experience if they worked in the criminal justice system (i.e., judges or police officers) or had relevant professional training (i.e., law students or police trainees; 35% or 7 studies) or no professional experience or training if they did not work in the criminal justice system (i.e., students or community members; 65% or 13 studies).

Meta-Analytic Statistical Methods

The meta-analysis and meta-regression were conducted by calculating the standardized mean difference (Hedges' *g*) within a random effects model using the metafor package in R (Viechtbauer, 2010). Standardized mean difference was selected as the aggregate effect size measure as there was variability in complainant credibility measures. As

Cohen’s *d* can overestimate the population level effect size in small samples, we used the Hedges’ *g* correction (Borenstein, Hedges, Higgins & Rothstein, 2009). Positive values for effect size indicated that emotional distress shown by the complainant increased perceptions of complainant credibility.

Heterogeneity was assessed using the *Q* statistic and the associated *p* value, I^2 and the associated confidence interval. The *Q* statistic is a ratio of observed variation (heterogeneity) to within study error or the weighted sums of squares. A significant *Q* statistic indicates that studies do not share a common effect size and the value of the *Q* statistic necessarily increases as the number of studies included in the analysis increases (Borenstein et al., 2009; Schwarzer, Carpenter & Rucker, 2015). I^2 is the proportion of between-study variation rather than sampling error (Borenstein et al., 2009; Deeks, Higgins & Altman, 2008; Schwarzer et al., 2015).

The effect of the two proposed moderators, stimulus modality and sample type, were assessed through meta-regression using a random effects model and the restricted maximum-likelihood estimation model (REML) for between-study heterogeneity. The REML estimator was selected because it is efficient and unbiased with a modest amount of effect sizes from studies with reasonable power (Viechtbauer, 2005). Stimulus modality and sample type were entered simultaneously as predictors (moderators) into the model. The outcome variable was the effect size estimate for the EVE.

Assessing Bias Using *P*-curve Analysis

A *p*-curve analysis creates a distribution of *p* values associated with the statistical tests of the key hypotheses reported in a set of articles. The distribution is created for statistically significant *p* values only. If selective reporting is a viable explanation for the statistically significant results included in the analysis, there will be more high (i.e., close to $p = .05$) than low *p* values making the *p*-curve left-skewed. If selective reporting is not a viable explanation for the results included in the *p*-curve, there should be more low (i.e., close to p = .01) than high *p* values and the curve will be right-skewed. Two distributions are created, one for *p* values under .05 (the full *p*-curve) and one for *p* values under .025 (the half *p*-curve). The half *p*-curve is a more rigorous test of whether selective reporting (via the file drawer problem or questionable research practices) can explain the statistically significant results in the analysis (Simonsohn, Simmons & Nelson, 2015). We conducted a *p*-curve analysis by specifying a study selection rule and extracting data from published papers using a *p*-curve disclosure table (Simonsohn et al., 2014a).

Study selection rule. To be included in our *p*-curve analysis, studies needed to be included in the meta-analysis and met the guidelines for inclusion in *p*-curve analysis (Simonsohn et al., 2014a; Simonsohn et al., 2015).

Data extraction. A *p*-curve disclosure table was used to extract data to perform the *p*-curve analysis as recommended by Simonsohn et al. (2014a). The disclosure table identifies test statistics selected for inclusion in the *p*-curve from each article. The disclosure table requires meta-analysts to identify and extract the original research hypothesis of interest, study design, key statistical result testing the hypothesis, and the statement of this result from the article. The *p*-curve guidelines specify which key statistical result testing the hypothesis is entered into the *p*-curve, depending on the form of interaction predicted by the original study authors in their hypotheses (i.e., in some cases the interaction term is selected and in others the simple effects; Simonsohn et al., 2014a). We encountered several difficulties as we extracted data for the *p*-curve analysis using the disclosure table.

First, our meta-analysis focused on the main effect of complainant emotional demeanor on credibility. To test the reliability of the literature included in this analysis, ideally, we would have extracted *p*-values associated with the main effect of complainant emotional demeanor in all studies. However, some studies included in our meta-analysis used factorial designs and did not make a prediction concerning the main effect of complainant emotional demeanor on credibility. As the *p*-values included in *p*-curve analysis must be associated with a prediction made by the original study authors (Simonsohn et al., 2014a), in these studies we extracted the test statistic associated with the hypothesis about the complainant's emotional demeanor. This was usually a hypothesis about complainant emotional demeanor interacting with other factors included in the design.

Second, some authors reported non-directional exploratory hypotheses and research questions in their paper, which meant that test statistics from these papers could not be extracted for the *p*-curve analysis according to Simonsohn et al.'s (2014a) guidelines.

Third, we could not extract the key statistical result for some papers according to Simonsohn et al.'s (2014a) guidelines. Some papers did not report the key statistic stipulated by the *p*-curve guidelines for the experimental design and hypothesis. In some papers, the test statistic which corresponded to the hypothesis of interest was non-significant (as *p*-curve analysis investigates the distribution of *p* values below the statistical significance criteria of $p < .05$, these values are excluded from the analysis). Of the 20 studies included in the meta-analysis, we were able to extract data for the *p*-curve analysis from 9 studies (see Table 1 for the *p*-curve disclosure table). Studies were excluded from the *p*-curve analysis because the key result testing the hypothesis was non-significant ($n = 5$) or the research hypotheses were exploratory ($n = 6$). We analyzed these data using the *p*-curve app (http://p-curve.com).

Robustness analyses. We ran additional *p*-curve analyses when the hypotheses reported by the original study authors addressed more than one dependent variable (Simonsohn et al., 2014a). To ensure that *p*-values were statistically independent as required to protect the integrity of the *p*-curve analyses, we only entered one *p*-value from any study into each *p*-curve.

Unregistered Analyses of Publication Bias

In addition to the pre-registered *p*-curve analyses, we also conducted some further tests of publication bias including a funnel plot and tests of asymmetry (Begg & Mazumdar, 1994; Egger, Smith, Schneider & Minder, 1997), a trim and fill analysis (Duval & Tweedie, 2000a), and fail-safe N analyses (Rosenthal, 1979; Rosenberg, 2005).

Funnel plot and tests of asymmetry. A funnel plot is a scatterplot of standard errors and effect size estimates (Sutton, 2009). Asymmetry in a funnel plot can indicate that studies are systematically missing from a meta-analysis as a function of study size or effect size (and can be an indication of publication bias however asymmetry can be caused by other factors; Sterne, Becker & Egger, 2006; Sutton, 2009). Asymmetry in a funnel plot can be statistically assessed through the rank correlation test (Begg & Mazumder, 1994) and the linear regression test (Egger et al., 1997). The linear regression test is generally considered to be more powerful than the rank correlation test (Sutton, 2009).

Trim and fill analysis. In trim and fill analysis, studies are simulated and imputed to make the funnel plot symmetrical and an adjusted effect size estimate is calculated (Duval & Tweedie, 2000a; 2000b). This means the analysis assumes that small effects are suppressed by publication bias (Sutton, 2009) which may not always be the case (Simonsohn et al., 2015). When high levels of between-study heterogeneity are present, the trim and fill technique does not perform well to estimate the effect of publication bias (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau & Olkin, 2003).

Fail-safe N. In the fail-safe N test (or file drawer analysis), the number of null result studies required to reduce the significance of the aggregate effect size estimate is calculated (Sutton, 2009). The Rosenthal fail-safe N test calculates the number of null effect studies that would be needed in order for the effect size estimate to become just significant at p = .05. Larger numbers of studies required to reduce the effect is an indicator that the effect is more robust (Rosenthal, 1979). The weighted fail-safe N test (Rosenberg, 2005) provides a more conservative estimate of the number of studies required to reduce the aggregate effect size estimate to just significant at p = .05 (Rosenberg, 2005; Sutton, 2009).

Exploratory Study Quality Assessment

Study quality assessment criteria. To assess the internal validity of the studies included in the meta-analysis, we executed an adapted form of the Cochrane risk of bias assessment (Higgins et al., 2011; Higgins & Altman, 2008). The risk of bias assessment focuses on identifying protections against biases that threaten the internal validity of randomized controlled trials including selection, performance, attrition, detection, and reporting bias. As the studies included in this meta-analysis are cross-sectional psychology experiments, it was necessary to adapt the criteria in the assessment tool to make it appropriate to assess the internal validity of these types of studies (as recommended in the Cochrane guidelines; Higgins & Altman, 2008; Reeves et al., 2008). We excluded reporting bias from the assessment as we investigated this through *p*-curve analysis. We also excluded attrition bias because participant attrition information is not reported for cross-sectional studies.

Table 2 displays the criteria from the Cochrane risk of bias assessment and the adapted criteria we used. We adapted the criteria based on scholarship about methodological protections against internal validity threats for cross-sectional psychology studies (e.g., Goodwin & Goodwin, 2013; Haslam & McGarty, 2008; Kite & Whitley, 2018).

Coding procedure. The criteria were coded by two independent raters and inter-rater reliability was assessed using Cohen's measure of agreement (kappa). This measure assesses the similarity of raters' categorizations accounting for chance agreement (Neuendorf, 2004). All criteria had acceptable inter-rater reliability (i.e., over .80; Krippendorff, 2004; 2013) on first coding.

Exploratory Content Analysis

We investigated the operationalization of complainant emotional demeanor and the measurement of complainant credibility in studies included in our meta-analysis through content analysis. Content analysis is a method of systematically evaluating and coding the content of text (Neuendorf, 2002). It can be used to classify the content of text based on frequently occurring words, themes or concepts within text (Krippendorff, 2013). Categories developed to classify text through content analysis must be reliable, so the analysis can be replicated by independent coders (Krippendorff, 2013; Neuendorf, 2002).

We executed two separate content analyses. The first analysis focused on evaluating how the independent variable of complainant emotional demeanor had been operationalized in included studies. We were specifically interested in the behaviors or descriptions of the complainant's behavior used to portray distress and controlled affect by the complainant. The second analysis focused on evaluating the type of questionnaire items used to measure complainant credibility in the included studies to assess the face validity of these measures.

Data extraction. The descriptions of the operationalization of complainant emotional demeanor and complainant credibility measures were extracted from full-text reports of included studies. For questionnaire items, sample items and scale anchors were extracted for coding.

Category development. The first coder developed the categories for both content analyses. For the first content analysis, focused on complainant emotional demeanor, each behavior described as part of the operationalization of complainant distress or controlled affect was extracted into a separate category (14 categories in total). For the second content analysis, focused on complainant credibility measures, nine categories were developed addressing the content of sample items and scale anchors reported. Categories for both content analyses were data driven (i.e., based on reports of the operationalization of complainant emotional demeanor and complainant credibility measures respectively).

Coding procedure. The first coder developed the coding categories and coded the data. The second coder was provided with a coding form and explanation of the coding categories. The second coder coded the data according to the coding scheme. Inter-rater reliability, via Cohen's kappa, was calculated to check the reliability of the coding scheme.

Categories which have a Cohen's kappa value of above .80 are considered to have acceptable agreement beyond chance (Krippendorff, 2004; 2013) and values between .40-.75 indicate fair to good agreement beyond chance (Banerjee, Capozzoli, McSweeney & Sinha, 1999). Categories with agreement in the fair to good range were reviewed by both coders and the definition of the category was revised. After recoding based on the revised category definitions, inter-rater reliability was re-calculated. All reported categories have kappa values above .85, suggesting acceptable reliability (Krippendorff, 2013). The first coder's ratings were reported for categories where there was disagreement between the first and second coder.

Results

Pre-Registered Analyses

Overall effect size. The final analysis contained 20 effect sizes. To test H1, we conducted a random effects model to estimate the effect size of the EVE. The random effects model resulted in an estimated effect size of g = 0.38 (95% confidence interval, [0.25; 0.51], Q = 47.43, p <.001, I^2 = 59.62%, 95% CI [29.49, 84.34]), which indicated, consistent with H1, that distressed rape complainants were perceived to be more credible than their controlled counterparts. Figure 2 shows the effect size estimates and associated confidence intervals for each study included in the meta-analysis. Table 3 displays the sample size, sample type, and stimulus modality for all studies included in the meta-analysis. The between-study heterogeneity (I^2) was substantial according to the Cochrane guidelines (Deeks et al., 2008), indicating that it was appropriate to use meta-regression to assess potential moderators.

Moderator analysis. To test the relationship between the effect size estimate and the two proposed moderators, meta-regression was conducted using a random effects model. Contrary to H2, stimulus modality did not moderate the effect size estimate for the EVE (Z = -1.01, p = .313, k = 20). H3 was also not supported. Sample type did not moderate the effect size estimate for the EVE (Z = 0.22, p = .827, k = 20).

***P*-curve analysis.** The *p*-curve was right skewed, suggesting that the studies included in the *p*-curve had evidential value (see Figure 3). The continuous tests (using Stouffer's Method) which examine whether the *p*-curve is significantly right skewed were significant for the full *p*-curve (Z = -3.61, p < .001) and the half *p*-curve (Z = -3.59, p < .001). According to the criteria provided by Simonsohn et al. (2015), if the half *p*-curve is significantly right skewed (p < .05) or the half and full *p*-curve are significantly right skewed (p < .10) then the *p*-curve indicates that the included studies have evidential value. The *p*-curve analysis suggested that the literature on included in the *p*-curve analysis had evidential value as both criteria were met.

Robustness *p*-curve analyses. We ran two robustness *p*-curve analyses. In the first *p*-curve analysis, we included *p*-values associated with dependent variables measuring negative attitudes towards the complainant that were addressed by the focal hypotheses selected. The full (Z = -2.58, p = .005) and half *p*-curve (Z = -3.90, p = .003) in this analysis were significantly right skewed. In the second analysis, we included *p*-values associated with dependent variables measuring positive attitudes towards the complainant and again the full (Z = -2.93, p = .002) and half *p*-curve (Z = -2.72, p = .003) were significantly right skewed. This suggested that the findings of the *p*-curve analysis were robust.

Unregistered Analyses of Publication Bias

Funnel plot. A scatterplot (funnel plot) of effect size estimate and standard error for all studies included in the meta-analysis is presented in Figure 4. Two statistical tests of funnel plot asymmetry, the nonparametric correlation test ($\tau = 0.11$, $p = .542$) and the regression test ($Z = 0.88$, $p = .380$) were both non-significant and suggested that the funnel plot was not asymmetric. This was consistent with the results of the *p*-curve analyses.

Trim and fill analysis. We conducted a trim and fill analysis (Duval & Tweedie, 2000a; 2000b). As there was substantial between-study heterogeneity in these data (by the Cochrane guidelines; Deeks et al., 2008), the results of the trim and fill analysis, particularly concerning the adjusted effect size estimate, should be treated with caution (Sterne, Egger & Moher, 2011; Viechtbauer, 2010). The trim and fill analysis resulted in an adjusted effect size estimate of $g = 0.34$ (95% CI, [0.21; 0.57], $Q = 55.94$, p <.001, $I^2 = 63.39\%$, 95% CI [38.09, 85.71]). Two simulated studies were imputed into the funnel plot for the adjusted analysis (see figure 5). The adjusted effect size estimate from a trim and fill analysis should not be considered the true estimate of the effect without publication bias (Viechtbauer, 2010). There is negligible change in the effect size estimate on the basis of a trim and fill analysis. This was consistent with the results of the *p*-curve analyses.

Fail-safe N. The Rosenthal fail-safe *N* test (1979) indicated that 588 studies averaging a non-significant effect would be needed to reduce the current effect size estimate to just significant ($p = .05$). Rosenthal (1979) suggested that if the number of studies needed to reduce the effect size estimate to just significant was over 500 this was an indication that effect was robust. This criterion has been criticized as being arbitrary (Rosenberg, 2005; Sutton, 2009). The Rosenberg fail-safe *N* test (2005) indicated that 414 studies averaging a non-significant effect would be needed to reduce the overall effect size estimate to just significant ($p = .05$). This was also consistent with the results of the *p*-curve analyses.

Exploratory Study Quality Analysis

Overall, the majority of studies (80% or 16 studies) had one or more methodological protections against all three (20% or four studies) or two biases (60% or 12 studies) considered in the risk of bias assessment. Only 20% (or four studies) reported protections against a single form of bias which threatens internal validity. Manipulation checks were the most common protection employed against performance bias (by 12 studies or 60%), reporting a reliability analysis for the credibility measure was the most common protection employed against detection bias (by 11 studies or 55%) and random allocation was used to protect against selection bias in half of studies (11 studies or 55%). The number and proportion of studies that met each risk of bias criteria is reported in Table 4.

Exploratory Content Analyses

Operationalization of complainant emotional demeanor. Two studies were excluded from this analysis as no description of the behaviors used to operationalize complainant distress or controlled affect were reported. Both coders were in agreement that these studies should be excluded. The proportion (and number) of studies, as well as the kappa inter-rater reliability coefficients for each category (i.e., complainant behavior) is presented in Table 5.

All studies described using more than one behavior to operationalize complainant distress (18 studies or 100%). The number of behaviors used to operationalize complainant distress ranged from two to five (mode = 4). All studies but one used crying to operationalize complainant distress. Around half of the studies also used either distressed facial expression, hesitations in speech or trembling voice to manipulate distress. For studies that used text-based stimulus to manipulate complainant distress, half had the complainant describe herself as distressed or had another person describe the complainant as distressed. No video studies used this form of complainant behavior to operationalize distress.

Two-thirds of studies (12 studies or 66.7%) described using one or more behavior to operationalize controlled emotional demeanor. Of the studies which did not describe any specific behaviors used to operationalize controlled emotional demeanor (6 studies or 33.3%), they provided a general description (e.g., the complainant was calm). The number of behaviors described to operationalize controlled demeanor ranged from one to five (mode = 1). To operationalize controlled affect, half of the studies reported that complainant's demeanor was factual. Just over a quarter of all studies also operationalized controlled affect as having a steady voice or the complainant appearing confident (see Table 5).

Measurement of complainant credibility. The proportion (and number) of all studies to use various items in their complainant credibility measure is reported in Table 6. All studies used one or more items describing a face valid concept related to credibility. The most commonly used face valid concepts were believability, credibility, and/or truthfulness. Some other concepts used as items in credibility measures were accuracy and whether the complainant had attempted to hide the truth in some way (see Table 6, category hiding truth). Some complainant credibility measures contained items tapping concepts not directly related to complainant credibility including defendant guilt (which was reversed coded) and the observer's confidence in their decision (see Table 6).

Further Exploratory Analysis

We were unable to include categories from the risk of bias assessment and content analyses as moderators of the EVE because many were skewed (e.g., one or more methodological protections against all three biases; crying to manipulate complainant distress). Meta-regression is sensitive to skew in categorical moderators which can undermine the reliability of the analysis (Borenstein et al., 2009; Schwarzer et al., 2015).

Discussion

Our aim in this review was to estimate the size and robustness of the effect of complainant emotional demeanor on credibility judgments for female adult rape complainants. We discuss our pre-registered analyses and suggest conceptual and theoretical directions for future research. Then we discuss our unregistered and exploratory analyses and suggest methodological improvements for future research.

Complainant Distress and Credibility

We expected a significant and positive effect size estimate, such that a distressed rape complainant would be judged more credible than her emotionally controlled counterpart (H1). This hypothesis was supported, and our meta-analysis suggested that there is a small to moderate effect, by Cohen's (1988; 1992) guidelines, of complainant emotional demeanor on credibility judgments. However, a review of effect sizes in social psychology called for empirically based cut-offs on the 25th, 50th, and 75th percentiles of effect sizes to be 0.15, 0.40, and 0.70 for Cohen's *d* or Hedges' *g* (Lovakov & Agadullina, 2017). By these guidelines, our effect size estimate is moderate.

In our review, we used the Heuristic-Systematic Model to explain how complainant emotional demeanor might influence credibility judgments (Chaiken, 1980). We suggested that perceivers in rape cases were likely to be highly motivated to engage in information processing (e.g., Thomas, 2010; White et al., 2010). In everyday credibility judgments, motivated perceivers engage in systematic processing which overrides the effect of heuristic cues, like complainant emotional demeanor, on their judgments (the attenuation hypothesis; Chaiken & Maheswaran, 1994; Reinhard & Sporer, 2008; 2010). However, because typical rape cases have ambiguous evidence (i.e., cases lack corroborative evidence; Cox, 2015; Quadara et al., 2013), highly motivated perceivers are likely to be influenced by heuristic cues in their systematic processing of information about the case (the bias hypothesis or biased systematic information processing; Chaiken & Maheswaran, 1994). The bias hypothesis is one plausible explanation for the observed effect of complainant emotional demeanor on perceiver's credibility judgments. The bias hypothesis predicts in rape cases, perceivers may be influenced by a misleading heuristic (complainant emotional demeanor) not because they are not motivated to engage in effortful information processing, but because the available evidence means decision-making is inherently difficult. However, other underlying mechanisms, such as expectancy violation (e.g., Hackett et al., 2008; Ask & Landström, 2010), perceivers' emotional responses to the complainant (e.g., Ask & Landström, 2010), or perceiver empathy (Ask, 2018) may also explain the relationship between complainant emotional demeanor and credibility judgments.

We extracted data associated with the main effect of complainant emotional demeanor on credibility judgments for our meta-analysis. Several included studies found interactions between complainant emotional demeanor and another factor on credibility. By selecting the main effect, it is possible the effect size was underestimated in these studies. However, selecting main effects allowed us to assess the size of the effect of complainant emotional demeanor on credibility judgments, irrespective of other moderating factors, which was our focus in this review.

We used *p*-curve analysis to determine whether selective reporting, via publication bias or questionable research practices, could explain the statistically significant findings reported in this literature. The *p*-curve, and robustness *p*-curves, were significantly right-skewed, and suggested that selective reporting (via the file drawer problem or questionable research practices) is not a viable explanation for the literature on the effect of female adult complainant emotional demeanor on credibility judgments.

Tests of Moderators

We considered whether stimulus modality or sample type modified the effect size estimate for the emotional victim effect. We expected the effect of complainant emotional demeanor would be stronger when participants watched a video of the complainant compared with reading a synopsis of complainant evidence (H2). This hypothesis was not supported. This means complainant distress increases credibility judgments even through less ecologically valid experimental paradigms using written complainant evidence. This is consistent with the results of other meta-analytic reviews of jury decision-making studies which suggest that stimulus modality has limited impact on decisions made about the case (e.g., verdict and guilt likelihood; Bornstein et al., 2017).

We reasoned when complainant emotional demeanor was more salient, it would be a more accessible heuristic cue for perceivers to use to judge credibility (Chaiken & Eagly, 1983). When a heuristic cue is more accessible, perceivers view it as more reliable and relevant to use to make the judgment (Chen & Chaiken, 1999). If complainant emotional demeanor is viewed as relevant to judging credibility, then it may have a larger effect on biased systematic information processing by guiding attention in systematic processing to information consistent with the heuristic cue of complainant emotional demeanor. However, some evidence suggested that complainant emotional demeanor may be more salient in video evidence (e.g., Chaiken & Eagly, 1976; 1983) and other evidence indicated that complainant emotional demeanor may be more salient in written evidence (e.g., Bless et al., 1993; Grice, 1975; Kite & Whitley, 2018; Schwarz et al., 1991). Our results suggest no difference in the effect of complainant emotional demeanor on credibility between video and written complainant evidence. Our content analysis of complainant emotional demeanor manipulations suggests that many studies using written complainant evidence included the complainant, or another witness, identifying the complainant's emotional demeanor for the perceiver. For example, the complainant would say she was distressed, and complainant emotional demeanor would have been highly salient for perceivers.

We also expected that the effect of complainant emotional demeanor on credibility judgments would be smaller for perceivers with professional experience or training (e.g., police officers or judges) compared to those without professional experience or training (i.e., community members or university students; H3). This hypothesis was not supported. There is debate about whether professional experience or training prevents judges, lawyers, and police officers from relying on heuristic cues to make judgments in their professional roles (e.g., Reinhard et al., 2012 cf. Ask, 2010). Our results suggest that being a member of a professional group does not reduce the influence of potential heuristic cues, like complainant emotional demeanor, on credibility judgments of rape complainants. This is consistent with research which suggests that professional expertise or training does not prevent criminal justice professionals from being influenced by motivated cognition or heuristic cues in their judgments (e.g., Ask, Grahag, & Rebelius, 2011; Miller, 2018; Rachlinski & Wistrich, 2017).

We reasoned that increased familiarity with the judgment context, which criminal justice professionals and trainees possess relative to lay perceivers, may make them more willing to engage with conflicting complex information in making credibility judgments (Reinhard et al., 2011; 2012). This may make professionals and trainees more likely to challenge the reliability of complainant emotional demeanor to judge credibility within biased systematic information processing, reducing the effect of complainant emotional demeanor on their credibility judgments relative to lay perceivers. However, survey evidence suggests that some police officers and prosecutors view complainant emotional demeanor as relevant and reliable information to judge credibility (e.g., Ask, 2010; Heenan & Murray, 2006). If complainant emotional demeanor is viewed as a reliable and therefore relevant cue to judge credibility, it is unlikely to be challenged in systematic processing, and biased systematic information processing may take place. Within biased information processing the evidence focused on in information processing will align with the heuristic cue (i.e., when the complainant is distressed perceivers will focus on other evidence in favor of her credibility and vice versa for complainants who present with controlled affect). This may be a viable explanation for why sample type did not moderate the effect of complainant emotional demeanor on credibility judgments.

An alternative explanation is, if experience rather than training encourages professionals to challenge the relevance of complainant emotional demeanor to judge credibility, sample type coded at the study level may not be a sufficiently precise measure of professional expertise. Several included studies used police trainees (e.g., Ask & Landström, 2010; Baldry, Winkel & Enthovan, 1997) or law students (i.e., Bohner & Schapansky, 2018). These types of participants may be too early in their criminal justice careers to have developed a sufficient level of expertise in decision-making, thus reducing the differences between the two groups in expertise. We were unable to do further exploratory sub-group analyses on differences between police trainees and officers or law students and judges as we did not have sufficient studies to power the analysis (Schwarzer et al., 2015).

Conceptual and Theoretical Issues for Future Research

Our meta-analytic review suggested that there is a substantial amount of between-study heterogeneity present in this literature and there may be powerful unidentified moderators of the effect of complainant emotional demeanor on credibility judgments (Deeks et al., 2008). We suggest that future research should further investigate moderators and theoretically-derived mediators of the effect of complainant emotional demeanor on credibility judgments. This would move the literature towards a comprehensive understanding of the operation and boundaries of the effect.

Conceptual moderators. Future research should continue to study criminal justice professionals focusing on senior lawyers, judges, and police officers so that any influence of professional experience on the emotional victim effect can be fully understood. We also suggest that future research study trainee groups so the effect of modern professional training (which may include updated specialized training for rape cases) on the emotional victim effect can be understood. The random effect model meta-regression analysis we used to examine professional experience was low powered. This means a non-significant result should not be considered conclusive evidence that the moderator does not have an influence on the effect size (Borenstein et al., 2009; Deeks et al., 2008).

Future research should also aim to understand how complainant emotional demeanor interacts with other factors which influence perceptions of rape complainants, like perceiver gender or rape victim stereotypes. Whereas there is limited existing evidence to suggest that perceiver gender influences how complainant emotional demeanor is used to judge credibility (Bollingmo et al., 2007; 2009; Calhoun et al., 1981; Kaufmann et al., 2003; Klippenstine, 2010; Schuller et al., 2010; Wessel et al., 2006 cf. Bohner & Schapansky, 2018), perceiver gender does influence perceptions of rape complainants more generally (e.g., Anderson et al., 1997 cf. Gravelin et al., 2019). Gender differences in emotion recognition, as suggested by the emotional sensitivity hypothesis, could drive gendered differences in the use of complainant emotion as a reliable and relevant heuristic cue to judge credibility (cf., a recent meta-analytic review did not find support for gendered differences in emotion perception; Fischer, Kret & Broekens, 2018). Similarly, few studies have investigated or shown that rape victim stereotypes influence the effect of complainant emotional demeanor (Kaufmann et al., 2003; Schuller et al., 2010), despite rape victim stereotypes being a strong influence on how rape complainants are perceived generally (Davies et al., 2009; Masser et al., 2010). However, existing psychology studies of these effects may be underpowered to detect interactions (Giner-Sorolla, 2018), so properly powered investigations of these potential moderators should be undertaken in future research. Exploring whether other heuristic cues about rape complainants modify the effect of complainant emotional demeanor is critical to a parsimonious understanding of how perceivers judge rape complainant credibility.

Theoretical explanations. In our review, we used the Heuristic-Systematic Model (Chaiken, 1980) to explain the effect of complainant emotional demeanor on credibility judgments. Although perceivers might be motivated to process information carefully to judge complainant credibility, the bias hypothesis suggests that in the ambiguous context of rape cases heuristic cues, such as complainant emotional demeanor, influence evidence focused on in systematic information processing (Chaiken & Maheswaran, 1994). Our results suggest that complainant emotional demeanor consistently influences credibility judgments, regardless of sample type or stimulus modality. However, our review cannot provide evidence of the mechanisms which might explain how complainant emotional demeanor is used to judge credibility.

We suggest that future research investigate the bias hypothesis as one possible explanation for the effect that complainant emotional demeanor has on credibility judgments. The majority of research adopting the Heuristic-Systematic Model has focused on the attenuation hypothesis and demonstrates that systematic processing has a curative effect on the potentially biasing effects of heuristics (e.g., Reinhard et al., 2011; 2012). By comparison, the bias hypothesis has been investigated in research substantially less (Chen & Chaiken, 1999; Todorov et al., 2002). Rape cases offer an inherently ambiguous real-life context in which both the operation and potential boundaries of the bias hypothesis can be investigated and understood.

Ask and Landström (2010) examined mediators of the emotional victim effect in the context of cognitive load, an ecologically valid moderator for criminal justice professionals.

Although not intended as a test of the Heuristic-Systematic Model, by examining the effect of cognitive load and rape complainant emotional demeanor on police trainees' complainant credibility judgments, Ask and Landström (2010) also provided evidence for the attenuation hypothesis (Chaiken, 1980). That is, when trainees were not under cognitive load (and likely to systematically process case information), the effect of complainant emotional demeanor on credibility judgments was attenuated.

However, for a complete test of the attenuation hypothesis, which assumes that attenuation of heuristic cues occurs only for highly motivated perceivers (Chaiken & Chen, 1999; Todorov et al., 2002), perceivers' motivation to engage in information processing needs to be measured to check this assumption. Perceivers' quality and quantity of information processing also need to be measured to confirm perceivers used systematic information processing to judge complainant credibility. Measuring the quality of perceivers' information processing also allows for confirmation that complainant emotional demeanor had a limited impact on systematic information processing (i.e., attenuation through systematic processing occurred). Neither perceiver motivation to engage in information processing nor quantity and quality of information processing was measured by Ask and Landström (2010). We suggest that future research should investigate the attenuation hypothesis as a possible explanation for the relationship between complainant emotional demeanor and credibility judgments. We also recommend that future research account for perceiver motivation to engage in information processing (e.g., through survey questions; Reinhard & Sporer, 2008) and measure quality and quantity of information processing (e.g., through a cognitive response task; Chaiken & Maheswaran, 1994) to allow for a full examination of the attenuation or bias hypotheses as mechanisms through which the emotional victim effect occurs.

The Heuristic-Systematic Model, and the bias hypothesis specifically, offer one plausible explanation for how complainant emotional demeanor influences credibility judgments. However, there are other plausible explanations for the mechanisms underpinning this effect. Only a few studies to date have empirically investigated possible theoretical mechanisms (e.g., Ask, 2018; Ask & Landström, 2010; Bohner & Schapansky, 2018; Hackett et al., 2008). Moreover, fewer studies have examined potential mediators of the effect, for example empathy (Ask, 2018) or expectancy violation (Ask & Landström, 2010). We suggest that future research focus on investigating theoretically driven mediators of the effect.

Research investigating theoretically driven explanations of how complainant emotional demeanor is influencing credibility judgments is critical. Our review suggests that complainant emotional demeanor could be adversely, and prejudicially, influencing complainant credibility judgments made across the criminal justice system by police officers, lawyers, judges, and jurors. This means complainant emotional demeanor could be, in part, an explanation for rape case attrition (Alderden & Ullman, 2012; Spohn & Tellis, 2012). Understanding the mechanisms through which complainant emotional demeanor influences credibility judgments is the first step in designing effective intervention, by identifying what needs to be targeted in intervention attempts, to prevent complainant emotional demeanor from prejudicing credibility judgments. For example, if future research finds the bias hypothesis adequately explains the link between complainant emotional demeanor and credibility judgments, then interventions will need to focus on reducing or eliminating perceptions of complainant emotional demeanor as a reliable or relevant cue to judge credibility. Interventions like giving directions to jurors to be more accurate or impartial (e.g., in Australia; Queensland Courts, 2017), which affect motivation (e.g., Reinhard, 2010; Reinhard & Sporer, 2008; 2010; Chaiken & Maheswaran, 1994), are unlikely to change the use of complainant emotional demeanor as a heuristic cue within biased systematic information processing. Alternatively, if perceiver empathy explains the link between complainant emotional demeanor and credibility judgments, then empathetic responses would need to be targeted in such interventions.

Unregistered Analyses of Publication Bias

There are numerous analyses that can be used as indicators of the effect of publication bias within a particular literature. Many of these analyses make different assumptions about the causes of publication bias. For example, fail-safe N and *p*-curves, in part, assume publication bias suppresses non-significant findings (Rosenthal, 1979), whereas funnel plots and the trim and fill analysis assume that small effects are suppressed by publication bias (Sutton, 2009). All metrics of publication bias also operate poorly under certain conditions. For example, trim and fill analysis performs poorly to estimate publication bias when between-study heterogeneity is high (Peters et al., 2007; Terrin et al., 2003). By examining multiple indicators of publication bias, we were able to look at cumulative evidence for the effect of publication bias on our overall effect size estimate (a strategy used by meta-analysts; e.g., Pettigrew & Tropp, 2006). The two tests of funnel plot asymmetry and the trim and fill analysis suggest that there is a limited effect of publication bias on our overall effect size estimate. Further, the two tests of fail-safe N suggest that a large number of non-significant studies would be required to reduce the overall effect size estimate to just significant. In combination, these additional analyses and indicators of publication bias support the results of the *p*-curve and robustness *p*-curves. Across all these metrics, there is limited evidence to suggest that publication bias has exerted an effect on the meta-analytic results.

Risk of Bias Assessment

The risk of bias assessment suggested that included studies had reasonable internal validity. Most studies had methodological protections against two or three biases which threaten internal validity. Our assessment was based on study reports, which may underestimate internal validity, as sometimes reports omit methodological details (Higgins & Altman, 2008). For example, we found half of the reports specified participants were randomly assigned to conditions (which protects against selection bias). However, random assignment to experimental conditions is ubiquitous in psychology experiments (e.g., Kite & Whitley, 2018) so this information may simply have been omitted from the remaining study reports.

Methodological protections against detection bias could be strengthened. Most studies did not report any validity assessment for complainant credibility measures. This is consistent with a recent review of social psychology research, which suggested that validity assessments are under reported (Flake, Pek & Hehman, 2017). Also, most studies reported Cronbach's alpha to demonstrate measure reliability. Although Cronbach's alpha is used typically as the sole indicator of reliability, it is problematic as a measure of reliability (Flake et al., 2017) and does not assess measure structure (i.e., unidimensionality or homogeneity; Schmitt, 1996; Sijtsma, 2009).

More robust protections against performance bias could be used in future research. Just half of the studies we examined included information about manipulation checks (and indicated whether the complainant emotional demeanor manipulation was successful). Few studies used a suspicion probe to screen for demand characteristics which can be a viable explanation for the study results (Kite & Whitley, 2018).

Exploratory Content Analysis

We found that complainant distress was commonly operationalized by crying, distressed facial expression, trembling voice, and speech hesitations. In contrast, controlled affect was operationalized by a factual or confident manner and a steady voice. However, there were several other behaviors used to operationalize distress and controlled affect which varied across studies. We also found variability in complainant credibility measures. All studies used a measure that contained one or more face valid items about the believability, credibility or truthfulness of the complainant. Yet, some less face valid items were also included (for example, items about perpetrator guilt). Few measures included items about the accuracy of the complainant's testimony and no measures contained items about the reliability of the complainant's evidence. This matches a recent review of credibility measures for child sexual abuse complainants, which found that items about accuracy and reliability were rarely used (Voogt, Klettle & Crossman, 2016). This omission is problematic as jurors are typically directed to consider the complainant's reliability related to credibility (e.g., in Australia and Canada; McKimmie et al., 2014; Porter & ten Brinke, 2009; *White v R*, 1947).

Improving Simulation Methods for Future Research

We focused on studies which used an experimental methodology in this review. As the effect of complainant emotional demeanor on credibility judgments may inform policy changes, it is particularly important that the methodology used to investigate the effect is rigorous (Kerr, 2017). We make five suggestions for improving simulated decision-making studies for future research below including increasing internal validity, refining complainant emotion manipulations and credibility measures, performing direct replication studies, implementing ecologically valid protocols, and sharing data and study materials.

Strengthening internal validity. Based on the risk of bias assessment, we recommend increasing protections to prevent performance and detection biases from threatening internal validity. In particular, we suggest that manipulations of complainant emotional demeanor are assessed using manipulation checks (e.g., questions about the emotion portrayed). Suspicion probes, which ask participants to speculate about the study hypotheses, should be included to screen for demand characteristics (Kite & Whitley, 2018).

Future research should investigate the reliability and validity of complainant credibility measures. For multiple item measures, results of exploratory factor analyses (to investigate the structure of measures) and other reliability assessments (e.g., Cronbach's alpha) should be conducted (Haslam & McGarty, 2008). Validity assessments of credibility measures should be executed (Flake et al., 2017). For example, researchers should detail the item development procedure or demonstrate convergent or discriminant validity with other established measures (e.g., the Witness Credibility Scale; Brodsky, Griffin & Cramer, 2010). We suggest that authors report all methodological features used to protect internal validity (e.g., randomization to conditions) in publication of the work. This provides the reader with information to assess the strength of evidence the study contributes.

Manipulating emotional demeanor and measuring credibility. Given the variation in the operationalization of complainant distress and controlled affect, we suggest that future research should explore whether changes in the complainant's emotional display modify how complainant emotional demeanor influences credibility. Content of emotional displays can change recognition of emotional states (e.g., tears make sad faces seem sadder to perceivers; Balsters, Krahmer, Swerts, & Vingerhoets, 2013) and rape complainants who display anger are sometimes viewed as less credible (Bohner & Schapansky, 2018 cf. Vrij & Fischer, 1997). More studies should investigate whether different negative emotional states (e.g., anger, defiance, fear) influence credibility judgments, to explore the extent and reliability of this effect.

We suggest that credibility measures should focus on issues perceivers are legally required to consider when making credibility judgments. For example, if jurors are required to assess the honesty and reliability of the witness to judge credibility, credibility measures should include only these items. This would allow researchers to assess whether perceivers are using legal guidance in their credibility judgments, which is a research priority for criminal justice professionals (Kerr, 2017).

Performing direct replications. The reproducibility of experimental research findings in psychology (Open Science Collaboration [OSC], 2015) and other sciences (e.g., McNutt, 2014) has been queried. Two large-scale replication projects in psychology suggested that the effect size produced by replication studies is half the original effect size reported (Camerer et al., 2018; OSC, 2015). The size of the originally reported effect and the statistical power of the replication study were positively correlated with whether the effect replicates (Bakker, van Dijk & Wicherts, 2012; OSC, 2015). As the effect of complainant emotional demeanor on credibility judgments may be used to affect policy change, it is critical the effect is demonstrated to be robust. Our review offers an initial positive assessment of robustness of the emotional victim effect, but we suggest that more properly powered direct replications are needed to further investigate the robustness of the effect.

Using ecologically valid protocols. We recommend that researchers should use ecologically valid experimental protocols to replicate the effect of complainant emotional demeanor on credibility judgments (Bornstein, 2017; Koehler & Meixner, 2017 cf. Kerr, 2017; Kerr & Bray; 2005). Simulated decision-making study protocols are often criticized for lacking ecological validity, including inadequate sampling, inappropriate measurement of dependent variables or missing investigative interview or trial procedure (Bornstein, 1999; Diamond, 1997; Koehler & Meixner, 2017). Most of the studies we reviewed that used a jury simulation paradigm did not include deliberation (cf. Dahl et al., 2007) and used short complainant testimony extracts without examination by the prosecutor or cross-examination by defense counsel. Cross-examination can reduce credibility judgments of rape complainants by highlighting complainant behavior that does not match rape victim stereotypes (Zajac & Cannan, 2009; Zydervelt, Zajac, Kaladelfos, & Westera, 2016). We suggest that more trial procedures that may influence the credibility judgments of complainants be incorporated in future studies.

Sharing materials and data. Generally, ecological validity is positively correlated with external validity. As the similarity between the experimental context and the real world increases so does the likelihood the effect will reproduce (external validity; Wiener, Krauss & Lieberman, 2011). However, police investigations and trial procedures vary greatly between different legal jurisdictions. Using more ecologically valid protocols for a particular legal jurisdiction can introduce substantial variation into simulated decision-making studies. This variation may reduce the prospect of reproducibility (Krauss & Lieberman, 2017). By sharing materials, research groups would be able to see how other researchers have incorporated specific investigation and trial procedure into materials and adapt these for their own legal jurisdiction (for conceptual replication) or use the materials (for direct replication). Sharing data and materials would also enable validity assessments of credibility measures. Data and materials can now be easily shared through online platforms like the Open Science Framework (https://osf.io) or other repositories (see the Registry of Research Data Repositories; https://www.re3data.org). We suggest that researchers make their materials and data available to encourage collaborative work to improve the robustness and usefulness of simulated decision-making research.

Conclusion

The results of this meta-analysis and *p*-curve analysis suggest that complainant emotional demeanor has a small to moderate effect on credibility judgments and reporting bias is not a likely explanation for significant results reported within this literature. Complainant emotional demeanor is not diagnostic of witness honesty, the key component of a credibility decision (e.g. in Australia and Canada; McKimmie et al., 2014; Porter & ten Brinke, 2009). That a potentially prejudicial factor influences credibility judgments should be a concern, especially given that complainant credibility is a key determinant of whether a rape case proceeds in the criminal justice system (Brown et al., 2007; O'Neal, 2017). Future research should focus on identifying the mechanisms through which rape complainant emotional demeanor influences credibility judgments. This is important so that effective interventions can be designed to target the underlying mechanisms to prevent complainant emotion from influencing credibility judgments made by criminal justice professionals and jurors.

References

References marked with an asterisk indicate studies included in the meta-analysis

Abrams, D., Viki, G. T., Masser, B., & Bohner, G. (2003). Perceptions of stranger and acquaintance rape: The role of benevolent and hostile sexism in victim blame and rape proclivity. *Journal of Personality and Social Psychology, 84*(1), 111-125. doi:10.1037/0022-3514.84.1.111

Alderden, M. A., & Ullman, S. E. (2012). Creating a more complete and current picture: examining police and prosecutor decision-making when processing sexual assault cases. *Violence Against Women, 18*(5), 525-551. doi:10.1177/1077801212453867

Anderson, K. B., Cooper, H., & Okamura, L. (1997). Individual differences and attitudes toward rape: A meta-analytic review. *Personality and Social Psychology Bulletin, 23*(3), 295-315. doi:10.1177/0146167297233008

Angelone, D. J., Mitchell, D., & Grossi, L. (2015). Men's perceptions of an acquaintance rape: The role of relationship length, victim resistance, and gender role attitudes. *Journal of Interpersonal Violence, 30*(13), 2278-2303. doi:10.1177/0886260514552448

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes*, *52*(2), 292–306. doi: 10.1016/0749-5978(92)90040-E

Ask, K. (2010). A survey of police officers' and prosecutors' beliefs about crime victim behaviors. *Journal of Interpersonal Violence, 25*(6), 1132-1149. doi:10.1177/0886260509340535

*Ask, K. (2018). Complainant emotional expressions and perceived credibility: Exploring the role of perceivers' facial mimicry and empathy. *Legal and Criminological Psychology, 23*(2), 252-264. doi:10.1111/lcrp.12132

Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: the need for cognitive closure. *Journal of Investigative Psychology and Offender Profiling*, *2*(1), 43–63. doi: 10.1002/jip.19

Ask, K., Granhag, P. A., & Rebelius, A. (2011). Investigators under influence: How social norms activate goal-directed processing of criminal evidence. *Applied Cognitive Psychology*, *25*(4), 548–553. doi: 10.1002/acp.1724

*Ask, K., & Landström, S. (2010). Why emotions matter: Expectancy violation and affective response mediate the emotional victim effect. *Law and Human Behavior, 34*(5), 392-401. doi:10.1007/s10979-009-9208-6

Australian Bureau of Statistics. (2017). *Personal Safety in Australia: Key Findings.* Report retrieved from https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4906.0~2016~Main%20Features~Key%20Findings~1

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

*Baldry, A. C. (1996). Rape victims' risk of secondary victimization by police officers. *Issues in Criminological & Legal Psychology, 25*, 65-68.

*Baldry, A. C., & Winkel, F. W. (1998). Perceptions of the credibility and evidential value of victim and suspect statements in interviews. *Psychology and Criminal Justice: International Review of Theory and Practice*, 74-82.

*Baldry, A. C., Winkel, F., & Enthoven, D. (1997). Paralinguistic and nonverbal triggers of biased credibility assessments of rape victims in Dutch police officers: An experimental study of “nonevidentiary” bias. *Advances in Psychology and Law*, 163-174.

Balsters, M. J. H., Krahmer, E. J., Swerts, M. G. J., & Vingerhoets, A. J. J. M. (2013). Emotional tears facilitate the recognition of sadness and the perceived need for social support. *Evolutionary Psychology, 11*(1), 147470491301100114. doi:10.1177/147470491301100114

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics, 27*(1), 3-23. doi:10.2307/3315487

Bédard, J., & Chi, M. (1992). Expertise. *Current Directions in Psychological Science, 1,* 135-139.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088. doi: 10.2307/2533446

Bieneck, S., & Krahé, B. (2011). Blaming the victim and exonerating the perpetrator in cases of rape and robbery: Is there a double standard? *Journal of Interpersonal Violence, 26*(9), 1785-1797. doi:10.1177/0886260510372945

Bless, H., Strack, F., & Schwarz, N. (1993). The informative functions of research procedures: Bias and the logic of conversation. *European Journal of Social Psychology*, *23*(2), 149–165. https://doi.org/10.1002/ejsp.2420230204

Bluett-Boyd, N., & Fileborn, B. (2014). *Victim/survivor focused justice responses and reforms to criminal court practice*. Canberra, Australia: Australian Institute of Family Studies.

*Bohner, G., & Schapansky, E. (2018). Law students' judgments of a rape victim's statement: The role of displays of emotion and acceptance of sexual aggression myths. *International Journal of Conflict and Violence, 12,* 1-13.doi: 10.4119/UNIBI/ijcv.635

*Bollingmo, G. C., Wessel, E. O., Eilertsen, D. E., & Magnussen, S. (2007). Credibility of the emotional witness: A study of ratings by police investigators. *Psychology, Crime & Law, 14*(1), 29-40. doi:10.1080/10683160701368412

*Bollingmo, G., Wessel, E., Sandvold, Y., Eilertsen, D. E., & Magnussen, S. (2009). The effect of biased and non-biased information on judgments of witness credibility. *Psychology, Crime & Law, 15*(1), 61-71. doi:10.1080/10683160802131107

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214-234. doi:10.1207/s15327957pspr1003_2

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*: John Wiley & Sons, Ltd.

Bornstein, B. H. (1999). The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior, 23*(1), 75-91.

Bornstein, B. H. (2017). Jury simulation research: Pros, cons, trends, and alternatives. In M. B. Kovera (Ed.), *The psychology of juries* (pp. 207-226). Washington, DC: American Psychological Association.

Bornstein, B. H., Golding, J. M., Neuschatz, J., Kimbrough, C., Reed, K., Magyarics, C., & Luecht, K. (2017). Mock juror sampling issues in jury simulation research: A meta-analysis. *Law and Human Behavior, 41*(1), 13-28. doi:10.1037/lhb0000223

Bridges, J., & McGrail, C. (1989). Attributions of responsibility for date and stranger rape. *Sex Roles, 21*(3-4), 273-286. doi:10.1007/BF00289907

Brodsky, S. L., Griffin, M. P., & Cramer, R. J. (2010). The Witness Credibility Scale: An outcome measure for expert witness research. *Behavioral Sciences & the Law, 28*(6), 892-907. doi:10.1002/bsl.917

Brown, J. M., Hamilton, C., & O'Neill, D. (2007). Characteristics associated with rape attrition and the role played by skepticism or legal rationality by investigators and prosecutors. *Psychology, Crime & Law, 13*(4), 355-370. doi:10.1080/10683160601060507

Bruns, S. B., & Ioannidis, J. P. A. (2016). p-Curve and p-Hacking in observational research. *PLoS ONE*, *11*(2), e0149144. doi: 10.1371/journal.pone.0149144

Burgess, A. W., & Carretta, C. M. (2016). Rape and its impact on the victim. In R. R. Hazelwood & A. W. Burgess (Eds.), *Practical Aspects of Rape Investigation: A Multidisciplinary Approach* (pp. 3-18). United States: Taylor and Francis.

Burgess, A. W., & Holmstrom, L. L. (1974). Rape trauma syndrome. *American Journal of Psychiatry, 131*(9), 981-986.

Burgess, A. W., & Holmstrom, L. L. (1986). Adaptive strategies and recovery from rape. In R. H. Moos (Ed.), *Coping with Life Crises: An Integrated Approach* (pp. 353-365). Boston, MA: Springer US.

*Calhoun, L. G., Cann, A., Selby, J. W., & Magee, D. L. (1981). Victim emotional response: Effects on social reaction to victims of rape. *British Journal of Social Psychology, 20*(1), 17-21. doi:10.1111/j.2044-8309.1981.tb00468.x

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, *2*(9), 637-644.

Campbell, R. (1995). The role of work experience and individual beliefs in police officers' perceptions of date rape: An integration of quantitative and qualitative methods. *American Journal of Community Psychology, 23*(2), 249-277. doi:10.1007/BF02506938

Carretta, C., & Burgess, A. (2013). Symptom responses to a continuum of sexual trauma. *Violence and Victims, 28*(2), 248-258. doi:10.1891/0886-6708.vv-d-12-00011

Carroll, M. H., & Clark, M. D. (2006). Men's acquaintance rape scripts: A comparison between a regional university and a military academy. *Sex Roles, 55*(7-8), 469-480. doi:10.1007/s11199-006-9102-3

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752-766.

Chaiken, S., & Eagly, A. H. (1976). Communication modality as a determinant of message persuasiveness and message comprehensibility. *Journal of Personality and Social Psychology*, *34*(4), 605-614. doi: http://dx.doi.org/10.1037/0022-3514.34.4.605

Chaiken, S., & Eagly, A. H. (1983). Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of Personality and Social Psychology*, *45*(2), 241-256. doi: http://dx.doi.org/10.1037/0022-3514.45.2.241

Chaiken, S., & Ledgerwood, A. (2012). A theory of heuristic and systematic information processing. In P. A. M. V. Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *A handbook of theories of social psychology*. London: SAGE Publications.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212-252). New York: Guilford Press.

Chaiken, S., & Maheswaran, D. (1994) Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, *66*, 460–473. doi:http://dx.doi.org/10.1037/0022- 3514.66.3.460

Chambers, C. (2017). *The seven deadly sins of psychology*. New Jersey: Princeton University Press.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In *Dual-process theories in social psychology* (pp. 73–96). Guilford Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Cox, P. (2015). *Violence against women: Additional analysis of the Australian Bureau of Statistics Personal Safety Survey, 2012*. Australia's National Research Organisation for Women's Safety: Sydney.

*Dahl, J., Enemo, I., Drevland, G. C. B., Wessel, E., Eilertsen, D. E., & Magnussen, S. (2007). Displayed emotions and witness credibility: A comparison of judgements by individuals and mock juries. *Applied Cognitive Psychology, 21*(9), 1145-1155. doi:10.1002/acp.1320

Daly, K., & Bouhours, B. (2010). Rape and attrition in the legal process: A comparative analysis of five countries. *Crime and Justice, 39*(1), 565-650. doi:10.1086/653101

Darwinkel, E., Powell, M., & Tidmarsh, P. (2013). Improving police officers' perceptions of sexual offending through intensive training. *Criminal Justice and Behavior, 40*(8), 895-908. doi:10.1177/0093854813475348

Davies, M., Rogers, P., & Whitelegg, L. (2009). Effects of victim gender, victim sexual orientation, victim response and respondent gender on judgements of blame in a hypothetical adolescent rape. *Legal and Criminological Psychology, 14*(2), 331-338. doi:10.1348/978185408X386030

Deeks, J., Higgins, J., & Altman, D. G. (2008). Analysing data and undertaking meta-analyses. In J. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 243-296). Chichester, United Kingdom: John Wiley & Sons.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118. doi:10.1037/0033-2909.129.1.74

Diamond, S. S. (1997). Illuminations and shadows from jury simulations. *Law and Human Behavior, 21*(5), 561-571.

Duval, S., & Tweedie, R. (2000a). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463.

Duval, S., & Tweedie, R. (2000b). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98. doi:10.1080/01621459.2000.10473905

Edwards, H. D. (1984). A judge's review of juror misconduct. *Howard Law Journal*, *27*, 1519-1547.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. doi: 10.1136/bmj.315.7109.629

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203-235. doi:10.1037/0033-2909.128.2.203

Ellison, L., & Munro, V. E. (2009). Reacting to rape: Exploring mock jurors' assessments of complainant credibility. *British Journal of Criminology, 49*(2), 202-219. doi:10.1093/bjc/azn077

Eyssel, F., & Bohner, G. (2011). Schema effects of rape myth acceptance on judgments of guilt and blame in rape cases: The role of perceived entitlement to judge. *Journal of Interpersonal Violence, 26*(8), 1579-1605. doi:10.1177/0886260510370593

Fahsing, I., & Ask, K. (2016). The making of an expert detective: the role of experience in English and Norwegian police officers’ investigative decision-making. *Psychology, Crime & Law*, *22*(3), 203–223. doi: 10.1080/1068316X.2015.1077249

Feldman-Summers, S., & Palmer, G. (1980). Rape as viewed by judges, prosecutors, and police officers. *Criminal Justice and Behavior*, *7*(1), 19–40. doi: https://doi.org/10.1177/009385488000700103

Filipas, H. H., & Ullman, S. E. (2001). Social reactions to sexual assault victims from various support sources. *Violence and Victims, 16*(6), 673-692.

Fischer, A. H., Kret, M. E., & Broekens, J. (2018). Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis. *PLoS ONE*, *13*(1). doi: 10.1371/journal.pone.0190712

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science, 8*(4), 370-378. doi:10.1177/1948550617693063

Flatley, J. (2018). *Sexual offences in England and Wales: Year ending March 2017*. London: Office for National Statistics: London. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/sexualoffencesinenglandandwales/yearendingmarch2017

Frohmann, L. (1991). Discrediting victims' allegations of sexual assault: Prosecutorial accounts of case rejections. *Social Problems, 38*(2), 213-226.

Frohmann, L. (1997). Convictability and discordant locales: Reproducing race, class, and gender ideologies in prosecutorial decision-making. *Law & Society Review, 31*, 531.

Giner-Sorolla, R. (2018, January 24). Powering your interaction [Blogpost]. Retrieved from https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Gitter, A. G., Kozel, N. J., & Mostofsky, D. I. (1972). Perception of emotion: The role of race, sex, and presentation mode. *The Journal of Social Psychology, 88*(2), 213-222.

Goodman-Delahunty, J., & Graham, K. (2011). The influence of victim intoxication and victim attire on police responses to sexual assault. *Journal of Investigative Psychology and Offender Profiling, 8*(1), 22-40. doi:10.1002/jip.127

Goodwin, C. J., & Goodwin, K. A. (2013). Methodological control in experimental research *Research in psychology: Methods and design* (7 ed., pp. 183-214). New Jersey: Wiley.

Gravelin, C. R., Biernat, M., & Bucher, C. E. (2019). Blaming the victim of acquaintance rape: Individual, situational, and sociocultural factors. *Frontiers in Psychology*, *9*, 2422. doi:10.3389/fpsyg.2018.02422

Gray, J. M., & Horvath, M. A. H. (2018). Rape myths in the criminal justice system. In E. Milne, K. Brennan, N. South & J. Turton (Eds.), *Women and the criminal justice system* (pp. 15-41).

Grice, H. P. (1975). Logic and conversation. In R. Stainton (Ed.), *Perspectives in the philosophy of language: A concise anthology* (pp. 267-287).

Grubb, A., & Harrower, J. (2008). Attribution of blame in cases of rape: An analysis of participant gender, type of rape and perceived similarity to the victim. *Aggression and Violent Behavior, 13*(5), 396-405. doi:10.1016/j.avb.2008.06.006

Grubb, A., & Turner, E. (2012). Attribution of blame in rape cases: A review of the impact of rape myth acceptance, gender role conformity and substance use on victim blaming. *Aggression and Violent Behavior, 17*(5), 443-452. doi:http://dx.doi.org/10.1016/j.avb.2012.06.002

*Hackett, L., Day, A., & Mohr, P. (2008). Expectancy violation and perceptions of rape victim credibility. *Legal and Criminological Psychology, 13*(2), 323-334. doi:10.1348/135532507X228458

Hans, V. P., & Vidmar, N. (1986). Mr. prejudice or miss sympathy: A thirteenth juror? In V. P. Hans & N. Vidmar, *Judging the Jury* (pp. 131–148). Boston, MA: Springer US. doi:10.1007/978-1-4899-6463-2_9

Haslam, S. A., & McGarty, C. (2008). Experimental design and causality in psychological research. In C. Sansone, C. Morf & A. Panter (Eds.), The SAGE handbook of methods in social psychology (pp. 237-264).

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology, 28*(5), 661-676. doi:10.1002/acp.3052

Heenan, M., & Murray, S. (2007). *Study of reported rapes in Victoria 2000-2003: Summary research report.* Melbourne: Office of Women's Policy, Department for Victorian Communities.

Higgins, J. P. T., & Altman, D. G. (2008). Assessing risk of bias in included studies. In J. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 187-241). Chichester, United Kingdom: John Wiley & Sons.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., . . . Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. *BMJ, 343*. doi:10.1136/bmj.d5928

Higgins, J., & Deeks, J. (2008). Selecting studies and collecting data. In J. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions*. Chichester, United Kingdom: John Wiley & Sons.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. doi:10.1002/sim.1186

Hockett, J. M., Smith, S. J., Klausing, C. D., & Saucier, D. A. (2016). Rape myth consistency and gender differences in perceiving rape victims. *Violence Against Women, 22*(2), 139-167. doi:10.1177/1077801215607359

Hohl, K., & Stanko, E. A. (2015). Complaints of rape and the criminal justice system: Fresh evidence on the attrition problem in England and Wales. *European Journal of Criminology, 12*(3), 324-341. doi:10.1177/1477370815571949

Jehle, J.-M. (2012). Attrition and conviction rates of sexual offences in Europe: Definitions and criminal justice responses. *European Journal on Criminal Policy and Research, 18*(1), 145-161. doi:10.1007/s10610-011-9163-x

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kahan, D. M. (2015). Laws of cognition and the cognition of law. *Cognition, 135*, 56-60. doi: 10.1016/j.cognition.2014.11.025

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515-526. doi:10.1037/a0016755

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

*Kaufmann, G., Drevland, G. C. B., Wessel, E., Overskeid, G., & Magnussen, S. (2003). The importance of being earnest: Displayed emotions and witness credibility. *Applied Cognitive Psychology, 17*(1), 21-34. doi:10.1002/acp.842

Kerr, N. L. (2017). Suggested do's and don'ts for future jury research: A swan song. In M. B. Kovera (Ed.), *The psychology of juries* (pp. 257-285). Washington DC: APA.

Kerr, N. L., & Bray, R. M. (2005). Simulation, realism, and the study of the jury. In N. Brewer & K. D. Williams (Eds.), *Psychology and law: An empirical perspective* (pp. 322-364). New York: Guilford Press.

Kerstetter, W. A. (1990). Gateway to justice: Police and prosecutorial response to sexual assaults against women. *The Journal of Criminal Law and Criminology, 81*(2), 267-313. doi:10.2307/1143908

Kite, M. E., & Whitley, B. E. (2018). The internal validity of research. *Principles of research in behavioral sciences* (4 ed., pp. 277-309). New York: Routledge.

Kleinke, C. L., & Meyer, C. (1990). Evaluation of rape victim by men and women with high and low belief in a just world. *Psychology of Women Quarterly, 14*(3), 343-353. doi:10.1111/j.1471-6402.1990.tb00024.x

*Klippenstine, M. A. (2010). *Perceptions of Sexual Assault: Expectations Regarding the Emotional Response of a Rape Victim* (Unpublished doctoral dissertation). York University, Toronto.

*Klippenstine, M. A., & Schuller, R. (2012). Perceptions of sexual assault: expectancies regarding the emotional response of a rape victim over time. *Psychology, Crime & Law, 18*(1), 79-94. doi:10.1080/1068316X.2011.589389

Koehler, J. J., & Meixner Jr, J. B. (2017). Jury simulation goals. *The psychology of juries.* (pp. 161-183). Washington, DC, US: American Psychological Association.

Konradi, A. (1999). “I don't have to be afraid of you”: Rape survivors' emotion management in court. *Symbolic Interaction, 22*(1), 45-77. doi:10.1016/S0195-6086(99)80003-4

Konradi, A. (2007). *Taking the stand: Rape survivors and the prosecution of rapists*. Westport: Greenwood Publishing Group.

Kotiaho, J. S., & Tomkins, J. L. (2002). Meta-analysis, can it ever fail? *Oikos, 96*(3), 551-553.

Krahé, B. (2016). Societal responses to sexual violence against women: Rape myths and the “real rape” stereotype. In H. Kury, S. Redo, & E. Shea (Eds.), *Women and children as victims and offenders: Background, prevention, reintegration* (pp. 671-700): Springer.

Krahé, B., Temkin, J., & Bieneck, S. (2007). Schema-driven information processing in judgements about rape. *Applied Cognitive Psychology, 21*(5), 601-619. doi:10.1002/acp.1297

Krahé, B., Temkin, J., Bieneck, S., & Berger, A. (2008). Prospective lawyers' rape stereotypes and schematic decision making about rape cases. *Psychology, Crime and Law, 14*(5), 461-479. doi: https://doi.org/10.1080/10683160801932380

Krauss, D. A., & Lieberman, J. D. (2017). Managing different aspects of validity in trial simulation research. In M. B. Kovera (Ed.), *The psychology of juries* (pp. 185-205). Washington: American Psychological Association.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, California: Sage.

Krippendorff, K. (2013). *Content analysis: an introduction to its methodology* (3rd ed.). London: SAGE.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*(1), 24-34. doi:10.1186/s40359-016-0126-3

Lee, J., Lee, C., & Lee, W. (2012). Attitudes toward women, rape myths, and rape perceptions among male police officers in South Korea. *Psychology of Women Quarterly, 36*(3), 365-376. doi:10.1177/0361684311427538

Lefebvre, C., Manheimer, E., & Glanville, J. (2008). Searching for studies. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 95-150). Chichester, United Kingdom: John Wiley & Sons

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275. https://doi.org/10.1037/0033-2909.125.2.255

Lidén, M., Gräns, M., & Juslin, P. (2018). The presumption of guilt in suspect interrogations: Apprehension as a trigger of confirmation bias and debiasing techniques. *Law and Human Behavior, 42*(4), 336-354. doi:http://dx.doi.org/10.1037/lhb0000287

Lievore, D. (2005). *Prosecutorial decisions in adult sexual assault cases: An Australian study*. Canberra: Office for the Status of Women.

Littleton, H. L., & Axsom, D. (2003). Rape and seduction scripts of university students: Implications for rape attributions and unacknowledged rape. *Sex Roles, 49*(9-10), 465-475. doi:10.1023/A:1025824505185

Littleton, H., Tabernik, H., Canales, E., & Backstrom, T. (2009). Risky situation or harmless fun? A qualitative examination of college women's bad hook-up and rape scripts. *Sex Roles, 60*(11-12), 793-804. Doi:10.1007/s11199-009-9586-8

Lonsway, K. A., & Fitzgerald, L. F. (1994). Rape myths: In review. *Psychology of Women Quarterly, 18*(2), 133-164. doi:10.1111/j.1471-6402.1994.tb00448.x

Lovakov, A., & Agadullina, E. R. (2017). *Empirically derived guidelines for interpreting effect size in social psychology*. Manuscript submitted for publication.

Lynch, K. R., Jewell, J. A., Wasarhaley, N. E., Golding, J. M., & Renzetti, C. M. (2017). Great sexpectations: The impact of participant gender, defendant desirability, and date cost on attributions of a date rape victim and defendant. *Journal of Interpersonal Violence*. Advance online publication. doi: 0886260517709800

Maddox, L., Lee, D., & Barker, C. (2011). Police Empathy and victim PTSD as potential factors in rape case attrition. *Journal of Police and Criminal Psychology, 26*(2), 112-117. doi:10.1007/s11896-010-9075-6

Maddox, L., Lee, D., & Barker, C. (2012). The impact of psychological consequences of rape on rape case attrition: The police perspective. *Journal of Police and Criminal Psychology, 27*(1), 33-44. doi:10.1007/s11896-011-9092-0

Maheswaran, D., Mackie, D. M., & Chaiken, S. (1992). Brand name as a heuristic cue: The effects of task importance and expectancy confirmation on consumer judgments. *Journal of Consumer Psychology*, *1*(4), 317–336. doi: 10.1016/S1057-7408(08)80058-7

Masser, B., Lee, K., & McKimmie, B. M. (2010). Bad woman, bad victim? Disentangling the effects of victim stereotypicality, gender stereotypicality and benevolent sexism on acquaintance rape victim blame. *Sex Roles, 62*(7-8), 494-504.

McKimmie, B. M., Masser, B. M., & Bongiorno, R. (2014). Looking shifty but telling the truth: The effect of witness demeanor on mock jurors' perceptions. *Psychiatry, Psychology and Law, 21*(2), 297-310. doi:10.1080/13218719.2013.815600

*McKimmie, B. M., Masser, B. M., Nitschke, F. T., Schuller, R. & Goodman-Delahunty, J. (2018b). [Effect of sexual script elements on judgments of rape cases and complainants]. Unpublished raw data.

McKimmie, B. M., Masser, B. M., Nitschke, F. T., Schuller, R. & Goodman-Delahunty, J. (2018a). *Cues to consent: The role of rape and consensual sex schemas on judgments in cases of alleged rape*. Manuscript in preparation.

McNutt, M. (2014). Reproducibility. *Science, 343*(6168), 229-229. doi:10.1126/science.1250475

Miller, A. L. (2018). Expertise fails to attenuate gendered biases in judicial decision-making. *Social Psychological and Personality Science*. Advance online publication. doi: https://doi.org/10.1177/1948550617741181

Morabito, M. S., Pattavina, A., & Williams, L. M. (2016). It all just piles up: Challenges to victim credibility accumulate to influence sexual assault case processing. *Journal of Interpersonal Violence*. Advance online publication. doi: 10.1177/0886260516669164

Nee, C., & Ward, T. (2015). Review of expertise and its general implications for correctional psychology and criminology. *Aggression and Violent Behavior, 20*, 1-9. doi: 10.1016/j.avb.2014.12.002

Neuendorf, K. A. (2002). *The content analysis guidebook*. California: Sage Publications.

Newcombe, P. A., Van Den Eynde, J., Hafner, D., & Jolly, L. (2008). Attributions of responsibility for rape: Differences across familiarity of situation, gender, and acceptance of rape myths. *Journal of Applied Social Psychology, 38*(7), 1736-1754. doi:10.1111/j.1559-1816.2008.00367.x

Nitschke, F. T., Masser, B. M., McKimmie, B. M., & Riachi, M. (2018). Intoxicated but not incapacitated: Are there effective methods to assist juries in interpreting evidence of voluntary complainant intoxication in cases of rape? *Journal of Interpersonal Violence.* Advance online publication. doi: 10.1177/0886260518790601

Norris, F. H., & Kaniasty, K. (1994). Psychological distress following criminal victimization in the general population: Cross-sectional, longitudinal, and prospective analyses. *Journal of Consulting and Clinical Psychology, 62*(1), 111-123. doi:10.1037/0022-006X.62.1.111

O'Neal, E. N. (2017). "Victim is not credible": The influence of rape culture on police perceptions of sexual assault complainants. *Justice Quarterly*. Advance online publication. doi:10.1080/07418825.2017.1406977

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716

Page, A. D. (2007). Behind the blue line: Investigating police officers' attitudes toward rape. *Journal of Police and Criminal Psychology, 22*(1), 22-32. doi:10.1007/s11896-007-9002-7

Paluck, E. L., Green, S., & Green, D. P. (2017). *The Contact Hypothesis Revisited*. Manuscript submitted for publication.

*Peace, K. A., & Valois, R. L. (2014). Trials and tribulations: Psychopathic traits, emotion, and decision-making in an ambiguous case of sexual assault. *Psychology, 5*(10), 1239.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*(25), 4544–4562. doi: 10.1002/sim.2889

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751-783. doi:10.1037/0022-3514.90.5.751

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123-205). New York: Academic Press.

Porter, S., & ten Brinke, L. (2009). Dangerous decisions: A theoretical framework for understanding how judges assess credibility in the courtroom. *Legal and Criminological Psychology, 14*(1), 119-134. doi:10.1348/135532508X281520

Quadara, A., Fileborn, B., & Parkinson, D. (2013). *The role of forensic medical evidence in the prosecution of adult sexual assault.* Canberra: Australian Institute of Criminology.

Queensland Courts. (2017). *Supreme and District Courts Criminal Directions Benchbook.* Brisbane, Queensland: Queensland Courts.

Rachlinski, J. J., & Wistrich, A. J. (2017). Judging the judiciary by the numbers: Empirical research on judges. *Annual Review of Law and Social Science*, *13*, 203-229.

Raganella, A. J., & White, M. D. (2004). Race, gender, and motivation for becoming a police officer: Implications for building a representative police department. *Journal of Criminal Justice*, *32*(6), 501–513. doi: 10.1016/j.jcrimjus.2004.08.009

Reeves, B. C., Deeks, J. J., Higgins, J. P. T., & Wells, G. A. (2008). Including non-randomized studies. In J. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 391-432). Chichester, United Kingdom: John Wiley & Sons.

Reinhard, M. A. (2010). Need for cognition and the process of lie detection. *Journal of Experimental Social Psychology*, *46*(6), 961-971. doi: 10.1016/j.jesp.2010.06.002

Reinhard, Marc-Andre, Scharmach, M., & Sporer, S. L. (2012). Situational familiarity, efficacy expectations, and the process of credibility attribution. *Basic and Applied Social Psychology*, *34*(2), 107–127. doi: 10.1080/01973533.2012.655992

Reinhard, M. A., & Sporer, S. L. (2008). Verbal and nonverbal behaviour as a basis for credibility attribution: The impact of task involvement and cognitive capacity. *Journal of Experimental Social Psychology*, *44*(3), 477-488. doi: 10.1016/j.jesp.2007.07.012

Reinhard, M. A., & Sporer, S. L. (2010). Content versus source cue information as a basis for credibility judgments: The impact of task involvement. *Social Psychology, 41*(2), 93-104. doi: 10.1027/1864-9335/a000014.

Reinhard, Marc-André, Sporer, S. L., Scharmach, M., & Marksteiner, T. (2011). Listening, not watching: Situational familiarity and the ability to detect deception. *Journal of Personality and Social Psychology*, *101*(3), 467–484. doi: 10.1037/a0023726

Reymond v Township of Bosanquet (1919) 59 CanSC 452.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464–468. doi: 10.1111/j.0014-3820.2005.tb01004.x

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi: 10.1037/0033-2909.86.3.638

Rotton, J., Foos, P.W., Van Meek, L. & Levitt, M. (1995). Publication practices and the file drawer problem: A survey of published authors. *Journal of Social Behavior and Personality, 10*(1), 1-13.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Schuller, R. A., & Hastings, P. A. (2002). Complainant sexual history evidence: Its impact on mock jurors' decisions. *Psychology of Women Quarterly, 26*(3), 252-261. doi:10.1111/1471-6402.00064

*Schuller, R. A., McKimmie, B. M., Masser, B. M., & Klippenstine, M. A. (2010). Judgments of sexual assault: The impact of complainant emotional demeanor, gender, and victim stereotypes. *New Criminal Law Review: An International and Interdisciplinary Journal, 13*(4), 759-780. doi:10.1525/nclr.2010.13.4.759

Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of “irrelevant” information. *Social Cognition*, *9*(1), 67–84. doi:10.1521/soco.1991.9.1.67

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. Germany: Springer.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, *81*(1), 75–86. doi: 10.1016/0001-6918(92)90012-3

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *P*-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547. doi: 10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). *P*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*(6), 666-681. doi: 10.1177/1745691614553988

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *p*-curves: Making *p*-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*(6), 1146-1152. doi: 10.1037/xge0000104

Sleath, E., & Bull, R. (2012). Comparing rape victim and perpetrator blaming in a police officer sample: Differences between police officers with and without special training. *Criminal Justice and Behavior, 39*(5), 646-665. doi:10.1177/0093854811434696

Sleath, E., & Bull, R. (2015). A brief report on rape myth acceptance: Differences between police officers, law students, and psychology students in the United Kingdom. *Violence and Victims, 30*(1), 136-147. doi: http://dx.doi.org/10.1891/0886-6708.VV-D-13-00035

Smith, S.G., Chen, J., Basile, K.C., Gilbert, L.K., Merrick, M.T., Patel, N., Walling, M., & Jain, A. (2017). *The National Intimate Partner and Sexual Violence Survey (NISVS): 2010-2012 State Report*. Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Smith, V. L. (1991). Prototypes in the courtroom: Lay representations of legal concepts. *Journal of Personality and Social Psychology, 61*(6), 857-872. doi:10.1037/0022-3514.61.6.857

Smith, V. L. (1993). When prior knowledge and law collide: Helping jurors use the law. *Law and Human Behavior, 17*(5), 507-536. doi:10.1007/BF01045071

Snipes, D. J., Calton, J. M., Green, B. A., Perrin, P. B., & Benotsch, E. G. (2017). Rape and posttraumatic stress disorder (PTSD): Examining the mediating role of explicit sex–power beliefs for men versus women. *Journal of Interpersonal Violence, 32*(16), 2453-2470. doi:10.1177/0886260515592618

Spohn, C., & Tellis, K. (2012). The criminal justice system's response to sexual violence. *Violence Against Women, 18*(2), 169-192. doi:10.1177/1077801212440020

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, *8*(5), 581–591. doi: 10.1177/1948550617693062

Stern Review. (2010). *A report by Baroness Vivien Stern CBE of an Independent Review into How Rape Complaints are Handled by Public Authorities in England and Wales*. Retrieved from http://webarchive.nationalarchives.gov.uk/20110608162919/http:/www.equalities.gov.uk/pdf/Stern_Review_acc_FINAL.pdf.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2006). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis* (pp. 73–98). Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/0470870168.ch5

Sterne, J. A. C., Egger, M., & Moher, D. (2011). Addressing reporting biases data. In J. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 297-333). Chichester, United Kingdom: John Wiley & Sons.

Strub, T., & McKimmie, B. (2012). Note takers who review are less vulnerable to the influence of stereotypes than note takers who do not review. *Psychology, Crime & Law*, *18*(10), 859-876. doi: https://doi.org/10.1080/1068316X.2011.581241

Suarez, E., & Gadalla, T. M. (2010). Stop blaming the victim: A meta-analysis on rape myths. *Journal of Interpersonal Violence, 25*(11), 2010-2035. doi:10.1177/0886260509354503

Süssenbach, P., Albrecht, S., & Bohner, G. (2017). Implicit judgments of rape cases: an experiment on the determinants and consequences of implicit evaluations in a rape case. *Psychology, Crime & Law, 23*(3), 291-304. doi:10.1080/1068316X.2016.1247160

Süssenbach, P., Bohner, G., & Eyssel, F. (2012). Schematic influences of rape myth acceptance on visual information processing: An eye-tracking approach. *Journal of Experimental Social Psychology, 48*, 660-686.

Süssenbach, P., Eyssel, F., & Bohner, G. (2013). Metacognitive aspects of rape myths: Subjective strength of rape myth acceptance moderates its effects on information processing and behavioral intentions. *Journal of Interpersonal Violence, 28*(11), 2250-2272. doi:10.1177/0886260512475317

Süssenbach, P., Eyssel, F., Rees, J., & Bohner, G. (2017). Looking for blame: Rape myth acceptance and attention to victim and perpetrator. *Journal of Interpersonal Violence, 32*(15), 2323-2344. doi:10.1177/0886260515591975

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 436-449). New York: Russell Sage Foundation.

Tasca, M., Rodriguez, N., Spohn, C., & Koss, M. P. (2013). Police decision making in sexual assault cases: Predictors of suspect identification and arrest. *Journal of Interpersonal Violence, 28*(6), 1157-1177. doi:10.1177/0886260512468233

Temkin, J. (2000). Prosecuting and defending rape: Perspectives from the bar. *Journal of Law and Society, 27*(2), 219-248. doi:10.1111/1467-6478.00152

Temkin, J., Gray, J. M., & Barrett, J. (2018). Different functions of rape myth use in court: Findings from a trial observation study. *Feminist Criminology, 13*(2), 205-226. doi:10.1177/1557085116661627

Temkin, J., & Krahé, B. (2008). *Sexual assault and the justice gap: A question of attitude*. Oxford: Hart Publishing.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113-2126. doi:10.1002/sim.1461

Thomas, C. (2010). *Are juries fair?* London: Ministry of Justice. Retrieved from http://www.ohrn.nhs.uk/resource/policy/arejuriesfair.pdf

Todorov, A., Chaiken, S., Henderson, M. D. (2002). The heuristic-systematic model of social information processing In J. P. Dillard & M. Pfau (Eds.), *The persuasion handbook: Developments in theory and practice* (pp. 195-212). Thousand Oaks: SAGE Publications.

Ullman, S. E., & Filipas, H. H. (2001). Predictors of PTSD symptom severity and social reactions in sexual assault victims. *Journal of Traumatic Stress, 14*(2), 369-389. doi:10.1023/A:1011125220522

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE, 9*(1), e84896. doi:10.1371/journal.pone.0084896

van der Bruggen, M., & Grubb, A. (2014). A review of the literature relating to rape victim blaming: An analysis of the impact of observer and victim characteristics on attribution of blame in rape cases. *Aggression and Violent Behavior, 19*(5), 523-531. doi:http://dx.doi.org/10.1016/j.avb.2014.07.008

Venema, R. M. (2013). Police officer decision making in reported sexual assault cases (Doctoral dissertation, University of Illinois). Retrieved from https://indigo.uic.edu/handle/10027/10227.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*(3), 261-293. doi:10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48. doi:10.18637/jss.v036.i03

Voogt, A., Klettke, B., & Crossman, A. (2016). Measurement of victim credibility in child sexual assault cases. *Trauma, Violence, & Abuse.* Advance online publication. doi:10.1177/1524838016683460

Vrij, A., & Fischer, A. (1997). The role of displays of emotions and ethnicity in judgments of rape victims. *International Review of Victimology, 4*(4), 255-265. doi:10.1177/026975809700400402

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). California: SAGE Publications.

Wentz, E., & Archbold, C. A. (2012). Police perceptions of sexual assault victims: Exploring the intra-female gender hostility thesis. *Police Quarterly, 15*(1), 25-44. doi:10.1177/1098611111432843

*Wessel, E., Drevland, G. B., Eilertsen, D. E., & Magnussen, S. (2006). Credibility of the emotional witness: A study of ratings by court judges. *Law and Human Behavior, 30*(2), 221-230. doi:10.1007/s10979-006-9024-1

*Winkel, F. W., & Koppelaar, L. (1991). Rape victims' style of self-presentation and secondary victimization by the environment. *Journal of Interpersonal Violence, 6*(1), 29-40. doi:10.1177/088626091006001003

White, M. D., Cooper, J. A., Saunders, J., & Raganella, A. J. (2010). Motivations for becoming a police officer: Re-assessing officer attitudes and job satisfaction after six years on the street. *Journal of Criminal Justice*, *38*(4), 520–530. https://doi.org/10.1016/j.jcrimjus.2010.04.022

White v R [1947] SCR 268.

Wiener, R. L., Krauss, D. A., & Lieberman, J. D. (2011). Mock jury research: Where do we go from here? *Behavioral Sciences & the Law, 29*(3), 467-479. doi:10.1002/bsl.989

Wrede, O., & Ask, K. (2015). More than a feeling: Public expectations about emotional responses to criminal victimization. *Violence and Victims, 30*(5), 902-915. doi: 10.1891/0886-6708.VV-D-14-00002.

Zajac, R., & Cannan, P. (2009). Cross-examination of sexual assault complainants: A developmental comparison. *Psychiatry, Psychology and Law, 16*(1), S36-S54. doi:10.1080/13218710802620448

Zydervelt, S., Zajac, R., Kaladelfos, A., & Westera, N. (2016). Lawyers' strategies for cross-examining rape complainants: Have we moved beyond the 1950s? *British Journal of Criminology, 57*(3), 551-569. doi:10.1093/bjc/azw023

Table 1

Disclosure Table for P-curve Analysis

Original Article	Quoted text from original article indicating prediction of interest to researchers	Study design	Key statistical result	Quoted text from original article with statistical results	Results	Robustness results
Ask & Landström (2010)	First, in line with previous demonstrations, we predicted that a rape victim who behaves in an emotional manner would be believed more readily than a victim who displays little emotion (Hypothesis 1).	2 (cognitive load: yes vs. no) x 2 (target demeanor: emotional vs. neutral) factorial design	Difference between means	A 2 (target demeanor: emotional vs. neutral) x 2 (cognitive load: yes vs. no) analysis of variance (ANOVA) revealed a significant main effect of target demeanor, *F*(1, 185) = 11.45, *p* <.001, partial eta2 = .06. Thus, in line with Hypothesis 1, participants who watched the emotional demeanor were more certain (*M* = 64.42, *SD* = 22.01) that the target person had been raped than did those who watched the neutral demeanor (*M* = 54.36, *SD* = 19.49).	*F*(1, 185) = 11.45, *p* = .00087	
Baldry (1996)	On the basis of what we have said, we then hypothesized: an alleged rape victim reporting to the police in an emotional	2 (emotional vs controlled woman) x 2 (powerful vs powerless speech	Difference between means	Our first hypothesis that a woman reporting to the police that she had been raped is believed more when she uses an	*F*(1, 122) = 4.89, *p* = .02888	

	self-style is going to be believed more readily than someone telling the same story in a controlled way.	style of the man) x 2 (low vs high empathy) factorial design		emotional style of communication, was confirmed (*F*(1, 122) = 4.89, *p* < .05.		
Baldry, Winkel, & Enthoven (1997)	A victim reporting a rape in an emotional way is believed more and held less responsible than a victim reporting the same crime in a controlled way	2 (emotional vs controlled victim) x 2 (powerful vs powerless perpetrator speech style)	Difference between means	This analysis resulted in a main effect - supporting hypothesis 1 - for victim's style of self-presentation (see Table 2). Table 2 suggests that emotional victims were perceived as more credible. Further support for hypothesis 1 comes from an analysis of covariance on groups 5 and 6, here too the perceived credibility of the emotional victim (*M* = 4.99) is significantly higher (*F* (1,19) = 17.08, *p* < .01) than the credibility of a victim exhibiting a controlled style.	*F*(1, 19) = 17.08, *p* = .00057	Alternate dependent variable (negative evaluation of complainant; responsibility for assault) *F*(1, 17) = .03, *p* =.86454
Baldry & Winkel (1998)	Main hypotheses tested were that an emotionally expressive style (unjustly) enhances the credibility of the victim, and the evidential value of her statements, while a powerless speech style (unjustly) reduces the	2 (presentation style of the victim: emotional vs controlled) x 2 (speech style of suspect: powerful vs powerless) x 2 (nationality of police officer:	Difference between means	Emotional victims made a much more favorable impression than emotionally controlled victims. Emotional victims generally made a more credible impression (*F* (1,196)=6.71, *p* <.01)	*F*(1, 196) = 6.71, *p* = .01031	Alternate dependent variable (evidential power of her statement). No significant effect of emotional demeanor so no

	credibility and evidential value of the suspect's statements. Moreover, the interaction effects due to the police officers' nationality were explored.	Dutch vs Italian) factorial design				statistics reported in paper.
Calhoun, Cann, Selby, & Magee (1981; Study 1)	The purpose of the present investigation was to examine the effects of the victim’s emotional style on social reactions to the rape victim. Of specific interest were the perceptions of the victim’s credibility, the degree to which she would be socially accepted, and the degree to which observers believed the victim found the rape unpleasant.	2 cell design (victim emotion: distressed vs calm)	Difference between means	When the victim was described as expressive she was viewed as significantly more credible (*M* = 17.36), than when she was described as calm (*M* = 13.96), *F* = 10.54, d.f. = 1, 45, *p* < .002.	*F*(1, 45) = 10.54, *p* = .00221	Alternate dependent variable (negative evaluation of complainant; victim’s perceived enjoyment of assault) *F*(1, 45) = 3.04, *p* = .08806 No significant effect of demeanor on social acceptance and no statistics reported in paper.
Calhoun, Cann, Selby, & Magee (1981; Study 2)	The purpose of the present investigation was to examine the effects of the victim’s emotional style on social reactions to the rape victim. Of specific	2 cell design (victim emotion: distressed vs calm)	Difference between means	There also was a significant impact on the perceived credibility of the victim. When the victim was controlled, she was rated as less credible (*M* =	*F*(5, 47) = 7.31, *p* = .00004[b]	Alternate dependent variable (negative evaluation of complainant;

	interest were the perceptions of the victim's credibility, the degree to which she would be socially accepted, and the degree to which observers believed the victim found the rape unpleasant.			14.21) than when she was expressed (M = 16.69). *F* = 7.31, d.f. = 5, 47, *p* < .009.		victim's perceived enjoyment of assault) *F*(5, 47) = 7.40, p = .00003 Alternate dependent variable (positive attitudes towards complainant; social acceptance) *F*(5, 47) = 3.71, *p* = .00652
Hackett, Day & Mohr (2008)	Based on the preceding discussion, it is therefore hypothesized that people who expect rape victims to behave in an emotionally expressive way will find an emotionally expressive victim as more credible than a victim who is not emotionally expressive. For people who do not expect rape victims to behave in an emotionally expressive way, it is hypothesized that there	2 (victim emotion: emotional vs unemotional) x 2 (participant expectation for emotion: expects emotion vs does not expect emotion)	2 x 2 interaction	To test the hypothesis, a 2 (expectation of typical behaviour vs. no expectation of typical behaviour) by 2 (victim emotionally expressive vs. victim not emotionally expressive) ANOVA was conducted. A significant two-way interaction with a moderate effect size was noted, (*F*(1, 133) = 5.43, *p* <.05, partial eta squared = .04).	*F*(1, 133) = 5.43, *p* = .02130	

	will be no difference in credibility judgements between an emotionally expressive victim and a victim who is not emotionally expressive.					
Klippenstine & Schuller (2012)	While past research has not examined the impact of the victim's emotions at multiple points in time, it was expected that a complainant who responded consistently over time, regardless if that response was tearful/upset or calm/controlled, would be more likely to be believed than a complainant who responded inconsistently.	2 (victim emotion day after event: upset vs calm) x 2 (victim emotion at trial: upset vs calm)	Both simple effects	Simple main effects revealed that when the complainant was tearful/upset the day following the assault, participants were more likely to believe her claim of rape when she was also tearful/upset during her testimony compared to when she was calm/controlled at trial. Specifically the complainant's claim of non-consent was more believable, F(1,120) = 4.00, p =0.048, eta squared = 0.07,	*F*(1, 120) = 4.00, *p* = .04776[a]	
Winkel & Koppelaar (1991)	We hypothesise that rape victims whose self-presentation is emotional run less risk of secondary victimisation by the environment than do those victims whose self-presentation is numbed.	2 (victim emotional state: numbed vs controlled) x 2 (participant ethnicity: Dutch vs Turkish)	Difference between means	The numbed victim was less often perceived as credible and more often blamed than the emotional victim. (Statistics provided in Table 2)	*F*(1, 76) = 14.06, *p* = .00034	Alternate dependent variable (negative evaluation of complainant; victimizing reactions)

	$F(1, 76) = 13.74$, $p = .00040$

Note. [a]In Klippenstine & Schuller (2012), the second simple effect was non-significant and not reported in the original article. This does not impact the reliability of the *p*-curve as non-significant test statistics are excluded from *p*-curve analysis. [b] Calhoun et al. (1981) report the degrees of freedom for this analysis (a univariate ANOVA) as (5, 47) however, these degrees of freedom do not match the reported analysis for the design. If the degrees of freedom are adjusted to match the description of the study design and statistical test reported (i.e., 1, 47), and are included in the *p*-curve and robustness *p*-curves, the results of these analyses are unchanged and all *p*-curves remain significantly right skewed.

Table 2

Adapted risk of bias criteria to assess internal validity in cross-sectional experimental psychology studies

Type of bias	Cochrane bias definition[a]	Cochrane criteria for randomized controlled trials	Adapted criteria for cross-sectional experimental psychology study
Selection bias	Bias present if there are systematic differences between participants allocated to each group in trial	What is the procedure for sequence generation and allocation concealment in the trial?	Were participants randomly allocated to experiment conditions?
Performance bias	Bias present if there is a problem with the intervention fidelity	Whether a procedure is used to appropriately blind participants, intervention personnel and outcome assessors to trial condition? Whether other steps are taken to protect internal validity?	Was a manipulation check used? Was the manipulation effective? Was a suspicion probe used to assess realism of the simulated testimony?
Detection bias	Bias present if there are problems in the assessment of outcomes	Was the outcome accurately assessed? Were outcome assessors blind to trial condition?	Was the reliability of the dependent variable measured reported? Was the validity of the dependent variable measured reported?

Note. [a]Cochrane bias definitions and criteria are drawn from Higgins & Altman (2008) and Reeves et al. (2008).

Table 3

Sample size, sample type and stimulus modality for each study included in the meta-analysis

Paper (reference)	Sample size	Sample type	Stimulus modality
Ask (2018)*	362	No experience (Students)	Video
Ask & Landström (2010)	189	Experienced (Police trainees)	Video
Baldry (1996)	130	Experienced (Police officers)	Video
Baldry, Winkel & Enthoven (1997)	98	Experienced (Police officers)	Video
Baldry & Winkel (1998)	205	Experienced (Police officers)	Video
Bohner & Schapansky (2018)*	120	Experienced (Law students)	Video
Bollingmo, Wessel, Eilertsen & Magnussen (2007)	69	Experienced (Police officers)	Video
Bollingmo, Wessel, Sandvold, Eilertsen & Magnussen (2009)	334	No experience (Students)	Video
Calhoun, Cann, Selby & Magee (1981, Study 1)	49	No experience (Students)	Text
Calhoun, Cann, Selby & Magee (1981, Study 2)	55	No experience (Students)	Video
Dahl, Enemo, Drevland, Wessel, Eilertsen & Magnussen (2007)	174	No experience (Students)	Video
Hackett, Day & Mohr (2008)	137	No experience (Students)	Video
Kaufmann, Drevland, Wessel, Overskeid, & Magnussen (2003)	169	No experience (Students)	Video
Klippenstine (2010, Study 2)*	85	No experience (Students)	Text
Klippenstine & Schuller (2012)	124	No experience (Students)	Text
McKimmie, Masser, Nitschke, Schuller & Goodman-Delahunty (2018)*	100	No experience (Community members)	Text
Peace & Valois (2014)	383	No experience (Students)	Text
Schuller, McKimmie, Masser & Klippenstine (2010)	210	No experience (Students)	Text

Wessel, Drevland, Eilertsen & Magnussen (2006)	53	Experienced (Judges)	Video
Winkel & Koppelaar (1991)	80	No experience (Students)	Video

Note. * denotes unpublished works. Ask (2018) and Bohner & Schapansky (2018) were unpublished at the time of data analysis and subsequently were accepted for publication.

Table 4

Number and proportion of studies to meet risk of bias assessment criteria

Type of bias	Adapted criteria	κ	% of all studies to meet criteria
Selection bias	Participants randomly allocated to conditions	1.00	55.0 (11)
Performance bias	Reported manipulation check	.90	60.0 (12)
	Manipulation check successful	.80	55.0 (11)
	Reported suspicion probe	1.00	25.0 (5)
Detection bias	Single item measure	1.00	25.0 (5)
	Reported reliability analysis	1.00	55.0 (11)
	Reported validity analysis	-	0.0 (0)

Note. All studies were included in this analysis (*N* = 20). Number of studies to meet risk of bias criteria is reported in parentheses.

Table 5

Behaviors used to operationalize complainant distress and controlled affect in all studies

Complainant emotional state	Complainant behavior in stimulus	κ	% of all studies which use behavior	% of video stimulus studies which use behavior	% of text stimulus studies which use behavior
Distressed	Crying	1.00	94.4 (17)	91.7 (11)	100.0 (6)
	Trembling voice	1.00	55.6 (10)	58.3 (7)	50.0 (3)
	Hesitations in speech	1.00	55.6 (10)	66.7 (8)	33.3 (2)
	Distressed facial expression	0.89	44.4 (8)	58.3 (7)	16.7 (1)
	Struggling to maintain control	1.00	44.4 (8)	58.3 (7)	16.7 (1)
	States or another states victim was distressed	1.00	16.7 (3)	0.00 (0)	50.0 (3)
	Avoiding eye contact	1.00	11.1 (2)	8.33 (1)	16.7 (1)
	Speaking quietly	1.00	11.1 (2)	8.33 (1)	16.7 (1)
	Shock	1.00	5.56 (1)	8.33 (1)	0.0 (0)
Controlled	Factual manner	0.89	55.6 (10)	83.3 (10)	0.0 (0)
	Steady voice	1.00	27.8 (5)	25.0 (3)	33.3 (2)
	Confident	0.86	27.8 (5)	33.3 (4)	16.7 (1)
	Maintained eye contact	1.00	5.6 (1)	8.3 (1)	0.0 (0)
	Hushed voice	1.00	5.6 (1)	8.3 (1)	0.0 (0)

Note. Two studies were excluded from this analysis as no description of the behaviors used to operationalize complainant distress or controlled affect were provided ($N = 18$, $n_{video} = 12$, $n_{text} = 6$). Number of studies reported in parentheses.

Table 6

Item content for complainant credibility measures in all studies

Item type	Credibility measure contained an item about	κ	% of all studies with item
Face valid items	Believability	1.00	50.0 (10)
	Credibility	1.00	45.0 (9)
	Truthfulness	1.00	30.0 (6)
	Hiding truth	1.00	10.0 (2)
	Honesty	1.00	5.00 (1)
	Accuracy	1.00	5.00 (1)
Other constructs	Perpetrator blame or punishment or guilt	1.00	20.0 (4)
	Complainant distress	1.00	10.0 (2)
	Decision confidence	1.00	10.0 (2)

Note. All studies were included in this analysis (*N* = 20). Number of studies with item reported in parentheses.

Figure 1. Flow chart of the study selection process for the meta-analysis and *p*-curve analysis.

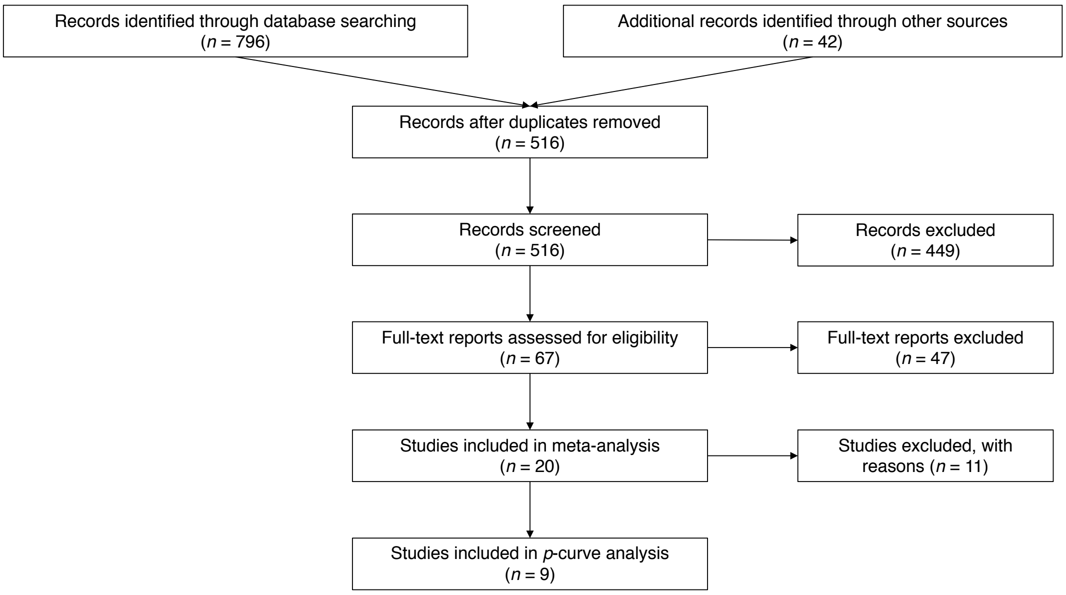

Records identified through database searching
(*n* = 796)
Additional records identified through other sources
(*n* = 42)
Records after duplicates removed
(*n* = 516)
Records screened
(*n* = 516)
Records excluded
(*n* = 449)
Full-text reports assessed for eligibility
(*n* = 67)
Full-text reports excluded
(*n* = 47)
Studies included in meta-analysis
(*n* = 20)
Studies excluded, with reasons (*n* = 11)
Studies included in *p*-curve analysis
(*n* = 9)

Figure 2. Forrest plot with effect size estimate and confidence intervals for each study.

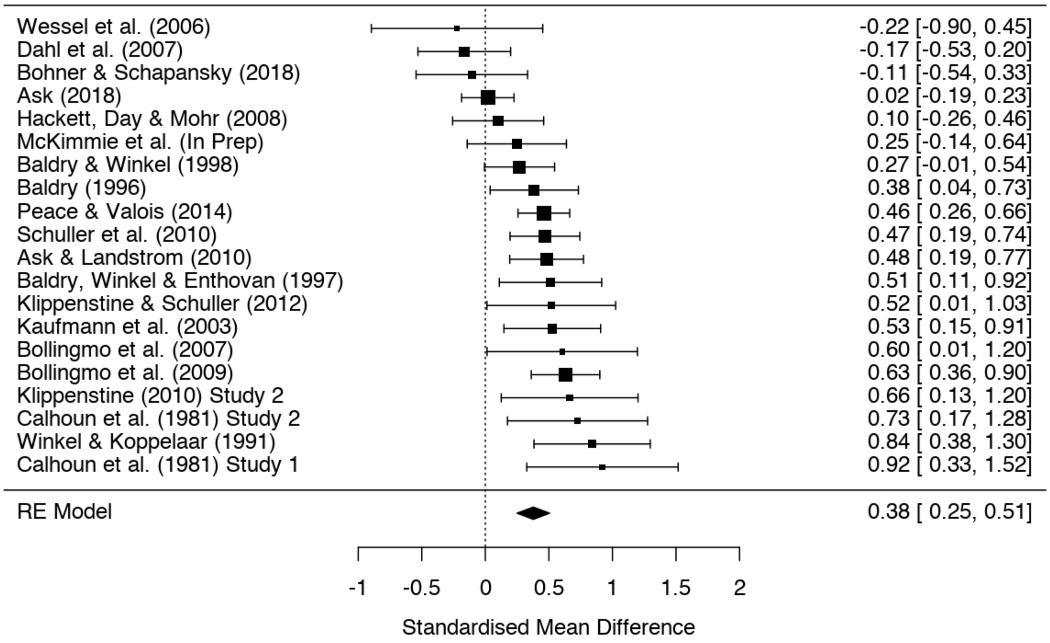
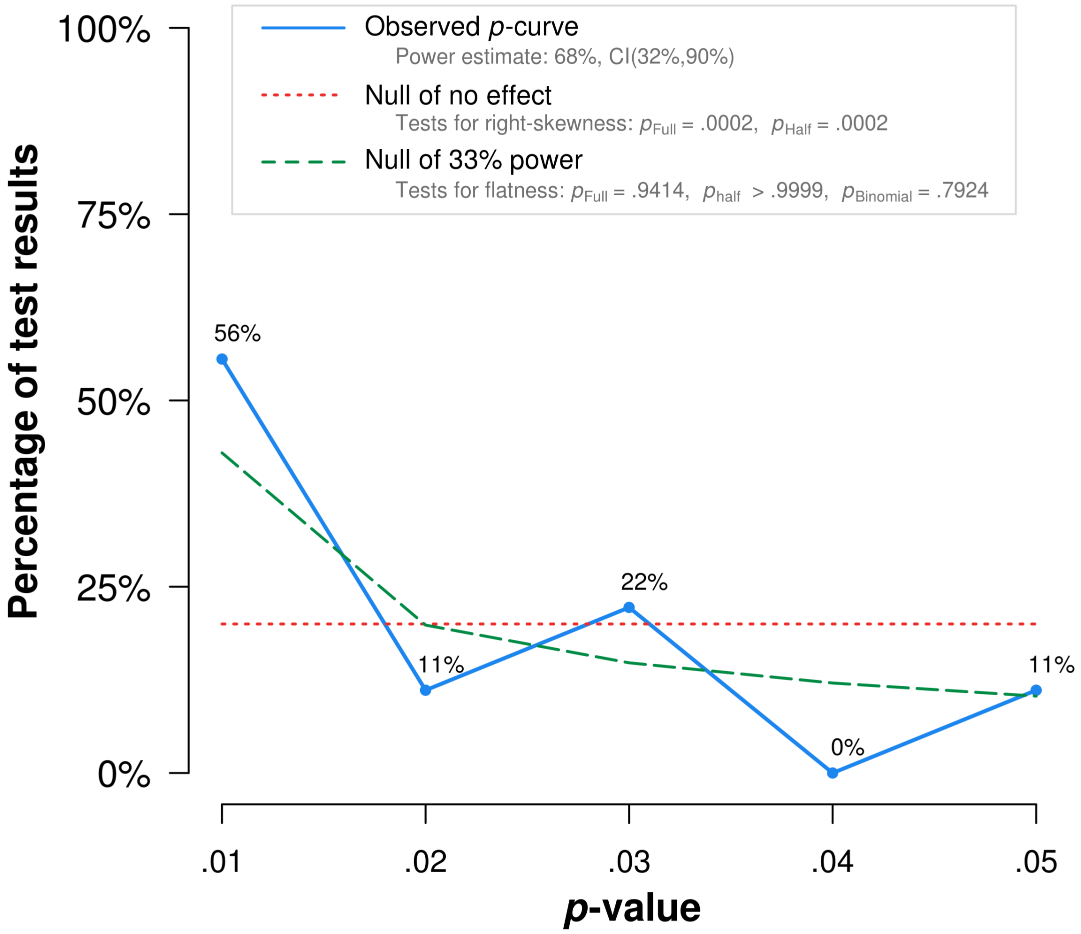


Figure 3. Plot of the distribution of *p*-values under .05 for published studies included in the *p*-curve analysis.



Note: The observed *p*-curve includes 9 statistically significant (p < .05) results, of which 7 are p < .025. There were no non-significant results entered.

Figure 4. Funnel plot of effect size estimate and standard error for all studies included in the meta-analysis.

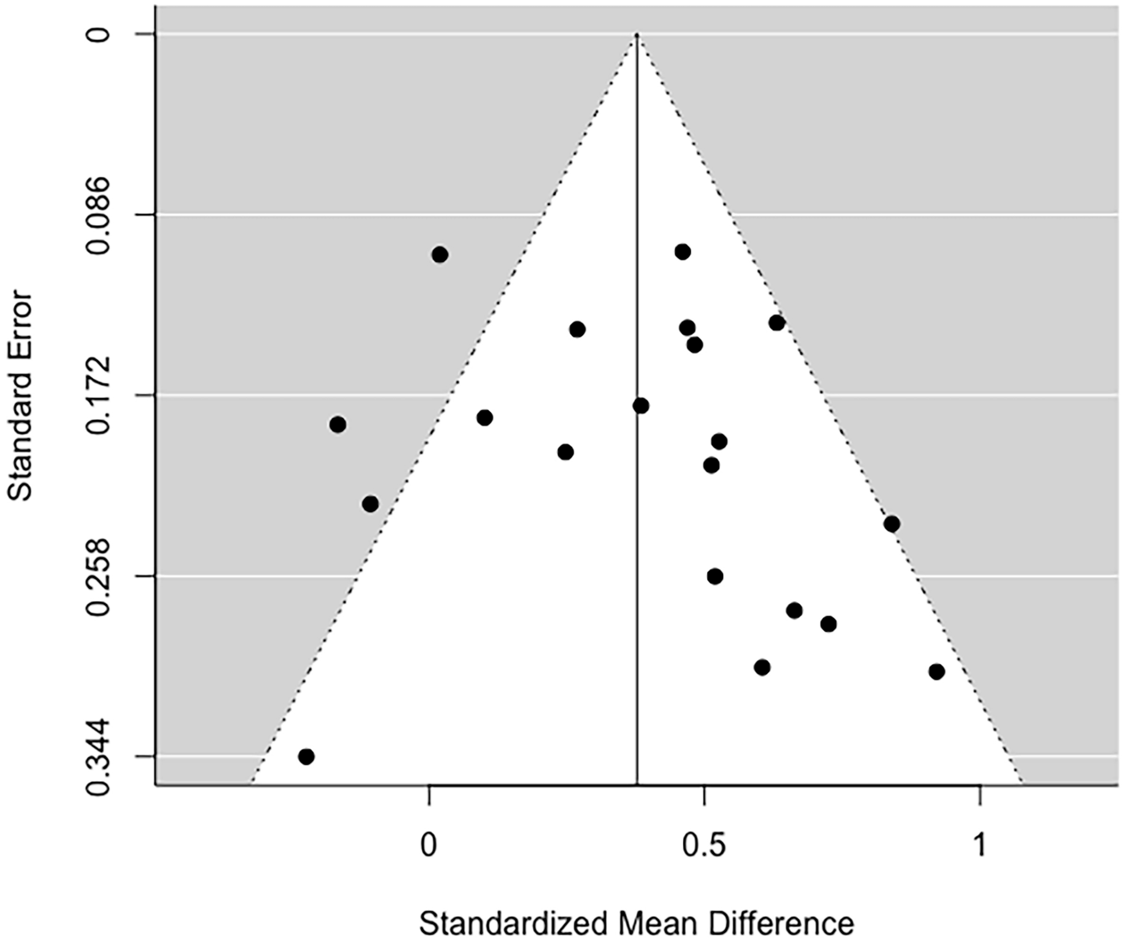
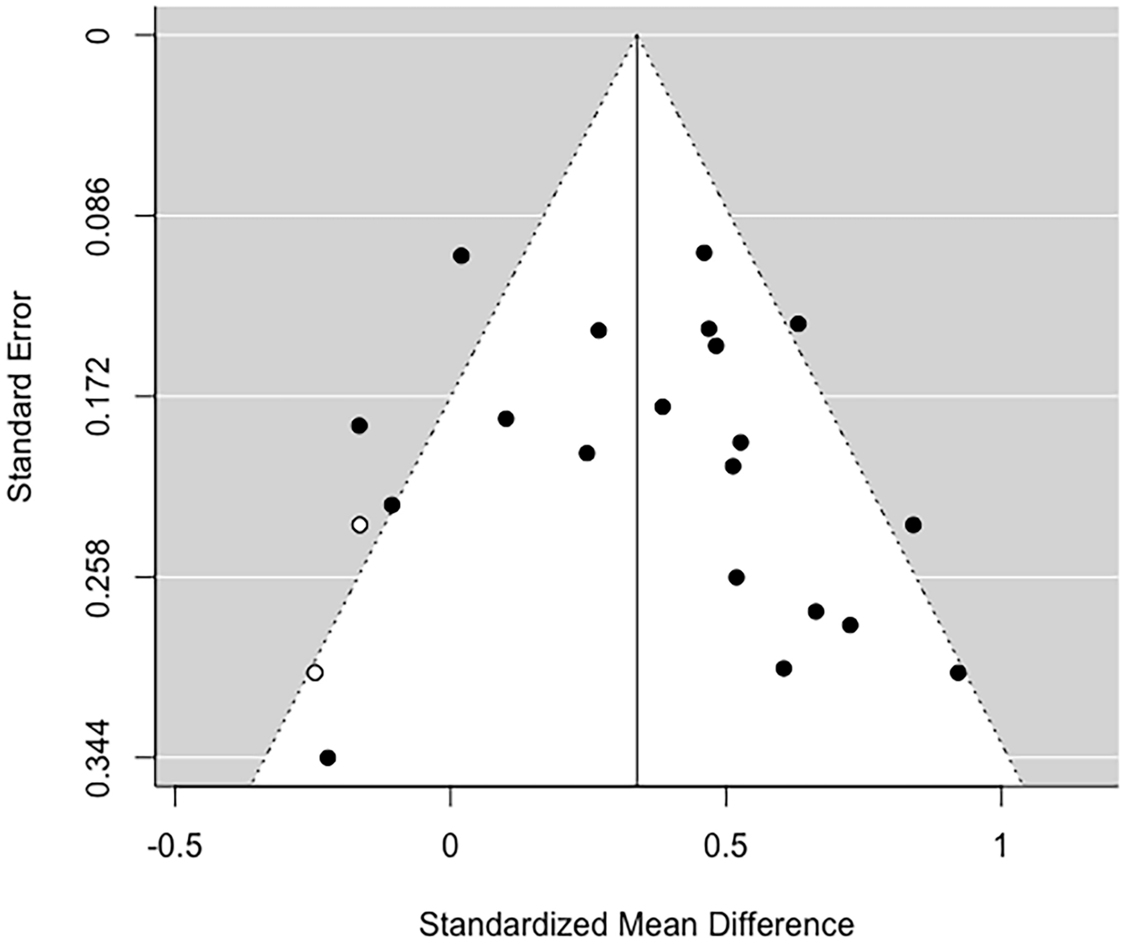


Figure 5. Funnel plot of effect size estimate and standard error for studies included in the meta-analysis and imputed through trim and fill analysis. White circles indicate imputed studies.


Standard Error
Standardized Mean Difference