Queen Margaret University
EDINBURGH

I AM PRETTY SURE BUT NOT 100%: OBTAINING, INTERPRETING AND PRESENTING EYEWITNESS CONFIDENCE STATEMENTS

PIA PENNEKAMP

Thesis submitted in partial fulfilment of the degree of Doctor of Philosophy

QUEEN MARGARET UNIVERSITY

2022

Acknowledgements

I would like to thank my supervisor, Dr. Jamal Mansour, for her unwavering support, encouragement, and continued guidance throughout the years. Words cannot express how grateful I am to have you as a mentor and role model. Thank you to my supervisory team, Dr. Stephen Darling and Dr. Stuart Wilson, for all their feedback and assistance. Thank you to Rhiannon Batstone for sharing her knowledge and supporting me in times when I needed it the most. Lastly, thank you to my family: My husband Ben, my dad, and my two bulldogs Honey and Bear for their great patience and unconditional love. Without you, this project would not have been possible.

Table of Contents

Dissemination

Pennekamp, P., & Mansour, J. K. (May 23, 2022). *Confidence Lexicon: An evidence-based approach for interpreting eyewitness confidence* [Poster presentation]. Scottish International Policing Conference, Edinburgh, United Kingdom.

Pennekamp, P., & Mansour, J. K. (2022, April 28-30). *Confidence Lexicon: An evidence-based approach for interpreting eyewitness confidence* [Conference presentation]. NOWCAM conference, Vancouver, BC, Canada.

Pennekamp, P., & Mansour, J. K. (2022, March 17-19). *Confidence Lexicon: An evidence-based approach for interpreting eyewitness confidence* [Conference presentation]. American Psychology-Law Society Conference, Denver, CO, United States.

Pennekamp, P., & Mansour, J. K. (2021, August 24). *Does order matter? Numeric and verbal eyewitness confidence* [Poster presentation]. European Association for Psychology and Law Conference (Virtual Edition). United Kingdom.

Pennekamp, P., & Mansour, J. K. (2021, March 21). *In your own words: Variability in verbal eyewitness confidence* [Poster presentation]. American Psychology-Law Society Conference (Virtual Edition). United States.

Mansour, J. K., **Pennekamp, P.**, & Batstone, R. J. (2020, November 17). *"I'm PRETTY SURE" vs "I'm pretty sure?" Using verbal confidence to evaluate the reliability of eyewitness identifications* [Conference presentation]. Face Science Symposium (Virtual Edition). South Africa.

Pennekamp, P., Batstone, R. J., & Mansour, J. K. (2019, December 9-10). *Eyewitness Identification Confidence: Requesting, articulating, and apperceiving* [Poster presentation]. Scottish Institute for Policing Research Conference, Edinburgh, U.K.

Abstract

Triers of fact must interpret the intended level of confidence expressed by eyewitnesses to judge the accuracy of identifications. There has been limited research on how to best obtain and interpret confidence judgements. Eyewitness identification confidence is typically studied using scales (generally numeric); in practice, eyewitnesses typically provide confidence in their own words. Verbal and numeric confidence similarly predict accuracy, but verbal confidence is difficult to interpret reliably (Mansour, 2020). To minimize miscommunication, eyewitnesses could provide scale ratings after verbal judgements or vice versa, but we do not know if the order in which such confidence statements are obtained affects the confidence-accuracy relationship. We (i.e. myself and supervisor) tested the utility of requesting both verbal and numeric confidence and whether order effects exist. Participants (*N* = 198) viewed a mock-crime video with two perpetrators. After a delay, they viewed two simultaneous lineups with one perpetrator each and provided confidence for each perpetrator verbally (in their own words) and then numerically (0-100%) or numerically and then verbally. Numeric confidence in identifications was higher when provided first, *t*(393.82) = 2.40, *p* = .02, *d* = 0.24. Confidence-accuracy characteristic (CAC) curve analysis indicates the effect is driven by medium-confidence judgements (numeric range). No order effect was found for verbal confidence (*p* = .32). However, for low and high numeric confidence, verbal followed by numeric was better calibrated than numeric followed by verbal. When the numeric judgement came first, none of the subsequent verbal judgements could be categorized as high confidence using our coding scheme. These data provide preliminary evidence that eyewitnesses should provide only a single confidence judgement. Given that verbal confidence statements are commonly used in practice and generally preferred, we aimed to improve the interpretation of verbal confidence statements to minimize miscommunication. In two studies, participants rated how well percentages (0%, 10%...100%) represented each of 13 common verbal confidence statements (e.g., moderately confident) on a scale (0 = *Not at all* to 100 = *Absolutely*). From the numeric distributions (membership functions) derived from each phrase's ratings, we identified four phrases with clear boundaries that together spanned the entirety of the 0-100% confidence scale. We developed a lexicon (i.e. translation tool) of four phrases and their ranges (including three synonyms). Understandings of verbal confidence statements are shared and quantifiable, facilitating common ground for reporting and interpreting eyewitness identification confidence.

To validate the lexicon, we tested 1) the replicability of the rank order, and 2) the (dis)similarity between the 13 phrases. Participants rank ordered phrases from the lowest to the highest level of confidence expressed. Interpretations were stable for low (not very confident; not sure) and high confidence phrases (very confident; confident). People have stable rank orders for some medium confidence phrases (such as quite confident; fairly confident; moderately confident), but not for others (e.g., he/she looks like the criminal). To test (dis)similarity between phrases, participants rated the (dis)similarity between the 13 phrases on a visual scale. Similarity was highest (>75%) for one low-confidence pairing (Not very confident/ Not sure), two medium-confidence pairings (Pretty sure/ Fairly confident; Quite confident/ Fairly confident) and one high-confidence pairing (Very confident/ Confident). We conclude that people consistently interpret verbal confidence phrases representing low and high confidence, but only some phrases representing medium confidence. Our research provides common ground for eyewitnesses and triers of fact when asked to provide and interpret verbal statements of confidence.

Keywords: confidence, eyewitness identification, lexicon, probabilities

Literature review

Eyewitness confidence

> One night in 1984, a stranger broke into Jennifer Thompson-Cannino's apartment and raped her. After the assault, Thompson-Cannino, then a 22-year-old college student, helped police sketch artists create a composite picture of her attacker. Later, in a photo lineup, she identified Ronald Cotton—a 22-year-old man who looked strikingly like her sketch and had previous run-ins with the law. Then, she picked Cotton from a live lineup. Cotton was convicted of rape and sentenced to life in prison. When Thompson-Cannino was first shown photos of possible suspects, she spent several minutes deliberating between two candidates. When she finally chose Cotton, she stated, "*I think that is him*" (Weir, 2016, p. 40).

By the time the case went in front of the courts, Jennifer Thompson-Cannino was "*absolutely sure*" that she had identified the perpetrator. But Jennifer was mistaken in her identification. A decade later, Ronald Cotton was exonerated by DNA evidence. Devastatingly, the case of Ronald Cotton is not an anomaly. Eyewitness misidentification has played a role in more than 70 percent of wrongfully convicted individuals (later exonerated by DNA evidence; Innocence project, 2022). Had investigators taken Jennifer Thompson-Cannino's initial statement of her uncertainty, namely "I think that is him", and accurately interpreted it, Ronald Cotton may have never been wrongfully convicted.

Eyewitnesses are often asked to make an identification from a lineup containing a suspect among fillers (i.e. people that are not suspected of having committed a crime). An identification of a suspect can be a decisive factor in investigations as eyewitness evidence provides a direct indication of guilt. When an eyewitness identifies an individual as the perpetrator, they provide evidence that the person is guilty. Eyewitnesses accounts, especially when given with high confidence, are often more persuasive than any other type of evidence (Semmler, Brewer & Douglass, 2011; Devlin, 1976; Boyce, Beaudry, & Lindsay, 2007). Memory, however, is reconstructive, easily influenced by external sources and, above all, prone to error (Loftus, 1981, Wells et al., 1998). In legal settings, this fallibility of memory has detrimental consequences. Mistaken eyewitness identifications are a leading factor in wrongful convictions (Innocence Project, 2022). Nevertheless, due to the nature of the criminal justice system, triers of fact continue, and will continue, to rely on eyewitness evidence.

Accompanying identifications, eyewitnesses often provide additional information when deciding on a lineup (e.g., "I think that is the guy"). Identification confidence, or the eyewitness' confidence in the lineup decision, has garnered attention from researchers in the last 40 years. Eyewitness confidence is often relied on in court settings to determine the reliability (i.e. accuracy) of eyewitness evidence (Cutler, Penrod, & Stuve, 1988). For example, high confidence eyewitness testimonies lead to a higher rate of convictions than low confidence eyewitness testimonies (Cutler & Penrod, 1995). Even when witnessing conditions differ, jurors rely heavily on confidence statements (Key et al., 2022; Slane & Dodson, 2022). Importantly, laypersons and even members of the legal profession consider confidence to imply accuracy (e.g., Brigham & Bothwell, 1983; Deffenbacher & Loftus, 1982). After all, "people *should* be a good judge of what they know and what they do not know" (Shaw et al., 2007, p. 371). But like memory, eyewitness confidence is malleable and subject to suggestion. There are many factors that can influence eyewitness confidence and once distorted, the weight jurors give it can become problematic. Nevertheless, eyewitness evidence continues to be a crucial source of evidence, especially in criminal investigations (e.g., Kebbell & Milne, 1998). Given that eyewitness confidence can provide valuable information about an eyewitness' lineup decision and, under certain conditions, does predict accuracy (Jusslin et al., 1996; Wixted & Wells, 2017), we need reliable ways to obtain, interpret and present eyewitness evidence.

Confidence and accuracy

Confidence has been studied as a predictor of accuracy since the 1980s. Early work found a weak to moderate correlation between eyewitness confidence and eyewitness accuracy (Leippe & Eisenstadt, 2007; Sporer, Penrod, Read, & Cutler, 1995). Correlation analyses "correlate a binary outcome (correct vs. incorrect) with a confidence rating across different participants, effectively averaging across all levels of confidence" (e.g., Brewin, Andrews, & Mickes, 2020, p. 122). Correlation analyses demonstrate that the relationship between confidence and accuracy appears to be stronger for choosers (i.e. individuals that make an identification) compared to non-choosers (i.e. individuals that reject the lineup; Sporer, Penrod, Read, & Cutler, 1995). Differentiating the evaluation of choosers (as opposed to all decisions) is of importance to the justice system since eyewitnesses that make an identification from a lineup are more likely to testify in a courtroom than eyewitnesses who reject a lineup that includes a suspect. However, point-biserial correlations do not effectively distinguish between correct and incorrect decisions (Juslin, Olsson, & Winman, 1996), regardless of whether the correlation is for choosers alone or all lineup decisions. Accordingly, researchers have shifted to using calibration curves for differentiating between accurate and inaccurate eyewitnesses. Calibration curves plot accuracy at each level of the confidence scale, in contrast to correlations where accuracy is collapsed across all confidence levels (Brewer & Wells, 2006). Like correlation analyses, calibration curves indicate that the confidence-accuracy (CA) relationship is stronger for identifications than rejections or both combined (Sporer, 1993; Sporer et al., 1995). The CA relationship for identifications (choosers) has since been well documented to be stronger than the CA correlation indicates (e.g., Juslin, Olsson, & Winman, 1996; Brewer & Wells, 2006, but current correlational analyses illustrate this also, cf. Lindsay, Read, & Sharma, 1998; Semmler, Brewer, & Wells, 2004). Brewer et al. (2002) demonstrated that CA calibration can be improved by experimental manipulations (such as asking individuals to engage in reflection and hypothesis disconfirmation) in the laboratory, further supporting the usefulness of confidence as a predictor of accuracy, particularly of those who make an identification from a lineup. But while calibration curves provide useful information about accuracy at each level of confidence, their diagnostic value in applied settings is limited since triers of fact are specifically interested in *suspect* identifications: calibration curves evaluate the accuracy of all choosers (filler and suspect identifications). Thus, Mickes (2015) proposed confidence-accuracy characteristic curves (CAC): The adjustment of calibration curves to use suspect identifications only. CAC curve analyses further confirm that confidence can predict accuracy (Mickes, 2015; Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017; Mansour, 2020).

Pristine conditions hypothesis

A recent review on the CA relationship literature proposed the *pristine conditions hypothesis* (Wixted & Wells, 2017). Wixted and Wells (2017) suggests that an eyewitness' initial confidence in a lineup decision conducted under "pristine conditions" is highly predictive of accuracy (i.e. eyewitness reliability). Pristine conditions refer to the control of all relevant system variables (Wells, 1978), or in other words, variables that the criminal justice system can control. Pristine conditions include fair lineup administration (i.e. the suspect does not stand out), a double-blind procedure (i.e. neither the administrator nor the eyewitness knows which lineup member is the suspect), and that confidence statements are gathered *at the time of identification*. Research has shown time and time again that non-pristine conditions interfere with the CA relationship. For example, poorly administered lineups (i.e. post-identification feedback and questioning, e.g., Bradfield, Wells, & Olson 2002, Wells & Bradfield, 1999, Wells & Bradfield, 1998, Shaw, J. S. III. & McClure, 1996), a suspect that stands out (e.g., Malpass, Tredoux, & McQuiston-Surrett, 2007), inappropriate or biased witnessing instructions (failing to instruct the witness that the culprit may not be present, e.g., Steblay, 1997, Malpass & Devine, 1981, Brewer & Wells, 2006), non-blind testing (i.e. administrator knows who the suspect is, Bull Kovera & Evelo, 2020), and a confidence statement not obtained immediately (e.g., confidence statement at trial, e.g., Greenspan & Loftus, 2020; Wells et al., 1998) have been shown to negatively affect the CA-relationship.

Initial identifications made with low confidence, regardless of testing conditions, should be seen as highly prone to error (Wixted & Wells, 2017; Berkowitz, Garrett, Fenn, & Loftus, 2020). Wixted and Wells (2017) argue that the legal system should clearly distinguish between initial confidence taken immediately under pristine conditions and confidence taken at a later point under potentially compromised conditions. Most notably, Wixted and Wells suggests that high confidence, when obtained under pristine conditions, indicates high accuracy.

While there is work that supports the notion that “highly confident eyewitnesses are remarkably accurate” (e.g., Carlson et al., 2017; Wixted et al., 2015; Wixted & Wells, 2017), we cannot conclude that confidence is “undeniably diagnostic of accuracy” even when “fair lineups were created” and when confidence was obtained “solely during the first and only lineup test (…)” (Wixted et al., 2021, p. 3). Pristine conditions do improve the likelihood that confidence predicts accuracy, but they cannot account for or mitigate pre-identification contamination of an eyewitness’ memory. Even though system variables and lineup testing conditions may be controllable, eyewitness memory and subsequent confidence judgements can become corrupted by factors beyond the control of investigators *prior to* lineup administration (e.g., exposure to sources of misinformation via social media, television or other). It is therefore important to continue to investigate factors that might compromise the diagnosticity of confidence statements, even when pristine testing conditions are met. For example, currently, it is unclear how differences in individual procedures may affect the CA relationship for suspect identifications. What happens when one (or more) identification procedures are not pristine (or, not pristine enough)? We do not yet know the full extent to which system factors and memory quality affect the CA relationship. The reliability of identifications made at high confidence under pristine conditions might be justifiable on a collective level but could be misleading when evaluating the accuracy of individual statements (Sauer, Palmer, & Brewer, 2019).

Some research supports the proposition that high confidence can be indicative of high accuracy even when encoding conditions vary. Semmler, Dunn, Mickes, and Wixted (2018) hypothesized that people are aware of factors that may affect the quality of their memories and can thus adjust how many memory cues may be needed to decide with high confidence when witnessing conditions are suboptimal (i.e. their response criterion). Stretch and Wixted (1998) proposed that individuals adjust their decision criterion following a constant likelihood ratio. That is, individuals adjust their response criterion in a way that the probability that an item has been previously seen (when it was judged as such) remains constant under varying conditions. Although individuals attempt a constant likelihood ratio, they are imperfect in doing so (Stretch & Wixted, 1998). For eyewitnesses, Semmler et al. (2018) suggests that signal detection theory can explain the eyewitness' decision process (also cf. Wixted & Mickes, 2014; Smith, Yang, & Wells, 2020; Lee & Penrod, 2019; Ayala, Smith, & Ying, 2022). As encoding conditions grow less favourable and retrieval becomes more complex, "overall accuracy will decline, but the accuracy of a suspect ID made with a particular level of confidence will remain unchanged" (Semmler et al., 2018, p. 400). In other words, suspect identification accuracy made at high confidence levels will remain unchanged even when witnessing conditions vary because people can adjust their decision criteria to match the conditions of encoding (Semmler et al., 2018). However, the extent to which this holds true for high confidence identifications has not been established. Recent work by Giacona, Lampinen & Anastasi (2021) suggests that high-confidence identifications become less precise when estimator variables (variables outside the control of the justice system) are suboptimal. Identifications made with high confidence under poor viewing conditions (e.g., long-distance, weapon presence, and long delay) were significantly less accurate than the high-confidence identifications made under good viewing conditions (e.g., close distance, no weapon, and a short delay). Thus, estimator variables seem to have a larger effect on the CA relationship than perhaps Wixted, Semmler, and their colleagues may believe. More research is needed to test the boundary conditions of high-confidence identifications.

While a large body of literature supports the claim that “high confidence indicates high accuracy” in laboratory settings (e.g., Wixted & Wells, 2017), the applications of the CA relationship in the real-world seem to be limited (Berkowitz & Frenda, 2018, but cf. Wixted, Mickes, Dunn, Clark, & Wells, 2015). Even though the criminal justice system might be able to control for pristine conditions, it is difficult to ensure that such procedures have been implemented and are met in actual criminal investigations. To date, only 25 states in the United States have adopted reforms informed by peer-reviewed research for obtaining eyewitness evidence (Innocence project, 2022). Internationally, the prevalence of best-practice recommendations (and practices) paints an even grimmer picture. A review of guidelines from 54 countries suggests that only 13% of countries include a record of confidence in their provisions (Fitzgerald, Rubinova, & Juncu, 2021). At this point in time, there is no work (and no standard in practice) for how to best interpret eyewitness confidence statements. Given the variety in jurisdictional procedures and our incomplete understanding of their effect, it is premature to assume pristine high confidence universally indicates high accuracy.

Abstract

Eyewitness identification confidence is typically studied using scales (generally numeric); in practice, eyewitnesses typically provide confidence in their own words. Verbal and numeric confidence similarly predict accuracy, but verbal confidence is difficult to interpret reliably (Mansour, 2020). To minimize miscommunication, eyewitnesses could provide scale ratings after verbal judgements or vice versa, but we do not know if the order in which such confidence statements are obtained affects the confidence-accuracy relationship. I tested the utility of requesting both verbal and numeric confidence and whether order effects exist. Participants (*N* = 198) viewed a mock-crime video with two perpetrators. After a delay, they viewed two simultaneous lineups with one perpetrator each and provided confidence for each perpetrator verbally (in their own words) and then numerically (0-100%) or numerically and then verbally. Numeric confidence in identifications was higher when provided first, *t*(393.82) = 2.40, *p* = .02, *d* = 0.24. Confidence-accuracy characteristic (CAC) curve analysis indicates the effect is driven by medium-confidence judgements (numeric range). No order effect was found for verbal confidence (*p* = .32). However, for low and high numeric confidence, verbal followed by numeric was better calibrated than numeric followed by verbal. When the numeric judgement came first, none of the subsequent verbal judgements could be categorized as high confidence using our coding scheme. These data provide preliminary evidence that eyewitnesses should provide only a single confidence judgement.

Eyewitness Confidence

Research shows that verbal (e.g., very confident) and numeric confidence statements (e.g., 90%) are similarly predictive of accuracy (Budescu & Wallsten, 2003; Mansour, 2020; Smalarz, Yang, & Wells, 2021). However, people differ in their preferences for ways to communicate and to receive confidence statements. In research, participant-eyewitnesses typically rate their confidence on a scale (e.g., Mickes, Flowe, & Wixted, 2012; Sauer, Brewer, Zweck & Weber, 2010). When given a choice, people are more likely to express confidence verbally rather than numerically (Dodson & Dobolyi, 2015; Budescu, Karelitz, & Wallsten, 2003; Kenchel, Reisberg, & Dodson 2017, but cf. Kenchel, Greenspan, Reisberg, & Dodson, 2021 who finds individuals prefer expressing confidence numerically in the eyewitness context). Mansour, Batstone & Pennekamp (in preparation) found that mock eyewitnesses and jurors prefer eyewitnesses to express confidence verbally (56% - eyewitnesses, 47% - jurors) compared to numerically (21%, 9%), using both (21%, 42%), or another way (2%, 1%). In practice, confidence is typically obtained in the eyewitness' own words (NAS, 2014).

Windschitl and Wells (1996) theorized that probability estimates derived from deliberative, rule-based reasoning differ from those that do not require deliberation. That is, rule-based probability estimates are likely to be intuitively conveyed using numbers (e.g., 65% chance of precipitation) while associative judgements that do not necessitate deliberation may be better assessed using verbal probability estimates (e.g., unlikely to pass without revision). It may thus be inferred that verbal measures of eyewitness confidence might be superior to numeric confidence measures as they may be more intuitive. Verbal probability estimates (e.g., probably) allow for overlap in associative meaning while numbers do not. For example, “probably” and “likely” could describe similar probability estimates while “30%” universally defines a distinct probability estimate. People are familiar with the use of verbal probability estimates. That is, “when describing their own uncertainty, most people in most everyday situations use words rather than numbers (e.g. “Will you be home by 5?”)” (p. 346, Windschitl & Wells, 1996).

The obtainment of verbal confidence statements seems to be instinctual and practical, but there are caveats to the diagnostic utility of verbal expressions. While verbal confidence statements can offer unique diagnostic information (Seale-Carlisle, Grabman & Dodson, 2020), they are more easily misinterpreted than numeric estimates (Dodson & Dobolyi, 2015; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986).

People frequently underestimate the variability in other peoples’ interpretation of verbal probability statements. For example, Brun and Teigen (1988) asked participants to estimate the range of numeric probabilities covered by a phrase for 90% of the population. The mean range given by participants varied from 50-75% of the actual range, empirically determined in the sample. They also found that different words carry a different emotional charge, which may additionally affect how they are interpreted. Compounded with variability at the receiver’s end, eyewitnesses also use a variety of phrases when judging their confidence (Mansour, 2020). Myself and my supervisor analyzed 3976 verbal confidence statements from 3 data sets and identified 938 unique responses (Pennekamp & Mansour, 2021).

Mansour (2020) provides primary evidence of the challenges inherent in interpreting eyewitness’ verbal confidence statements. First, there is substantial deviation between eyewitness' and individuals' numeric interpretation of an eyewitness' verbal confidence rating. Approximately 15-25% of the time, confidence was interpreted differently from how eyewitnesses meant it—sometimes higher, sometimes lower. Second, for eyewitnesses, Mansour (2020) found that verbal confidence statements were only interpreted in a way consistent with what the eyewitness intended to communicate when confidence was high. To further exacerbate the issue of misinterpretation, Mansour (2020) also found considerable intra-individual variability in translations of verbal confidence statements on a numeric scale. Individuals interpret verbal statements differently, even when statements are provided "in their own words". Similar interpretive difficulties were found by Smalarz, Yang, & Wells (2021). Across three experiments, evaluators systematically underestimated eyewitnesses' verbal confidence.

There are differences in the use of verbal versus numeric expressions of uncertainty. Verbal expressions of uncertainty are generally preferred when situations are deemed unimportant, expressing to the audience that the results are inconsequential or based on weak data (Wallsten et al., 1993, p. 138). On the other hand, individuals prefer communicating uncertainty numerically when situations are deemed important or based on strong data (Mandel, Wallsten & Budescu, 2021). Thus, verbal communication of uncertainty (i.e. confidence) might not accurately convey the significance or intended meaning of a statement.

Despite these challenges or perhaps in light of them, eyewitness identification evidence in the United Kingdom is currently being treated as one or the other, 0% confident (no identification) or 100% confident (identification). If an eyewitness spontaneously provides a confidence judgement, this information is recorded and is available as evidence. However, in the United States and Canada, where confidence is collected, it is typically collected verbally (in "the witness' own words"). Given that confidence statements can provide additional information about the likely accuracy of eyewitness evidence, we need a reliable approach to obtain and interpret eyewitness confidence.

To minimize miscommunications of verbal confidence, recent work proposes the collection of confidence in both ways, verbally and numerically (e.g., Tekin, Lin, & Roediger, 2018). There is limited work available that has tested the collection of both, verbal and numeric confidence, in the eyewitness area. Tekin, Lin and Roediger (2018) compared verbal only and verbal + numeric confidence on a two- and four-level scale. High confidence was associated with high accuracy, irrespective of scale range and method used to obtain confidence. Tekin et al. presented participants with labels (e.g., "not sure at all", "somewhat sure", "very sure", "absolutely sure") but did not assess confidence in the eyewitness' "own words". The verbal + numeric condition included the same labels, "with a corresponding number next to them" (e.g., "3-very sure"). The problem with collecting confidence this way is two-fold. First, there is adversity when individuals are asked to use others' definitions of verbal phrases (Wallsten & Budescu, 1990). The approach is unlikely to translate to practice as it would not be acceptable to "put words in the eyewitness' mouth" (as cited in Mansour, 2020). Second, it is not clear to what extent the "corresponding number" represents the *meaning* of each label. Does "3" accurately represent the meaning of "very sure", for example? Individuals use linguistic probabilities differently depending on context (Clark, 1990), event severity (Harris & Corner, 2011), and outcome valence (Mandel, 2015). Attaching arbitrary numbers to phrases oversimplifies the complexity of language, its natural use in context and subsequent interpretation. Several studies have demonstrated that numerically-bound linguistic probability schemes (such as the verbal + numeric condition in Tekin, Lin & Roediger, 2018) are not accurately interpreted by individuals (Budescu et al., 2009; 2012; 2014). That is, individuals do not associate terms within the attached ranges. Forcing the categorization of natural language use may thus not be a viable option.

Even though the combined collection of verbal and numeric confidence judgements may not affect the CA relationship individually (Tekin, Lin, & Roediger, 2018; Dodson & Dobolyi, 2015; Mansour, 2020), we do not know the extent to which the order in which confidence judgements are provided affects the CA relationship. While collecting both, verbal and numeric confidence judgements, may seem like a viable option, different cognitive mechanisms are used to recall fine-grain (precise, e.g., numeric) versus coarse-grain (broad, e.g., verbal) information (Brewer, Vagadia, Hope & Gabbert, 2018) and post hoc judgements of decision-making are influenced by the decision itself (Hall, Johansson, & Strandberg, 2012). Findings on order effects are mixed: pilot data from Smalarz et al. (2021) suggests no effect of order on the CA relationship but the authors noted limited generalizability due to small sample sizes. The results of the main experiment in Smalarz et al. suggest that the order in which numeric and verbal statements are obtained does have an effect. Numeric confidence statements were higher when provided after a verbal confidence statement (compared to numeric confidence statements before a verbal confidence statement). In addition, Smalarz et al. report that individuals naturally expressed confidence numerically, even when asked to "use words, not numbers". Smalarz et al. conclude that individuals may be "naturally predisposed to provide confidence statements numerically as opposed to verbally" (p. 143). It is crucial to determine best practices for obtaining confidence statements. We need to do so in a way that considers peoples' preferences and natural predispositions for providing confidence while simultaneously minimizing risks for misinterpretation.

As a first step, we (i.e. myself and supervisor) sought to test for order effects when obtaining both verbal and numeric confidence statements. Our methods, hypotheses, and analyses were pre-registered on the Open Science Framework and can be found via the following link: https://osf.io/ypt78. We expected that a numeric scale rating following a verbal statement would strengthen the relationship between confidence and accuracy (cf. numeric then verbal), based on evidence that 1) people prefer to give confidence verbally (Wallsten & Budescu, 1995), 2) that suggests the accuracy of people's memory reports is highest when they are allowed to choose how coarsely/precisely they report (Koriat & Goldsmith, 1996), and 3) that the cognitive interview results in more information than a standard interview, in part because it uses multiple retrieval opportunities, which involve obtaining a first narrative followed by probing for more information about that narrative (Fisher & Schreiber, 2007; i.e. verbal confidence statement may be similar to a first narrative). Based on preliminary data from our laboratory and previous work (Kenchel, et al., 2021; Pennekamp, Batstone, & Mansour, 2019), we expected individuals to prefer receiving ratings of confidence numerically but to prefer giving ratings of confidence verbally. We did not have specific hypotheses for the CA relationship for rejections. We also did not have specific hypotheses for the preference of other types of scales. We examined this relationship in an exploratory fashion.

Method

The study was approved by the university’s research ethics board.

Participants

Participants were adults with sufficient visual capacity to view a computer screen. Participants (*N* = 366) were recruited via the university‘s SONA system for course credit, social media, and word of mouth. The usable subsample (*n* = 198) did not include survey previews, duplicate IP addresses, cases where the participants failed 3 of the 3 attention checks, indicated technical difficulties, or did not provide at least one identification decision and one confidence decision (*n* = 168). We did not obtain further demographic information due to an oversight.

Design

This experiment used a 2 (Target presence: present, absent) x 2 (Order: verbal→numeric, numeric→verbal) x 2 (Target: perpetrator, accomplice) mixed design, involving a within-subjects manipulation of Target presence. Participants viewed two lineups (one for the perpetrator, one for the accomplice), with one lineup including the suspect (target-present) and one not (target-absent). All participants viewed one target-present and one target-absent lineup. The Order of confidence statements was manipulated between-subjects (verbal → numeric or numeric → verbal). We varied the role of the actors who served as Targets between the two videos (perpetrator or accomplice, perpetrator or victim) in the mock-crime video shown to ensure stimulus sampling (Wells & Windschitl, 1996). However, all analyses were collapsed across this perpetrator factor. Participants were randomly assigned to view one of two perpetrator videos.

Materials

The study was programmed on Qualtrics.

Mock-crime videos

The videos portrayed two males in a parking lot. One of two Caucasian males (the perpetrator) approaches the male Caucasian victim and steals a bag before exiting the scene. The second Caucasian male (accomplice) observes the interaction and then exits with the perpetrator (see Figure 22, Appendix 2). Two videos were used. In one, actor A was the perpetrator, actor B was the accomplice, and actor C was the victim. In the other, actor B was the perpetrator, actor C was the accomplice, and actor A was the victim. All of the videos were approximately 12 seconds in duration.

Intervening task

The participants were presented with an intervening task between viewing the mock-crime video and the lineups. Participants were asked to perform a two-minute “Where is Waldo/Wally” visual search task, to create a short delay between witnessing and the lineup decision. Due to low identification rates after initial data collection (19 correct IDs, 84 false alarms after 175 participants total), we reduced the delay time to 30 seconds. However, identifications rates were still low after more participants were collected (21 correct IDs, 97 false alarms after 195 participants total). We thus further reduced the delay to 10 seconds (no additional correct IDs, 1 additional false alarm). All participants were included for analyses. For the task, participants were asked to answer questions relating to a Where's Waldo?[1] image which was displayed on a screen and depicted a busy beach. The 12 questions were ones such as "How many open umbrellas are there?" and "What colour is the beach ball in the scene?". The questions were designed to be impossible to complete within two minutes.

Lineups

Participants were presented with two simultaneous lineups in a 3 x 2 array. Only the head of the lineup members was shown. Six fillers (non-target individuals known to be innocent), were selected using an iterative match-to-description process (Mansour, 2020) from the laboratory database. Following Oriet & Fitzgerald (2018), targets similar in appearance to each other were selected so that they could act as innocent suspects to each other and thereby provide a means by which to have target-absent lineups constructed for the suspect (also see Mansour, 2020). No person appeared in more than one lineup. The target-absent lineups consisted of all six fillers, whereas the target-present lineups included the culprit and five randomly chosen fillers (see Figure 23, Appendix 2). The culprit-present Suspect 1 lineup had a Tredoux's *e* of 5.07, 95%CI[4.10, 6.64] after being shown to 48 individuals. The culprit-absent Suspect 1 lineup had a Tredoux's *e* of 4.02, 95%CI[3.11, 5.68] after being shown to 49 individuals. The culprit-present Suspect 2 lineup had a Tredoux's *e* of 4.75, 95%CI[3.58, 7.02] after being shown to 52 individuals. The culprit-absent Suspect 2 lineup had a Tredoux's *e* of 2.71, 95%CI[1.85, 5.03] after being shown to 49 individuals. The culprit-present Suspect 3 lineup had a Tredoux's *e* of 4.48, 95%CI[3.68, 5.72] after being

[1] TM and © 2008 Entertainment Rights Distribution Limited. All rights reserved.

shown to 98 individuals. The culprit-absent Suspect 3 lineup had a Tredoux's *e* of 5.13, 95%CI[4.27, 6.41] after being shown to 100 individuals.

Each participant made two lineup decisions, equally distributed between target-present and target-absent lineups. Participants were instructed that the target may or may not be present in the lineup before each lineup decision (Perpetrator instructions: *Think back to the video you watched at the start of the study. Imagine that the police have contacted you because they have two suspects in custody. They would like you to look at a lineup for each suspect. Each lineup will contain ONE or NONE of the "criminals" from the video you watched. You will see a picture of all the lineup members at once. If you see one of the "criminals" from the video in the lineup, please click the number associated with his face. If you do not see any of the "criminals" in the lineup, please click "not there." We will first show you a lineup for the "criminal" who threatened the victim in the video.* Accomplice instructions: *We would now like to show you a lineup for the second "criminal" in the video (the one who did not interact with the victim). If you would like a reminder about the instructions for the lineup, they are repeated below: The lineup will contain ONE or NONE of the "criminals" from the video you watched. You will see a picture of all the lineup members at once. If you see one of the "criminals" from the video in the lineup, please click the number associated with his face. If you do not see any of the "criminals" in the lineup, please click "not there.*). The order of the presentation of lineups was unknown to participants. Participants could select one of the six lineup members, respond "not there" or respond "I do not know".

Immediately after each lineup decision, participants rated their confidence in their decision either verbally first, then numerically or numerically first, then verbally (dependent on condition). For the numeric judgement, participants were asked to rate how confident they were on a scale of 0 = Not at all confident to 100 = Completely confident. For the verbal judgement, they were asked to provide their confidence "in their own words" ("Please tell us how confident you are in the accuracy of your lineup decision using your own words").

Preferences

We asked participants for their preferences when asked to give confidence and to receive confidence as well as their preferences for different types of methods to obtain confidence statements ("In your opinion, which is the best way to ask someone for their confidence?").

Metacognitive awareness

We asked participants if they think they provided confidence automatic or deliberately ("Do you think your **first** judgement of confidence was deliberate (intentional, conscious) or do you think it was automatic (effortless, quick)?").

Attention checks

Participants were asked questions pertaining to video quality ("Did you complete this study on a phone? This will not affect your credit.") and technical difficulties ("Did you have any technical difficulties?"). Data from trials where a participant responded that they could not watch the video were excluded from analysis. The study included three attention checks (e.g., "What colour was the backpack shown on the ground that was later carried away?"). Participants that failed all three attention checks were excluded from analyses.

Procedure

After viewing the information sheet and providing informed consent (see Appendix 1), participants were asked to complete the study on a desktop computer. Participants were instructed that they would be seeing a video before actually watching the video. After watching one of the mock-crime videos, participants were informed that they had become a witness to a crime and that they would be asked to look at a lineup after completing an intervening task intended to mimic the delay during which police investigate the crime. They then completed the intervening Where’s Waldo? task. Next, participants read the lineup instructions and then made their lineup decision before providing their confidence judgements. Participants made lineup decisions on two lineups (one lineup for the perpetrator, one lineup for the accomplice) but provided confidence after each lineup decision. Next, participants were asked questions about their preferences regarding confidence and their metacognitive awareness. Finally, the participants were debriefed, thanked for their participation, and granted credit if they were recruited for credit.

Measures

Lineups

Lineup decisions were coded as correct (suspect identifications from target-present lineups, rejections of target-absent lineups) or incorrect (filler identifications, suspect identifications from target-absent lineups, rejections of target-present lineups) or “I do not know”. For confidence-accuracy characteristic (CAC) curves, we estimated the innocent suspect identification rate by dividing the total number of identifications from the target-absent lineups by the number of lineup members (six).

Confidence

Confidence judgements were coded as indicating low, medium, or high confidence following Mansour (2020) coding scheme (e.g., *Looks kind of like* = low confidence, *Pretty sure* = medium confidence, *Very certain* = high confidence). For the scale-based confidence judgements, data for confidence was collapsed from the original 0-100 scale into three categories: Low confidence = 0%-49%, Medium = 50%-79%, and High = 80%-100% (Mansour, 2020). We also looked at the numeric confidence judgements on a finer point scale (i.e. 5 categories instead of 3). That is, we partitioned the CAC curves into 5 confidence bins (0-19%; 20-49%; 50-69%; 70-89%; 90-100%). We partitioned the CAC curves into these bins to compare our data to prior work indicating "high confidence" to be 80%+ and 90%+. Verbal judgements of confidence were coded as low, medium, or high confidence according to Behrman and Richards' (2005) and Mansour's (2020) coding scheme[2]. We also calculated calibration curves (Brewer & Wells, 2006; Mansour, 2020; Vredeveldt & Sauer, 2015).

Own words confidence

We coded responses that participants provided in response to "Please tell us in your own words" for mode of confidence (whether participants provided a verbal statement, numeric statement, or both).

[2] Mansour (2020) presented participants with confidence phrases and asked them to provide a numeric estimate of the confidence rating represented by the phrase on a 0-10 scale. Low confidence judgements were associated with ratings below 5, medium confidence judgements were associated with ratings from 5-7 and high confidence judgements with ratings of 8 and higher.

Results

Below we separately analyse the results for the scale-based confidence judgements and the own-words confidence judgements. The sample sizes for these analyses differ slightly because some participants provided uninterpretable responses when asked to provide confidence in their own words (e.g., “When I observe something or someone with details my long-term memory is very strong.”). Throughout the results, effect sizes were calculated using R’s effectsize package (Benn-Shachar, et al., 2020).

Accuracy

Correlations

In addition to the analyses below, we also conducted confidence-accuracy Pearson’s correlations for comparison with older studies. For verbal confidence, the confidence-accuracy correlation was small, $r(193) = .21$, $p = .003$. For numeric confidence, the confidence-accuracy correlation was weak and non-significant, $r(192) = .12$, $p = .104$.

Confidence ratings

We conducted t-tests with confidence condition as the independent variable (two levels) and confidence rating as the dependent variable. Numeric confidence was higher when provided first, $t(393.82) = 2.40$, $p = .02$, $d = 0.24$. There were no significant differences for verbal confidence, $t(382.32) = 1.00$, $p = .32$, $d = 0.10$.

Identifications

Table 1a

Descriptive Statistics for Numeric Confidence Statements

Order	Confidence bin	Correct IDs	Target-present filler IDs	Target-absent filler IDs	Incorrect rejection	Correct rejection	I do not know	IDs
Numeric first	high	1	2	2	6	5	9	0.33
Numeric first	medium	4	15	23	6	15	6	0.15
Numeric first	low	5	22	20	12	10	31	0.20
Verbal first	high	3	1	3	3	8	5	0.50
Verbal first	medium	1	14	15	5	13	2	0.06
Verbal first	low	7	34	38	10	11	29	0.16

Table 1b

Descriptive Statistics for Verbal Confidence Statements

Order	Confidence bin	Correct IDs	Target-present filler IDs	Target-absent filler IDs	Incorrect rejection	Correct rejection	I do not know	IDs
Numeric first	high	N/A	2	3	3	N/A	1	N/A
Numeric first	medium	4	11	24	10	14	1	0.14
Numeric first	low	6	25	17	9	14	33	0.26
Verbal first	high	3	1	5	1	5	4	0.38
Verbal first	medium	2	19	16	9	15	1	0.11
Verbal first	low	5	28	34	7	12	28	0.13

CAC Curves

Scale-based Confidence Judgements. Figure 1 displays the CAC curves for confidence judgements for the two Order conditions tested in the current experiment. It is apparent that there is a difference in the degree of calibration when participants were medium confident. Contrary to our hypothesis of better calibration for the verbal →

numeric condition, the verbal → numeric condition resulted in poorer calibration than the numeric → verbal condition for medium-confidence identifications (see Figure 1A). Specifically, providing a verbal judgement first seems to have led participants to be overconfident in the medium-confidence category.

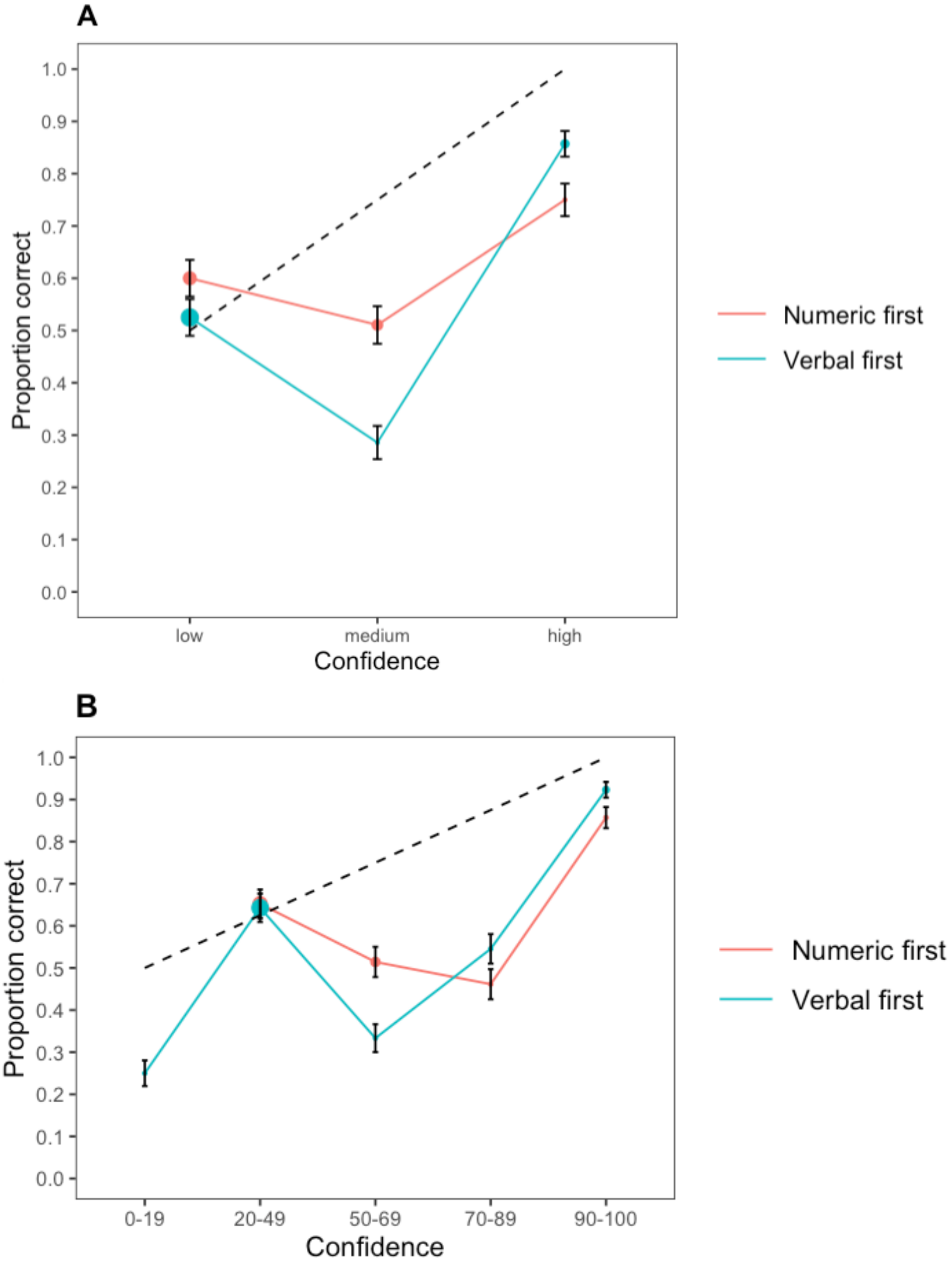
A different pattern was found for the low- and high-confidence categories for the verbal → numeric condition. We predicted that the verbal → numeric condition would show improved calibration relative to the numeric → verbal condition. This hypothesis was supported: For low and high numeric confidence, calibration was better when a verbal judgement versus a numeric judgement was given first. However, participants in both conditions were underconfident in the low confidence category and overconfident in the high confidence category. Highly confident participants were highly accurate (80-90%).

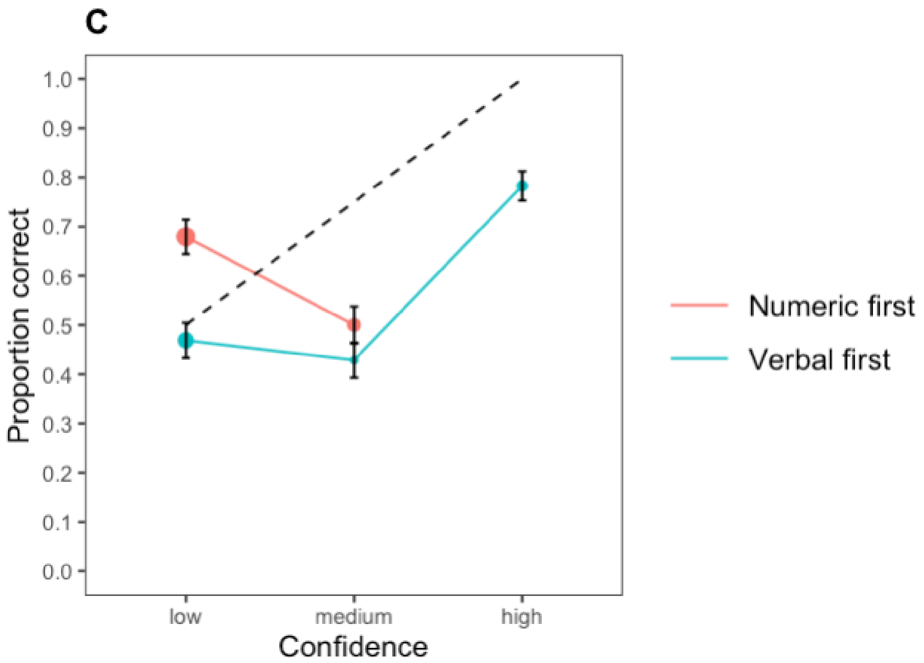
We looked at the CAC curves for numeric confidence in a more fine-grained way by splitting confidence into five (rather than three) bins. This allowed us to make more nuanced inferences about the differences in the CA relationship across the numeric confidence scale. Figure 1B presents the results. The difference in calibration due to order of judgement is driven by medium confidence judgements. While both conditions were well calibrated at 20-49%, both conditions showed overconfidence for the 50-69% and 70-89% bins.

Own-words Confidence Judgements. Figure 1C shows the CAC curves for the own-words confidence judgements. The pattern is similar to that for the scale-based judgements: Both conditions show overconfidence at medium confidence. Cummings, Findler, and Vaux (2007) note that if doubling the length of the standard error bars (when *n* > 3) does not result in overlapping bars, then the difference between the points will be significant when alpha = .05. Based on this rule of thumb, the difference between the Order conditions is unlikely to be significant. Unlike the scale-based judgements, participants were overconfident when they were low-confident in the verbal → numeric condition, but underconfident in the numeric → verbal condition. similar to the scale-based judgements, highly confident participants were 75% accurate. However, when numeric confidence was provided first, none of the subsequent verbal confidence judgements were judged to be of high confidence according to the coding scheme.

Figure 1

Confidence-Accuracy Characteristic Curves (CAC)





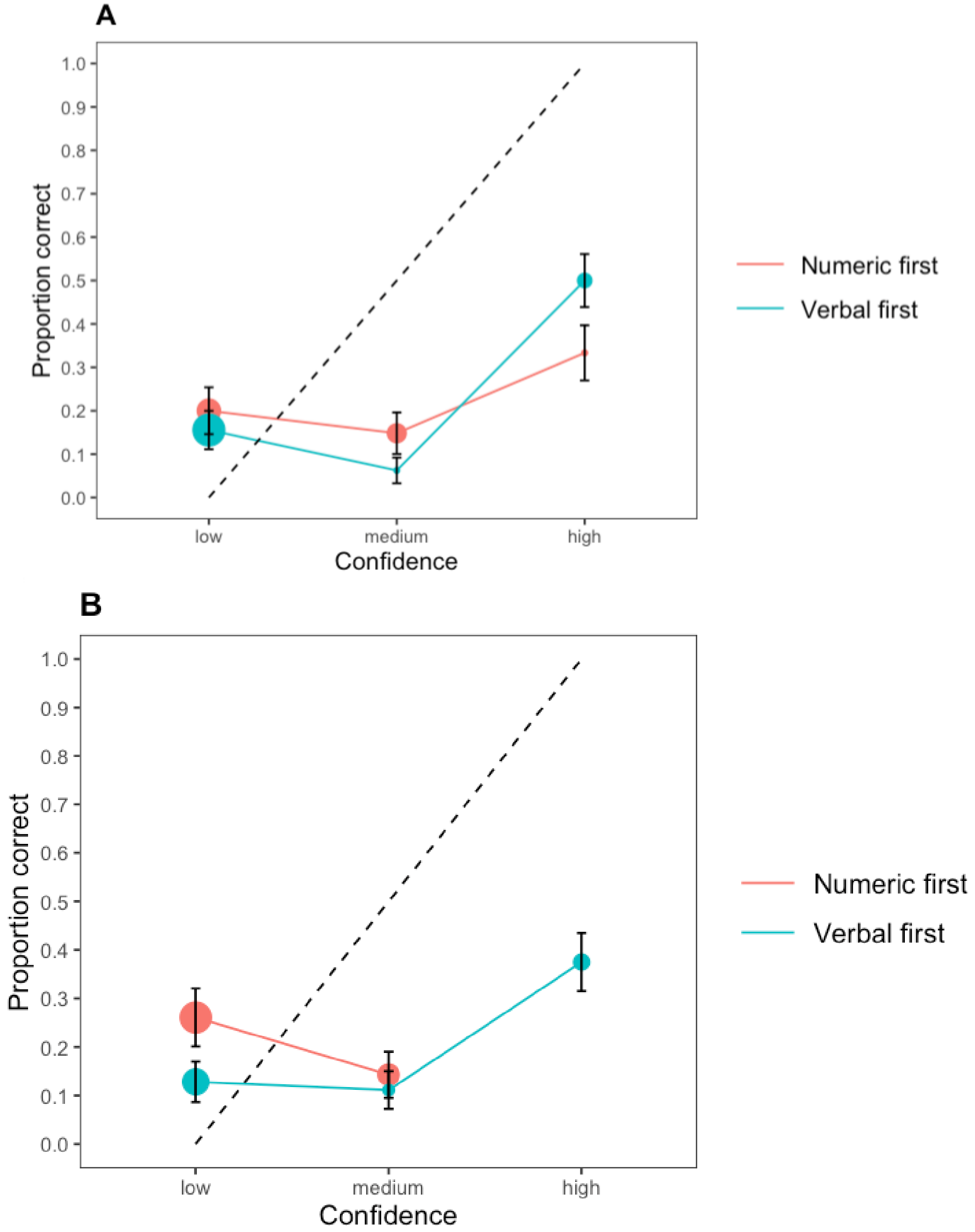
Note. A) Numeric confidence in three bins, B) Numeric confidence in five bins, C) Verbal confidence in three bins. Error bars reflect standard error. Innocent suspect identifications were estimated by dividing the number of target-absent lineup identifications by six. In panel C, none of the subsequent verbal judgements were categorized as high confidence when numeric judgements were provided first. Dotted identity line represents perfect calibration.

Calibration Curves

Scale-based Confidence Judgements. Figure 2A depicts the relevant calibration curves. As with the CAC curves, there is obviously no difference between the order conditions when confidence is low. For medium- and high-confidence decisions, participants in both conditions appear overconfident. For high-confidence decisions, the verbal→numeric condition appears to be slightly better calibrated.

Own-words Confidence Judgements. Figure 2B depicts the calibration curves for each condition for the own-words confidence judgements. The pattern is very similar to that of the scale-based judgements. Medium- and high-confidence participants were overconfident in both conditions. High-confidence decisions better calibrated for the verbal→numeric condition, but again there was overconfidence in both conditions. The standard error bars are sufficiently wide to suggest the differences are not reliable.

Figure 2
Calibration Curves for Identifications



Note. A) Numeric confidence in three bins, B) Verbal confidence in three bins. Error bars reflect standard error. As is common procedure (e.g., Brewer & Wells, 2006), filler identifications from target-present lineups were not used in constructing the calibration curves. Points on the graph indicate participant numbers in each bin. In panel B, there were no high-confidence numeric judgements when confidence was provided numerically first. Dotted identity line represents perfect calibration.

Logistic Regression

Scale-based Confidence Judgements. We conducted a logistic regression with scale-based confidence, order, and their interaction as predictors of accuracy. There were no significant effects. Neither confidence (p = .40), nor order (p = .84), nor the interaction (p = .68) predicted accuracy.

Own-words Confidence Judgements. A logistic regression with own-words confidence (low, medium, high), order, and their interaction as predictors of accuracy was conducted. There were no significant effects. Neither confidence (p = .32), nor order (p = .11), nor the interaction (p = .08) predicted accuracy.

Rejections

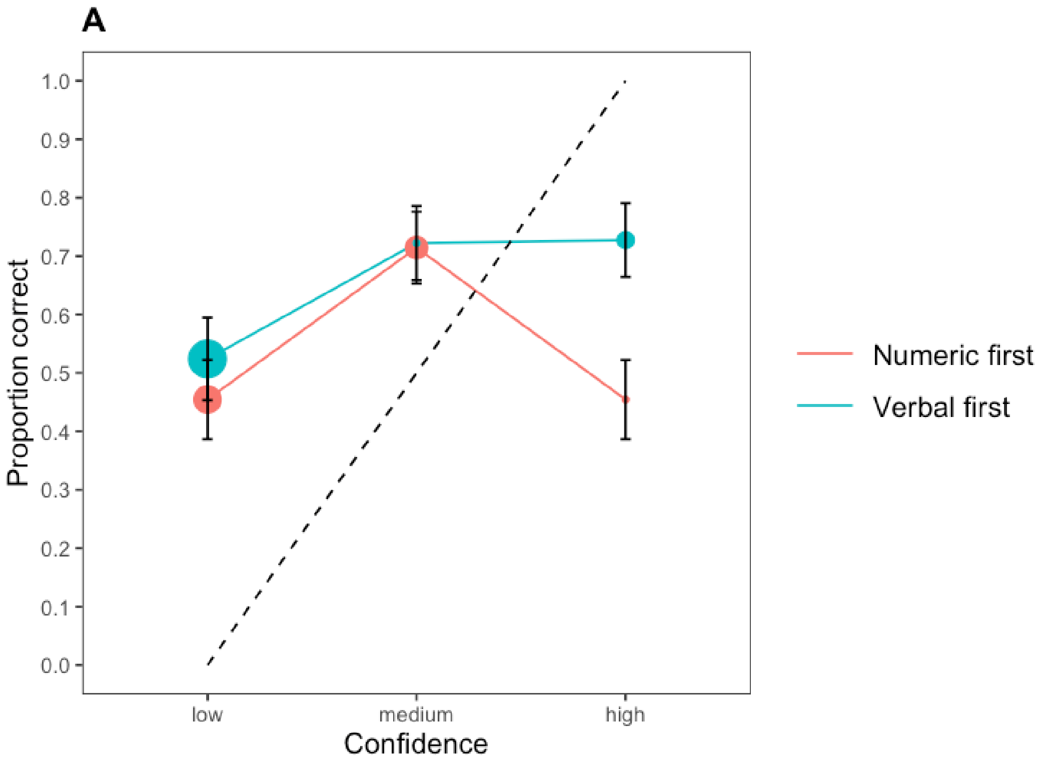
Calibration Curves

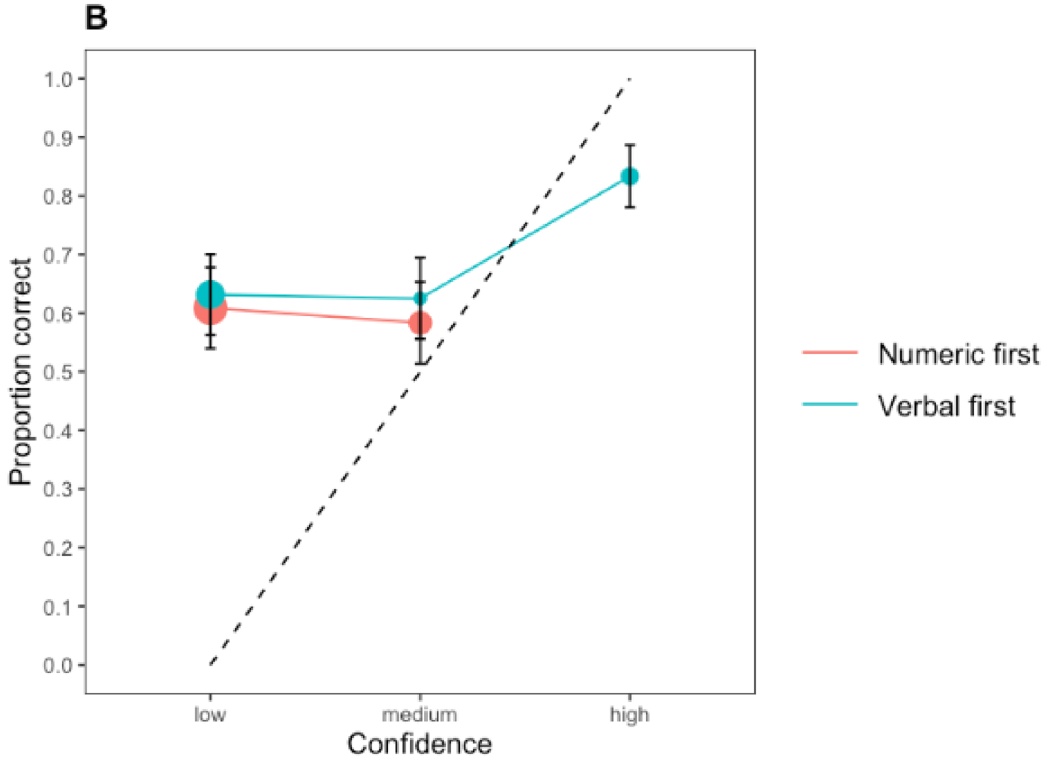
Scale-based Confidence Judgements. Consistent with prior research (e.g., Brewer & Wells, 2006), we generally found poor calibration for rejections; Figure 3A displays relatively horizontal lines across conditions. We do not find support for the idea that high confidence rejections may be more strongly associated with accuracy than low- or medium-confidence rejections (Wixted & Wells, 2017). Interestingly, underconfidence was present for low- and medium-confidence decisions, whereas high-confidence decisions tended towards overconfidence.

Own-words Confidence Judgements. Figure 3B illustrates the calibration of rejections when confidence was provided in the participants’ own words. Like for the scale-based confidence judgements, the curves for both Order conditions are relatively horizontal. There were no high confidence rejections when numeric confidence was obtained first. Medium-confidence participants were overconfident in both conditions.

Figure 3

Calibration Curves of Rejections





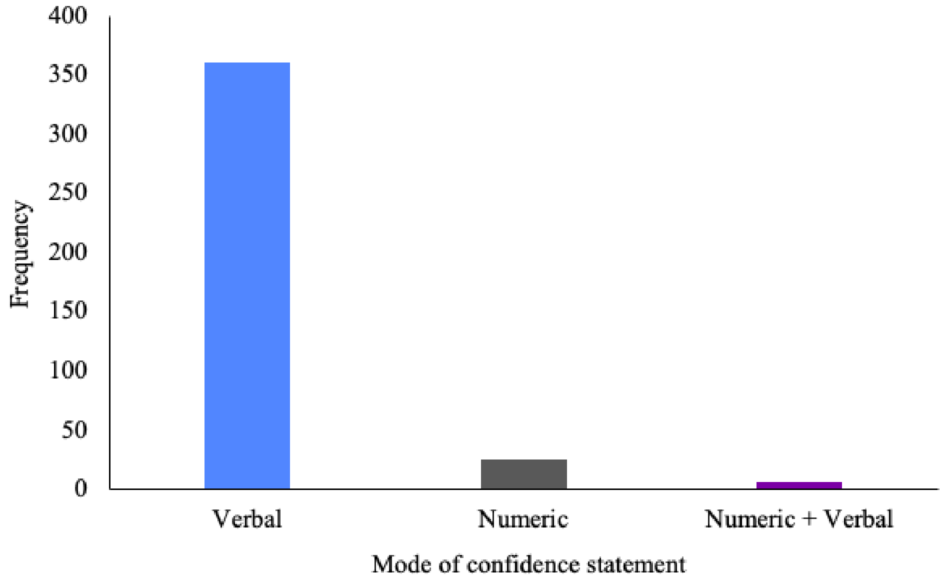
Note. A) Numeric confidence in three bins, B) Verbal confidence in three bins. Error bars reflect standard error. Calibration curves include all rejections. In panel B, there were no high numeric judgements when confidence was provided numerically first.

Preferences

Type of own words confidence. When asked to provide confidence in "one's own words" ("Please tell us in your own words"), participants most frequently provided a verbal statement. Participants rarely used numbers when asked for confidence in their "own words" (See Figure 4).

Figure 4

Responses to Confidence in "Own Words"

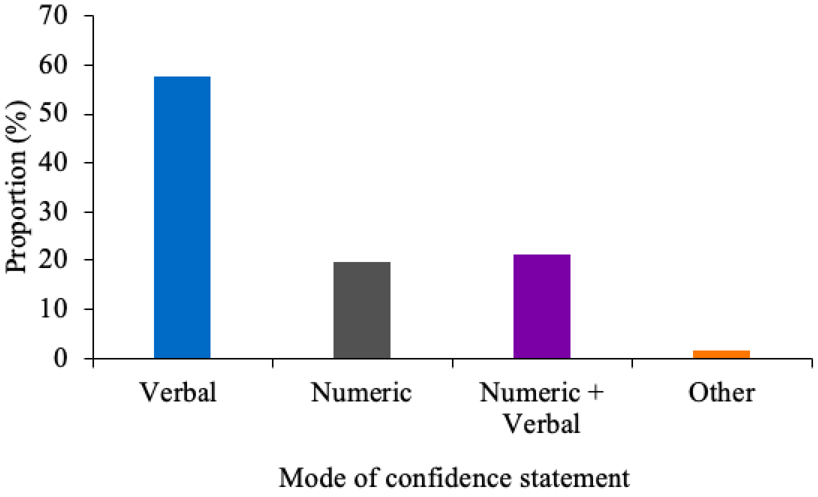


Note. Frequency of mode of confidence statement when participants were asked to provide confidence in their own words.

Obtaining and presenting confidence. When asked for their preferences to give and receive/hear confidence if this was a real crime, participants preferred to give confidence verbally (see Figure 5) and preferred to receive/hear confidence verbally (see Figure 6).

Figure 5

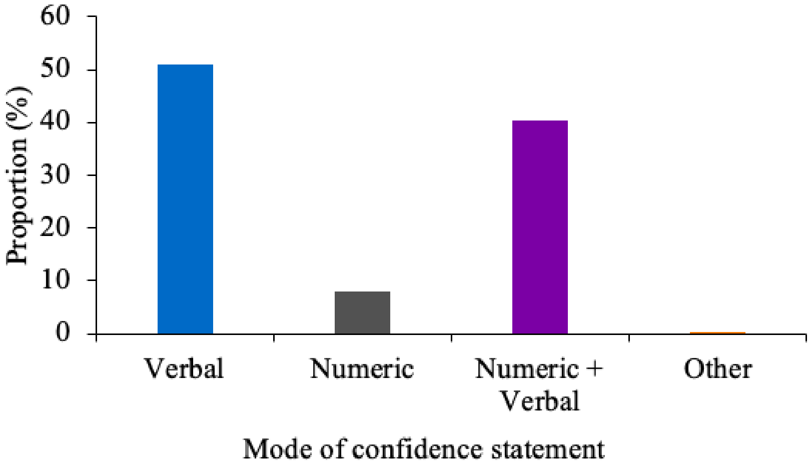
Preferences for Giving Confidence



Note. Frequency of mode of confidence statement when participants were asked for their preference to give a confidence statement.

Figure 6

Preferences for Receiving Confidence



Note. Frequency of mode of confidence statement when participants were asked for their preference to hear/receive a confidence statement.

Scale preferences. We asked for preferences of different types of methods for obtaining confidence. Participants indicated to prefer combined methods that provided both, verbal and numeric information (see Table 2).

Table 2

Preferences for Different Methods to Obtain Confidence

Scale type	Percentage (%)
Numeric scale	15.66
Likert-type scale	27.78
Visual Analogue scale	6.06
Numeric and verbal scale	36.36
Verbal	14.14

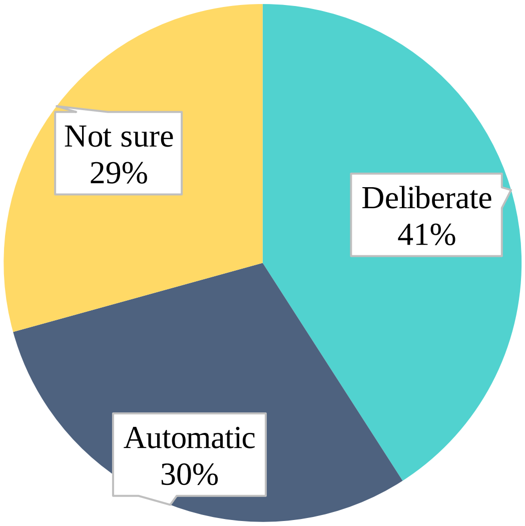
Note. Scale types were presented visually (see Figure 24, Appendix 3).

Metacognitive awareness

We asked participants if they thought they provided confidence automatic or deliberately. We report the proportion of responses by participants. More participants indicated that they thought their first confidence judgement was deliberate (41%) rather than automatic (30%; see Figure 7).

Figure 7

Metacognitive Awareness of Initial Confidence Statements



Note. Percentage of responses when participants were asked if their initial confidence statement was provided deliberately versus automatically.

Discussion

We tested verbal and numeric methods for obtaining confidence to investigate whether the order of confidence statements affects the CA relationship. We expected that a numeric scale rating following a verbal statement (verbal→numeric) would strengthen the relationship between confidence and accuracy compared to a verbal statement following a numeric scale judgement. We did not find support for this hypothesis. Confidence did not predict accuracy.

However, individuals were slightly better calibrated for low- and high-confidence when verbal confidence was obtained first. Importantly, our results suggest that the order in which confidence statements are obtained matters to some extent: Numeric confidence was higher when provided first. When numeric confidence was provided first, none of the subsequent verbal judgements were categorized as high confidence using our coding scheme. Our data differ from Smalarz et al. (2021) who found numeric confidence is higher when given second. There are possible explanations for these differences. CAC analysis indicated that our effect is driven by medium-confidence judgements. In our study, memory quality was poor. A majority of our participants made decisions with low confidence (447 low-confidence judgements total). In comparison, there were only 76 high-confidence judgements. It may be that there are differences in subsequent interpretations of different levels of confidence. For example, a verbal confidence statement of high certainty (e.g., *very confident*) may be easier to interpret (and to translate) numerically (e.g., 95%) versus a confidence statement of moderate or low certainty (e.g., *pretty sure*). Similarly, it may be easier for individuals to correctly interpret high numeric estimates (e.g., 95%) verbally, but not medium or low estimates.

When a verbal judgement came second, those verbal judgements were never interpreted as high confidence based on our coding scheme. It may be that individuals assign different meanings to verbal estimates as compared to numeric estimates. For example, one participant indicated to be “80%” confident in their decision (high confidence numerically) but “rather confident (…)” when asked to provide confidence in their own words (medium confidence). Another participant indicated they were “85%” confident in their decision (high confidence numerically) but when asked to provide confidence verbally indicated to be “(…) fairly sure (…)” (medium confidence). There is considerable variability when individuals translate numeric estimates into verbal statements. For example, Mansour (2020) found that individuals’ verbal judgements were only interpreted accurately when confidence was high. This difficulty may be exacerbated when others are asked to translate verbal statements into numeric estimates (such as police officers, judges and jurors), particularly if those individuals already have opinions about the case or the eyewitness. Confidence in one’s “own words” allows individuals to provide further reasons for their (un)certainty (e.g., references to features, “I'm fairly sure the second guy was black curly haired”). However, research shows that such references (or responses as to “why a decision was made”, as commonly asked in policing practice in Scotland) are even more challenging for individuals to interpret (Mansour, 2020) and predict false identifications (Grabman, Dobolyi, Berelovich, & Dodson, 2019).

Based on preliminary data from our laboratory and previous work (Kenchel, et al., 2021; Mansour, 2020), we expected individuals to prefer receiving ratings of confidence numerically but to prefer giving ratings of confidence verbally. Our hypotheses were partially supported. Participants preferred giving confidence verbally. However, participants also indicated to prefer receiving confidence verbally (as opposed to numerically) followed by combined methods (verbal + numeric). This finding differs from previous research (Wallsten, et al., 1993; Kenchel, et al., 2021; Pennekamp, Batstone, & Mansour, 2019). In the present study, we asked participants how they would prefer to “hear/receive confidence” if they were a juror. It may be that our instruction (“hear”) influenced participants’ indication of preference (priming). Similarly, asking participants to provide confidence in “their own words” prior to may have influenced indications of preference. It may also be that asking participants to provide both, preferences to give and to receive confidence, could have introduced response bias (i.e. participants may have tried to stay consistent in their answers). Nevertheless, these findings support current policing practice: Verbal confidence methods, such as obtaining confidence in the “witness own words”, seem to be preferred. When asked to express confidence “in their own words”, individuals rarely used numbers to express their confidence. This is contradictory to findings reported by Smalarz et al. (2021) who found that individuals report numbers instinctually when asked for a verbal judgement. “20% of participants in the verbal confidence statement condition provided a confidence statement using a number despite the instruction to use words and not numbers” (Smalarz et al., 2021, p. 143). We suggest it may be attributable to the differences in instruction: Smalarz et al. instructed participants to “use words, not numbers, (…)” (p. 142). In our study, participants were instructed to provide confidence “in your own words”. It is possible that the type of instruction influences the reasoning process that eyewitnesses engage in to provide confidence (Windschitl & Wells, 1996). There were also differences in who was sampled in these studies. We recruited a mix of students and laypeople (via social media) whereas Smalarz et al. recruited students (pilot studies) and Amazon Mechanical Turk workers. More research is needed to better understand people’s preferences in relation to expressing their confidence as an eyewitness.

We also assessed preferences for different types of methods to obtain confidence. We did not have specific hypotheses for the preference of other types of scales. We thus examined this relationship in an exploratory fashion. Participants preferred visual analogue scales that provide both verbal and numeric information (i.e. Likert-type scales, numeric and verbal scales). Combined formats are shown to be predictive of accuracy, similar to verbal-only or numeric-only methods (Tekin, Lin, & Roediger, 2018; Mansour, 2020) but do not overcome interpretive difficulties (cf., Budescu, 2009; 2012; 2014). This result is promising and deserves further investigation: Multi-modal formats, such as a visual analogue scale that combines verbal and numeric information, could provide alternatives (and possibly advantages) to verbal-only, numeric-only methods and numerically-bound linguistic schemes.

We assessed metacognitive awareness by asking participants if they thought their first confidence judgement was deliberate or automatic. Interestingly, individuals do not necessarily seem to know whether they engage in a deliberation process when making judgements of confidence (29% not sure), though we found somewhat more participants felt their judgement was deliberate (41%) versus automatic (30%). This finding has important theoretical implications. Research shows that accurate identifications are faster than inaccurate identifications (e.g., Brewer, Caon, Todd, & Weber, 2006). Specifically, automatic recognition (and reference to such, e.g., “the face popped out”) is more likely to be accurate than when people use a deliberate process (Dunning & Stern, 1994; Grabman et al., 2019). Windschitl and Wells (1996) suggests that numeric measures “elicit deliberate and rule-based reasoning from respondents, whereas verbal measures allow for more associative and intuitive thinking” (p. 343).

That said, our mixed results may be due to the design of our study: participants provided numeric and verbal estimates. It may be that eliciting both influences the reasoning processes that underly judgements of probability. However, people do not have perfect insight into internal metacognitive processes (e.g., Dunning & Stern, 1994). Yet, Semmler et al.’s (2018) constant likelihood ratio model suggests that individuals are aware of factors that influence their memory (and subsequent confidence judgements). If individuals in fact adjust their response criterion following such a constant likelihood ratio, they may also have knowledge about the processes (or factors that may affect) underlying their confidence statements. Future research should examine whether individuals that consider their initial confidence statement to be automatic are better calibrated than individuals that consider their initial confidence statement to be deliberate (or that are “not sure”).

While our study provides additional information about the utility of methods for obtaining confidence statements, there are limitations to this research. Our sample size was relatively small. Jusslin, Wilson, and Olsson (1996) suggest that the ideal sample per point in a calibration analysis is 200 participants, however, no empirical analysis has addressed the issue of sample size with confidence-accuracy calibration curves and it is common to have samples of 100 participants per point. Future research should aim to replicate these findings using a larger sample and especially, a higher proportion of correct suspect identifications.

A limitation of our experiment, however, is that only 21 lineup decisions were correct suspect identifications (out of 210 identifications; 396 decisions total). It is likely that our stimuli simulated suboptimal viewing conditions (e.g., short viewing time). Viewing conditions certainly affect the accuracy of eyewitnesses (Semmler, Dunn, Mickes, & Wixted, 2018; Cutler, Penrod, & Martens, 1987) and sometimes also the confidence accuracy relationship (Lockamyeir et al., 2020; Giacona, Lampinen, & Anastasi, 2021). We reduced the time of delay (from two minutes to 30 seconds to 10 seconds) to minimize the time between encoding and test in hopes that memory quality would be improved (e.g., Grabman et al., 2019; Sauer et al., 2010). However, correct suspect identification rates remained low. Future research should further investigate the role of estimator variables (such as viewing time) on confidence to address if high confidence is predictive of high accuracy when conditions are poor to determine possible boundary conditions of this relationship.

Our research tested the utility for obtaining both verbal and numeric eyewitness confidence statements. Numeric confidence was higher when provided first but people were better calibrated when confidence was obtained verbally first. Our findings highlight the necessity for replication and provide preliminary evidence that eyewitnesses should provide only a single confidence judgement. Importantly, individuals consider methods that present information in combined formats (i.e. numerically + verbally) to be superior to other methods for obtaining confidence statements from eyewitnesses.

Abstract

We (i.e. myself and supervisor) developed an evidence-based tool for assisting with communication and interpretation of eyewitness confidence. Participants rated how well percentages (0%, 10%...100%) represented each of 13 common verbal confidence statements (e.g., moderately confident) on a scale (0 = *Not at all* to 100 = *Absolutely*). From the numeric distributions (membership functions) derived from each phrase's ratings, we identified four phrases with clear boundaries that together spanned the entirety of the 0-100% confidence scale. The created tool thus includes the four phrases and their ranges. Understandings of verbal confidence statements are shared and quantifiable, facilitating common ground for reporting and interpreting eyewitness identification confidence.

Chapter 2

Confidence Lexicon: An Evidence-Based Approach for Interpreting Eyewitness Confidence

Alternative methods for obtaining confidence may be more effective than extant verbal or numeric methods. For example, asking eyewitnesses to express their confidence by selecting from a series of statements resulted in similar performance to verbal-only and numeric scale-only approaches (Mansour, 2020). However, people are generally hesitant to use others' definitions of verbal phrases (Budescu & Wallsten, 1990). Combined formats requesting or providing as options both numeric and verbal estimates, offer a solution. However, research in other fields highlights inconsistencies when combined formats are used in practice (Mandel & Irwin, 2021). For the general public, the translation of vague verbal statements into numbers may not be feasible due to individual differences in exposure (i.e. use of numbers in daily life) and cognitive abilities (e.g., numeracy skills). More importantly, Mandel and Irwin (2021) suggest that combined formats do not have an advantage over numeric-only probability estimates (cf. Tekin, Lin, & Roediger, 2018).

Visual analogue scales may eliminate variability in translations and overcome the shortfalls of combined formats but testing of visual scales necessitates further investigation. For example, icon arrays have been shown to improve comprehension by less numerate end users (Galesic et al., 2009) and by those with low graph literacy (Okan et al., 2015; Mandel, Wallsten, & Budescu, 2019). Previous work indicates that confidence ratings obtained using a scale predict accuracy, irrespective of the type of scale presented (Dobolyi & Dodson, 2016). However, we know little about the extent to which different approaches convey the *intended* meaning of a confidence statement. This is especially important given the necessity for transparency when asking others, such as triers of fact, to make judgements of guilt based on eyewitness evidence. For example, mock jurors judge eyewitnesses to be significantly less confident and the accused less likely to be guilty when exposed to verbal compared to a numeric confidence judgement. Belief of and subsequent evaluations of the credibility of eyewitnesses may be lower when confidence is verbal (and liable to misinterpretation) versus numeric. While the ability to distinguish between accurate and inaccurate eyewitnesses does not seem to be compromised when evaluators underestimate (i.e. misinterpret) verbal confidence statements (Smalarz, Yang, & Wells, 2021), we do not know the extent to which this variability in interpretation affects assessments of eyewitness evidence in practice, particularly for individual cases. For example, does misinterpretation occur at all levels of evaluation (e.g., police officer, jurors, judges, general public)? When do evaluators overestimate eyewitness confidence? When does diagnostic utility become compromised because of misinterpretation? Administration of justice must be predictable, consistent, and conducted to a high standard. To make eyewitness evidence reliable, the legal system should seek to eliminate confounding factors, such as the potential for systematic misinterpretation of eyewitness confidence.

A growing body of literature suggests that high confidence, when obtained under pristine conditions, indicates high accuracy (e.g., Wixted & Wells, 2017; Carlson et al., 2017; Wixted et al., 2016, Palmer et al., 2013; Dodson & Dobolyi, 2016). Ironically, we do not know what constitutes “high confidence” (or “low” confidence, or “medium” confidence for that matter). Wixted and Wells (2017) suggest that “it is visually apparent that in most cases, high confidence accuracy is very high (95%-100% correct), whereas low-confidence accuracy is obviously lower” (Wixted & Wells, 2017, p. 30). Some research considers high confidence to imply anything above 80% (Smith et al. 2021; Mansour, 2020; Brewer et al., 2002), and anything under 50% to imply low confidence (Mansour, 2020; Brewer et al., 2002). Other research considers “high confidence” as ratings of 90% or above (Wixted & Wells, 2017; Wixted et al., 2015), “medium confidence” to range from 70-80%, and “low confidence” to refer to ratings of below 60% (Wixted et al., 2015). One of the main caveats to the usefulness of confidence in research and in practice is the associated subjective judgement of “what constitutes sufficient evidence” (Sauer, Palmer, & Brewer, 2019, p. 44). The limitations of variability in interpretation and subjective judgements of evidence, as I will outline below, are likely even more pronounced in practice. Before we can determine if “high confidence” predicts “high accuracy” (under pristine and/or other conditions), we need to operationally define what constitutes “high”.

Our previous work suggests that eyewitnesses should only provide one confidence judgement (see Chapter 1). Given that verbal confidence statements are currently used in practice and generally preferred (Chapter 1; Dodson & Dobolyi, 2015; Budescu, Karelitz, & Wallsten, 2003; but cf. Smalarz et al., 2021; Kenchel et al., 2021), research aimed at improving the communication of verbal confidence statements could improve the administration of criminal justice. Reducing the variability in interpretations of verbal confidence while taking individual preferences and best practices for obtaining confidence into consideration can maximize the utility of such an approach. Thus, a key question we aim to address is how verbal confidence statements can be interpreted to minimize the miscommunication of eyewitness confidence.

Only one published study has used empirical means to interpret verbal confidence judgements. Behrman & Richards (2005) categorized 35 verbal confidence statements obtained from real eyewitnesses into low, medium, or high confidence based on 0 (No confidence) to 10 (Absolutely certain) ratings by participants. While this work provides a preliminary coding scheme for verbal confidence statements, we do not know the *intended* meaning of verbal confidence judgements. Behrman and Richards evaluated confidence statements made by real eyewitnesses, but those eyewitnesses did not translate their verbal statements onto the numeric scale later used by their participants to rate the verbal statements. However, it is worth noting that recently Mansour and Vallano (2022) replicated Behrman and Richards' findings. Using the same approach but based on new participants' ratings, Mansour and Vallano assigned 34 of the 35 statements to the same category (low, medium, or high confidence) as Behrman and Richards.

Other fields, such as climate science and intelligence analysis, use lexicons to communicate probabilities in verbal and numeric ranges in a way that minimizes opportunities for miscommunication. Ho, Budescu, Dhami & Mandel (2015) initially identified inconsistencies in the communication of uncertainty in climate science. According to Ho et al., these inconsistencies are rooted in the differences in preference when people, especially scientists versus the general public, are asked to express and interpret decisions of uncertainty. Ho et al. suggest that people believe they share the same interpretations as their peers, thus leading to miscommunications. Their findings indicate that evidence-based lexicons outperform lexicons developed by practitioners in the United Kingdom and in the United States. To develop their evidence-based lexicons, Ho et al. produced membership functions for common probability terms—essentially, they empirically established a separate numeric range for each term. The approach of Ho et al. is more nuanced than Behrman and Richards' (2005) as it allows for overlap in meaning between terms, which may in practice occur (e.g., one person may consider "pretty sure" as being between 60% and 80% confident whereas another might feel "pretty sure" spans 50% to 65%). Most importantly, we do not know the extent to which a verbal statement may represent a range of probability values. In sum, the interchangeability between terms used by eyewitnesses to express confidence has not been empirically tested—does "pretty sure" always imply medium confidence, for example?

We (i.e. myself and supervisor) sought to test the utility of this evidence-based approach in the eyewitness context, extending the existing literature on the use of lexicons as well as introducing this approach to the eyewitness area. To do this, we selected a set of phrases that were commonly reported by participants in eyewitness experiments where confidence was collected in the participant's own words (Mansour, 2020). We applied the membership function approach for developing a lexicon used by Ho et al. (2015) to these phrases. Based on the findings of Ho et al., we hypothesized that some phrases would indicate a range of values that represent that probability concept (i.e. have distinct membership functions). We further hypothesized that phrases indicative of extreme levels of confidence (e.g., *very confident, not very confident*) would have more narrow membership functions than those indicative of middle-range confidence (e.g., *pretty sure*). We also expected that there would be distinct probability peaks for terms that have distinct membership functions. Thus, we hypothesized that there would be a rank order for the phrases based on the range of probabilities they represent. We expected that, based on the distinct membership functions, it would be possible to select a subset of phrases that have membership functions spanning a full 0-100% probability range. We also hypothesized that the membership functions for some phrases would almost fully overlap, indicating interchangeability (synonyms). Specifically, we expected *somewhat confident* and *moderately confident*, and *quite confident* and *pretty confident* to overlap.

Based on previous work (Dhami, 2018; Ho et al., 2015; Renooij & Witteman, 1999; Zimmer, 1983; Merz, Druzdzel, & Mazur, 1991), we expected to be able to abbreviate the selected number of phrases (likely between four and eight) given their distinct membership functions and interchangeability. Moreover, we hypothesized that our lexicon (i.e. the selected number of statements) would include at least some of the 35 statements identified as commonly used by real eyewitnesses in Behrman and Richards' (2005) Table 1. We also hypothesized that the lexicon would include the most frequently used phrases from our prior work (Pennekamp & Mansour, 2021; Mansour, 2020; i.e. *fairly confident*, *pretty confident*, *very confident*, *not very confident*). Furthermore, we expected the terms selected for our lexicon to be similar to those that are used in climate science lexicons (i.e. *very unlikely, unlikely, likely, very likely*; IPCC reports, Ho et al., 2015).

Finally, we hypothesized that a validation sample (Study 2) would result in a similar lexicon (i.e. similar or the same number of terms and include nearly all the same or the same terms. Synonyms in the development sample (Study 1) were expected to also be synonyms in the validation sample). Our methods, hypotheses and analyses were pre-registered on the Open Science Framework and can be found via the following link: https://osf.io/dbncz/?view_only=94461dd285b44b728f7cc85857475a5e. Both studies were approved by our university's research ethics board.

General Method

Design

This study used a within-subjects design whereby all participants provided 11 ratings for each of 13 verbal confidence phrases, which are described in more detail below. The order that the 13 phrases were presented to participants was randomized.

Materials

The study was programmed on Qualtrics. The phrases chosen for the lexicon were the confidence statements that were most frequently provided by participants in our prior research (Mansour, 2020; see Table 3). The nature of the task is described in more detail below.

Table 3

Most Frequently Provided Confidence Phrases in Mansour (2020)

Phrase
Not very confident
Not sure
He/She resembles the criminal*
He/She looks familiar*
I think it is him/her*
Moderately confident*
Somewhat confident
He/She looks like the criminal*
Pretty sure*
Fairly confident*
Quite confident
Confident*
Very confident*

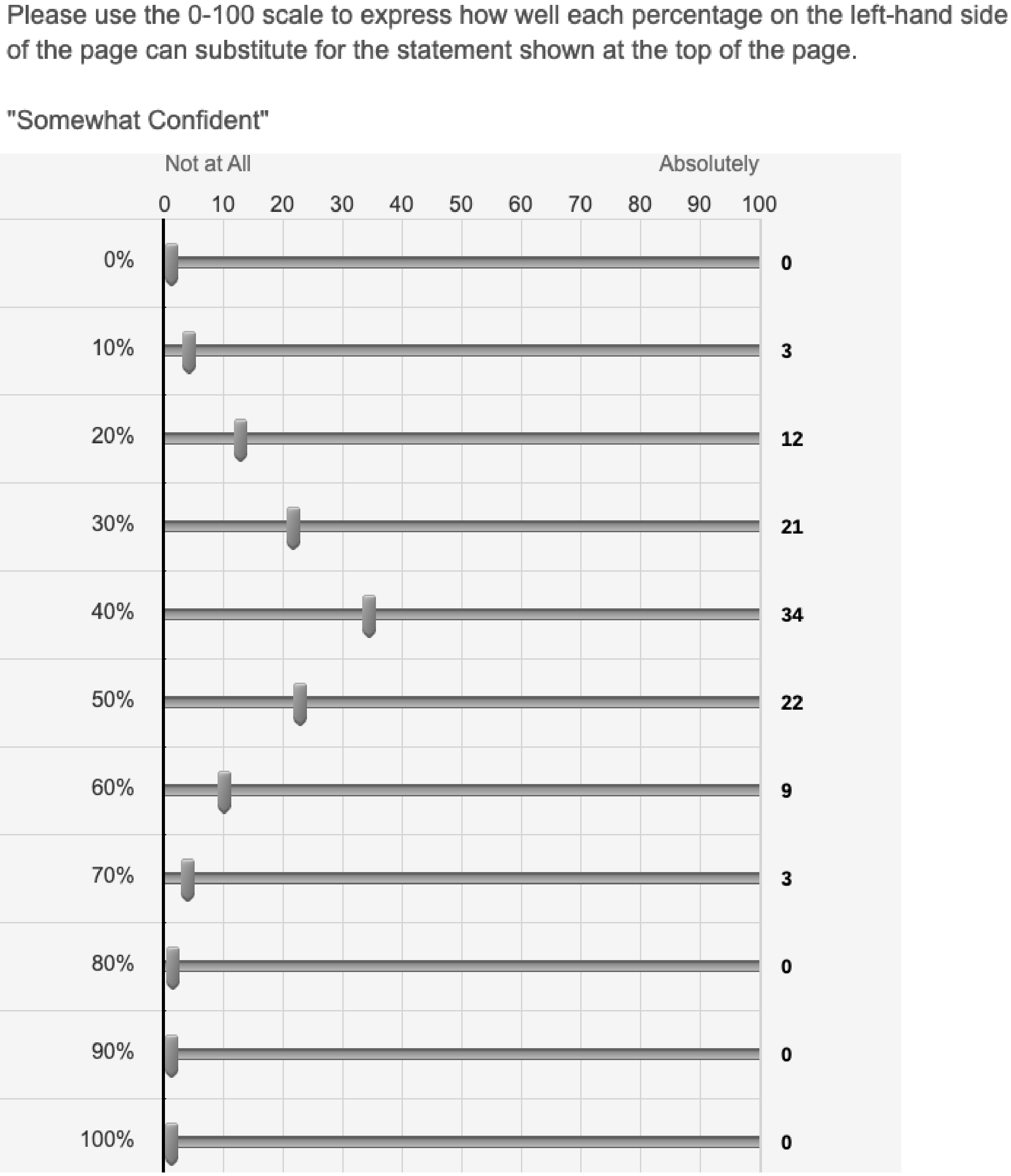
Note. Most frequently provided verbal confidence phrases by participant-eyewitnesses in Mansour (2020), *phrases reported as frequently used by real eyewitnesses in Behrman and Richards (2005), Table 1

Procedure

After providing informed consent, all participants read instructions for the task. They then completed three practice trials to ensure they understood what they were asked to do. Each practice trial was followed by a feedback screen that explained how the participant could have responded. Participants were then asked to rate how well each phrase was represented by 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. That is, participants rated how well each percentage value (e.g., 60%) represents each phrase (e.g., moderately confident) on a 0-100 scale (from “Not at all” to “Absolutely”; see Figure 8. Participants were asked to give 11 representativeness ratings for each of the 13 phrases.

Figure 8

Representativeness Rating Task



Please use the 0-100 scale to express how well each percentage on the left-hand side of the page can substitute for the statement shown at the top of the page.

"Somewhat Confident"

0%	0
10%	3
20%	12
30%	21
40%	34
50%	22
60%	9
70%	3
80%	0
90%	0
100%	0

Note. Example of the representativeness rating task for a phrase completed by one participant. Participants rated how well each of the 11 percentage values represented a phrase (from Not at All to Absolutely). For this participant, the phrase "somewhat confident" was best represented at 40%.

Measures

Representativeness

The key measure in our study was the value each participant assigned for how well a specific phrase was represented by each of the 11 percentages. Statements were presented with a limited context (i.e. "phrases given by eyewitnesses in response to lineups"). That is, phrases were presented without provision of the full, original statement obtained from eyewitness-participants (e.g., "I think I got the right guy, I'm fairly confident").

Data Quality Checks

There were two attention checks (e.g., "The water is freezing", i.e. "cold") and one manipulation check (i.e. "We asked you to evaluate judgements reported by who?" Multiple choice answer: "Eyewitnesses"). Participants were excluded if they failed both attention checks. Few participants (*n* = 4) failed the manipulation check indicating that they did not understand they were evaluating eyewitness judgements. Because we did not specify that we would exclude participants that failed the manipulation check, those that failed it were still included in the data analyses. We did not exclude these participants on the basis that they passed two attention checks.

Analytic Approach

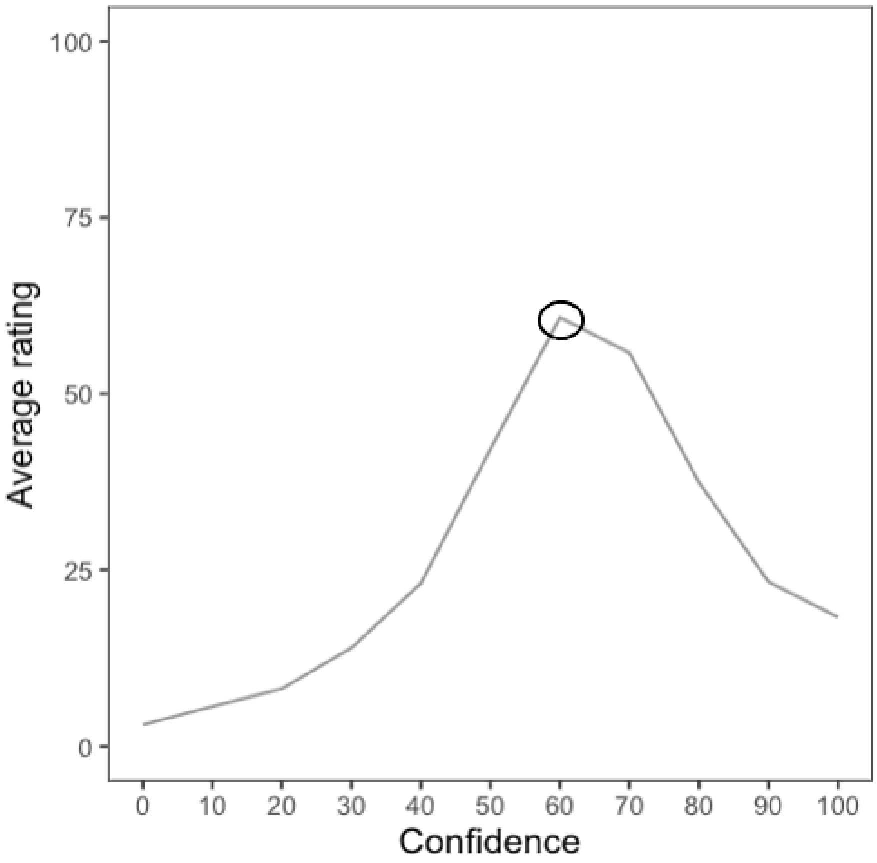
Membership functions

We aimed to quantify the meaning of the phrases in the phrase set by means of the membership functions (Wallsten & Budescu, 1990). A membership function indicates how well numerical values (0%, 10%, etc.) represent the probability expressed by a particular phrase. As described earlier, participants were asked to indicate how well each probability (0-100% in 10% increments) represents each phrase on a 0 to 100 scale. The score given to each increment for each phrase constitutes the membership value. A membership of 0 suggests the increment substitutes “not at all” while a membership of 100 indicates that the increment “absolutely” substitutes for the phrase. Assigning a value of 100 indicates that an increment is at the peak of the phrase’s membership function. Therefore, for the phrase *very confident*, we might expect the increment 0% to receive a 0, while the increment 10% might receive a slightly higher value such as 5, while 90% might receive a score of 100. Membership functions were calculated for each person for each phrase and then the membership functions for each phrase were averaged across the sample of participants to obtain a mean membership function for each phrase (i.e. we created average membership functions).

To create the lexicon, we analyzed our data in four steps: 1) determining the phrase peak, 2) determining phrase boundaries, 3) determining phrase synonymity, and 4) abbreviating the resulting lexicon.

1) Determining Peaks. We began visualizing our data by creating 13 histograms (i.e. one histogram for each phrase), where the bars in each histogram were the mean peak values across participants for a particular percentage increment (see Figure 9). We visually identified peaks and potential overlap between phrases. The percentage increment with the highest mean value (i.e. the peak value) was considered the value most representative of the phrase.

Figure 9
Histogram for a Phrase



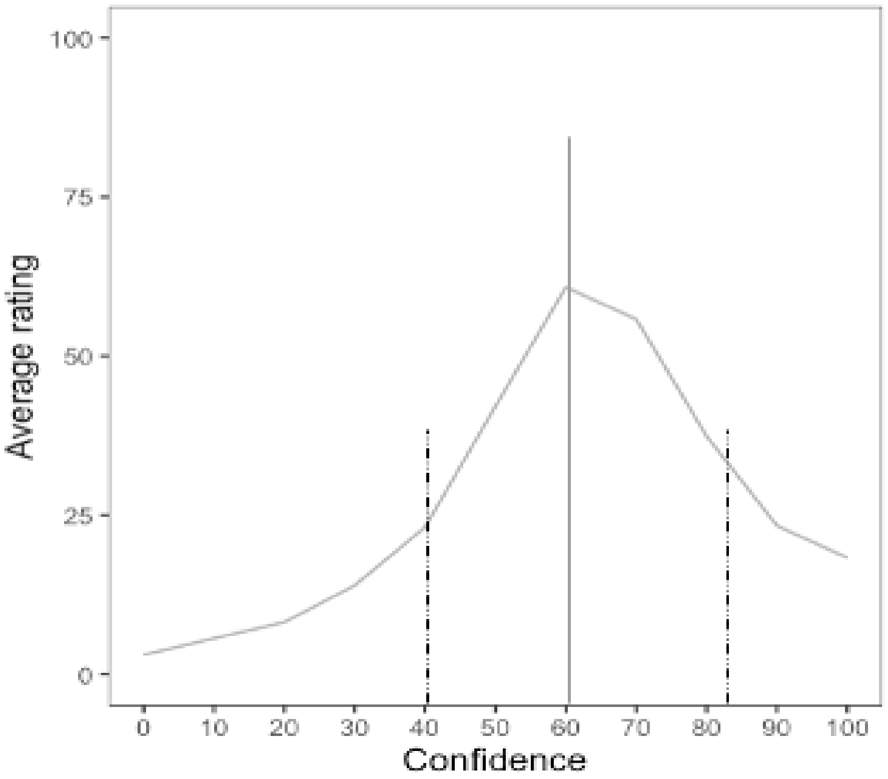
Note. Weighted mean peak value for a phrase (somewhat confident). Across the sample, somewhat confident was rated to be best represented at 60%. Points on the graph represent mean values for each increment (0%, 10%, …, 100%).

We also rank ordered all phrases using weighted mean peak values. The weighted mean indicates how well each numeric increment represents the phrase (i.e. the average number of votes for each increment as a representation of the phrase). More votes mean an increment better represents the phrase.

2) Determining Phrase Boundaries. For this step, we determined the range of the majority of the histogram's density, and therefore what percentage range best captured the phrase's range. We used 64% (i.e. one standard deviation in a normal distribution) as the majority percentage. To determine where the middle 64% of the density lay, we calculated weighted standard deviations. For each phrase, we created distribution curves using a probability density function with the weighted means to see at what increment of percentage the majority of votes lie (see Figure 10).

Figure 10

Example of a Distribution Curve for an Average Membership Function



Note. Example of a distribution curve for a mean membership function for one of the 13 phrases (somewhat confident). The solid vertical line represents the weighted mean and the dotted vertical lines surround the density representing one weighted standard deviation.

Cut offs. Ho et al. (2015) determined the cut off points between adjacent phrases by identifying the region of values for which the membership function of a given phrase was higher than all other phrases (see Ho et al.'s Figure 1, right panel for a clear illustration of this). We plotted average membership functions of all phrases to identify the region of values for which a membership function of one phrase was higher than the next. For phrase pairs to be deemed synonymous, the mean peaks of membership functions had to be in the same cut-off range.

Determining Phrase Synonymity. There are no statistical tests to compare membership functions. Ho et al. (2015) determined synonymity by visually comparing average membership functions and the numeric ranges they represented. However, Ho et al. visually examined membership functions for only three phrase pairs (remote chance/very unlikely, probably/likely, and very likely/almost certain). Since our visual comparison was going to consider a larger number of phrases, we also attempted to calculate the percentage of overlap in a more objective way. That is, we calculated the shared area under the curves for any two adjacent mean membership functions. We calculated a percentage of overlap between pairs of phrases by comparing the shared area under the curves to the total area under the curves.

3) Reducing the Number of Phrases. We separately examined all 13 phrases based on the range of each phrase to determine which eight phrases had the narrowest distributions. We calculated lower bounds and upper bounds for each phrase to determine the range of interpretation for each phrase. Upper bounds were defined as the highest rating of the phrase within the sample. Lower bounds were defined as the lowest rating of the phrase within the sample. A larger span (i.e. greater distance from lowest point to highest point) meant that phrases were more difficult to interpret. A narrow span (i.e. smaller distance from lowest point to highest point) meant that phrases were easier to interpret. We eliminated the five phrases with the largest ranges from this selection to establish a set of eight phrases with the narrowest ranges.

We then examined all 13 phrases based on their weighted standard deviations to determine which eight phrases had the smallest standard deviations. A lower standard deviation indicates that a phrase has a clearer meaning. The larger the standard deviation, the less clear the meaning of a phrase. We eliminated the five phrases with the largest standard deviations from this selection to establish a set of eight phrases with the lowest standard deviations.

4) Abbreviating the Lexicon. Finally, we compared the eight phrases with the narrowest distributions to the eight phrases with the lowest standard deviations to determine which phrases (including synonyms) would meet the criteria for our abbreviated lexicon. The criteria were that

a. The set of phrases must have membership functions that together clearly span 0-100%, such that peaks occur across the span.
b. Phrases should have average membership function peaks that are distinct from one another. When the average membership functions essentially overlap, the phrases will be treated as synonyms.

Study 1 (Development Sample)

Participants

Participants were adults with sufficient visual capacity to view a computer screen. Participants (*N* = 40) were recruited via CloudResearch (Amazon Mechanical Turk). The usable sample (*n* = 37) did not include duplicate IP addresses (*n* = 2) or cases where the participants failed 3 of the 3 attention checks (*n* = 0), indicated to have cheated (*n* = 1), or did not provide a rating (0-100) for all 11 of the increments (i.e. 0%, 10%, etc.) for a particular phrase. The usable sample of participants identified as primarily male (56.76%), white (72.97%), Asian (10.81%), black (8.11%), Hispanic (2.70%) and other (5.41%) with a mean age of 42.70 years (SD = 11.07, Range = 27-73).

Results

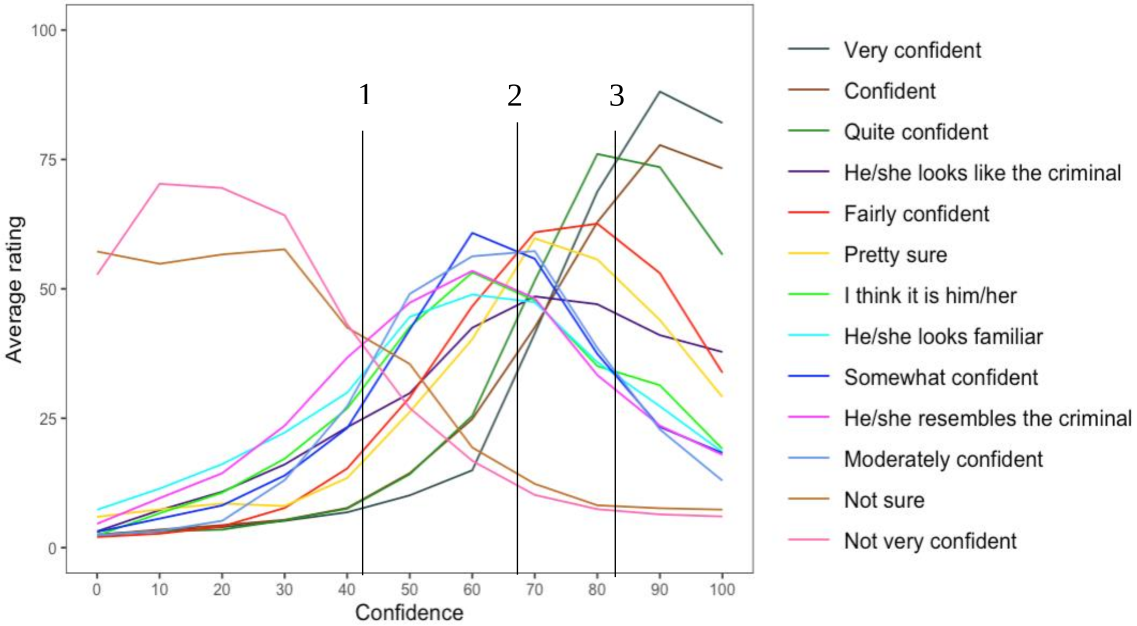
All 13 phrases had visually distinct probability peaks. Visual inspection suggested there were four sets of phrases that overlap. The histograms for eight phrases were primarily overlapping, potentially indicating synonymity for medium confidence.

Cut offs

We determined the cut off points between adjacent phrases by identifying the region of values for which the membership function of a given phrase was higher than all other phrases, following Ho et al. (2015) and our pre-registered plan. The first cut off was at 43%, the second at 68% and the third cut off at 83% (see Figure 11).

Figure 11

Average Membership Functions for all 13 Phrases in Study 1



Note. The vertical lines depict the cut offs used to determine synonymity.

1) Determining Peaks

Mean peak values. Using the mean peak values, we established a rank order for the sample (low to high confidence). Table 4 shows this ordering. As expected, all membership functions had numerically distinct peaks.

Table 4

Rank Ordered Phrases in Study 1 Based on Weighted Means

Phrase	Weighted *M* peak values	Weighted *SD*	Lower bounds	Upper bounds	Range
Not very confident	28.34	23.21†	5.13	51.55	46.42*
Not sure	30.68	24.79	5.79	55.37	49.58
He/She resembles the criminal	57.85	23.49	34.36	81.34	46.98
He/She looks familiar	58.01	24.84	33.17	82.85	49.68
I think it is him/her	61.73	22.69†	39.04	84.42	45.38*
Moderately confident	62.09	19.98†	42.11	82.07	29.96*
Somewhat confident	62.21	21.50†	40.71	83.71	43.00*
He/She looks like criminal	66.31	24.49	41.82	90.80	48.98
Pretty sure	67.89	23.61	44.28	91.50	47.22
Fairly confident	71.18	20.31†	50.87	91.49	40.62*
Quite confident	77.59	19.86†	57.73	97.45	39.72*
Confident	78.77	20.91†	57.86	99.68	41.82*
Very confident	80.93	20.11†	60.82	100	39.18*

Note. *indicates set of eight phrases with the smallest ranges. † indicates set of eight phrases with smallest standard deviations.

2) Determining Phrase Boundaries

Comparison of overlap for adjacent phrase pairs. To calculate the overlap (%) between phrase pairs, we calculated the shared area under the curve between two adjacent mean membership functions. Overlap was high (>80%) for nine phrase pairings (see Table 5).

Table 5

Overlap Between Adjacent Phrase Pairings in Study 1

Phrase 1	Phrase 2	Overlap (%)
Moderately confident*†	Somewhat confident*†	92.12
He/she resembles the criminal	He/she looks familiar	91.17
He/She looks familiar	I think it is him/her*†	89.54
Quite confident*†	Confident*†	89.09
Not very confident*†	Not sure	86.70
I think it is him/her*†	Moderately confident*†	86.57
He/she looks like the criminal	Pretty sure	83.13
Confident*†	Very confident*†	81.65
Somewhat confident*†	He/she looks like the criminal	80.06
Fairly confident*†	Quite confident*†	66.95
Pretty sure	Fairly confident*†	53.74
Not sure	He/She resembles the criminal	34.75

Note. Overlap represents the shared area under the curves between two adjacent phrases. *indicates phrases included in our set of eight smallest ranges and †smallest standard deviations.

Synonyms. To be considered synonymous, phrases' mean peak values had to be within the same cut off area. When phrases overlapped with one or more phrases, we chose the "best match" (i.e. highest overlapping phrase match). We found there to be six sets of synonyms in Study 1 (see Table 6).

Table 6

Synonyms in Study 1

Phrase 1	Phrase 2
Moderately confident*†	Somewhat confident*†
Confident*†	Very confident*†
He/she resembles the criminal	He/she looks familiar
Fairly confident*†	Quite confident*†
Not very confident*†	Not sure
He/she looks like the criminal	Pretty sure

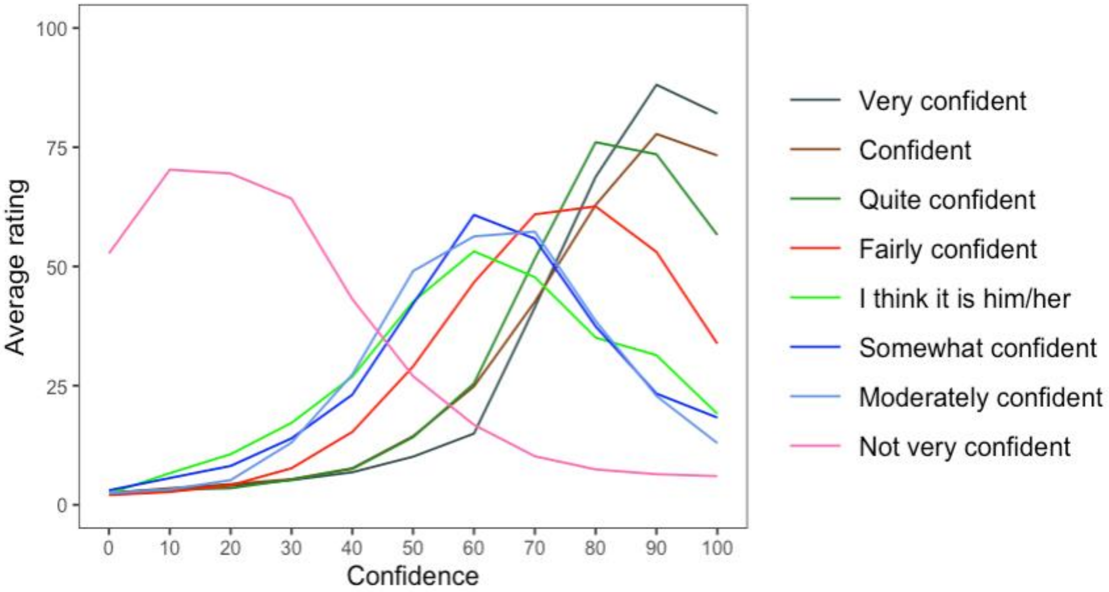
Note. Synonyms in Study 1. Phrases were adjacent with mean peaks in the same cut off area and had the highest overlap compared to other phrases.
*indicates phrases included in our set of smallest ranges and †smallest standard deviations.

3) Reducing Number of Phrases

Ranges. Based on the size of ranges (distance from lower bounds to upper bounds), we chose the eight phrases with the narrowest ranges (see Figure 12). The selected phrases are starred in Table 4.

Standard deviations. Based on the lowest weighted standard deviations, we chose the eight phrases had the least variation in their interpretation (see Figure 12). The eight phrases with the lowest standard deviations and narrowest ranges are identified with a cross in Table 4.

Figure 12
Eight Phrases Meeting our Criteria for Inclusion in the Lexicon

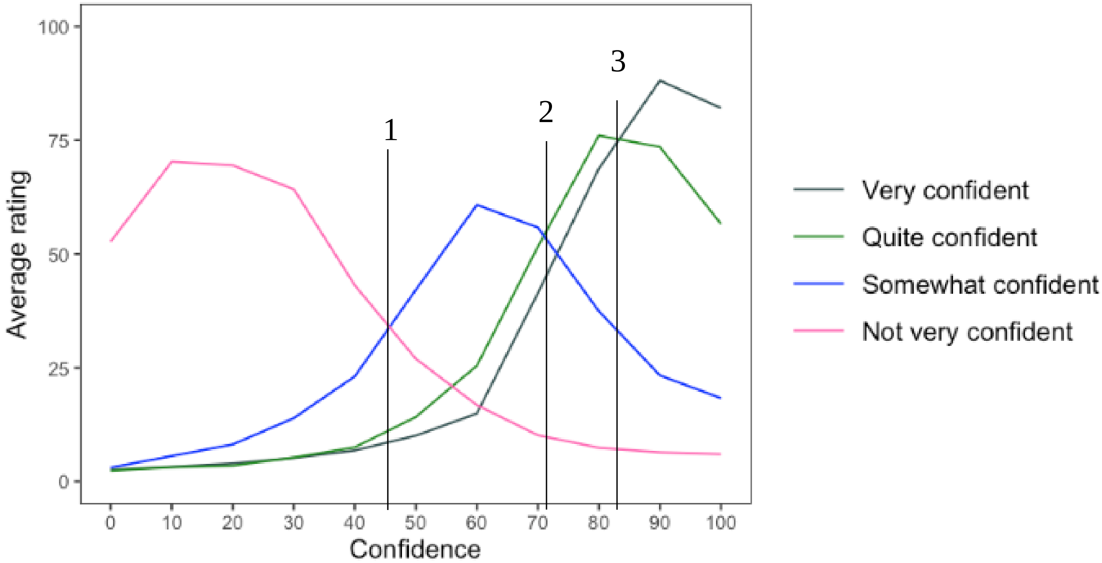


4) Abbreviating the Lexicon

There were four phrases that spanned 0-100%, such that distinct peaks occurred across the span. These phrases were also part of the eight narrowest spans (i.e. least difficult to interpret) and part of the eight lowest weighted standard deviations. The phrases that met these criteria were *not very confident, somewhat confident, quite* *confident* and *very confident*. We again determined cut offs by identifying the crossover between one membership function into the next (see Figure 13).

Figure 13

Abbreviated Lexicon From Study 1



Note. The four phrases with membership functions for phrases that had a) unique peaks, b) narrow membership function, c) small standard deviations. Black lines represent cut offs.

Synonyms in the abbreviated lexicon. Phrases that did not meet these criteria (i.e. phrases that did not span 0-100 with distinct peaks occurring across the span) but that appeared in both sets of 8 and were synonymous with one of the phrases meeting the abbreviated lexicon criteria were considered synonyms in our abbreviated lexicon. We compared the standard deviations and ranges of synonyms and phrases that met the criteria to decide which phrases were presented as synonyms (versus core phrases). Phrases that had comparatively lower standard deviations and narrower ranges were considered as core phrases in the abbreviated lexicon (versus synonyms).

The phrases considered synonyms in our abbreviated lexicon were *moderately confident* (synonymous with *somewhat confident*), *confident* (synonymous with *very confident*) and fairly confident (synonymous with *quite confident*).

Discussion

Building on the work of Ho et al. (2015), we constructed a lexicon to communicate and interpret eyewitness confidence by empirically determining the numeric ranges, overlap, and cut offs for 13 commonly used verbal confidence phrases. First, visual examination indicated overlap that resulted in four sets of phrases. Based on each phrase's distinct weighted mean peak value, we were able to establish a rank order for all of the phrases (low to high). Then in our second step, we reduced the phrase set based on their membership function narrowness and size of standard deviations. We identified the eight phrases with the narrowest ranges and the smallest standard deviations. Phrases included in the final lexicon had to be part of each set of eight and represent a distinct numeric range on a scale from 0-100%. We determined synonymity in our third step. To do this, we calculated the shared area under the curves between adjacent membership functions. If overlapping phrase pairs had mean peak values in the same cut off area, we considered them to be synonymous. And in our final step, we abbreviated the lexicon from eight to four phrases. The final lexicon included *not very confident, somewhat confident, quite confident* and *very confident*. Three phrases (*confident, moderately confident, fairly confident*) were considered synonyms as they overlapped with a phrase in the final four, were also included in each set of eight and had mean peak values in the same cut off range.

We hypothesized that some phrases would indicate a range of values that represent a probability concept (i.e. have distinct membership functions). This hypothesis was supported. In Study 1, each phrase had a distinct membership function (i.e. numerically distinct weighted means). Thus, participants given the context of eyewitness confidence completed the task in a similar way to participants who were given the context of climate science (Ho et al., 2015). However, as the phrases examined were different, the samples are not particularly comparable. Nonetheless, our final lexicon resembles the one produced by Ho et al. in that our final lexicon includes four phrases.

We also hypothesized that phrases indicative of extreme levels of confidence (e.g., *very confident, not very confident*) would have more narrow membership functions than those indicative of middle-range confidence (e.g., *pretty sure*), consistent with the findings of Kenney (1981). This hypothesis was partially supported. *Very confident* had the narrowest membership function compared to all other phrases. While *not very confident* represented a distinct numeric range on a scale from 0-100% and was included in the set of eight narrow membership functions in the sample, its membership function was not narrower than *moderately confident, fairly confident, somewhat confident, quite confident* or *confident*. There is variation in how the public interprets high versus low probabilities. Previous work suggests that the public interprets expressions of probability in a regressive manner (i.e. they underestimate high probabilities and overestimate low probabilities; Budescu, Broomell & Por, 2009). It may be that individuals vary more in their interpretation of phrases representing “low” confidence than “medium” confidence. Similarly, individuals do not interpret verbal “low” confidence as such when asked to interpret such statements numerically (Mansour, 2020). We found that 15% of judgements that would have been considered low based on the eyewitness' scale rating were judged as medium confidence by mock-jurors. It may thus be difficult for people to attach precise meanings (numbers) to "low confidence" phrases.

We hypothesized that there would be a rank order for the phrases based on the range of probabilities they represent. This hypothesis was supported. All 13 phrases had distinct numeric peaks, allowing us to rank order the phrases based on the mean peaks they represented. Establishing a rank order is important: although British criminal justice systems currently do not request confidence judgements, having ranking information can be helpful and informative when evaluating spontaneous expressions of eyewitness confidence. And for jurisdictions that do request eyewitness confidence, ranking information can assist in judgements of the relative likelihood of accuracy of an eyewitness, particularly when multiple eyewitnesses exist.

Based on lexicons assessed by Dhami (2018) and Ho et al. (2015), we expected that, it would be possible to select a subset of phrases that have membership functions spanning a full 0-100 probability range. This hypothesis was supported. The abbreviated lexicon consists of four phrases (not very confident, somewhat confident, quite confident, very confident). We also hypothesized that the membership functions for some phrases would almost fully overlap, indicating interchangeability (i.e. there would be synonyms). Specifically, we expected *somewhat confident* and *moderately confident*, and *quite confident* and *pretty sure* to overlap. This hypothesis was partially supported. *Somewhat confident* and *moderately confident* were rated to be synonymous. We did not find evidence for synonymity between *quite confident* and *pretty sure*. We will discuss these findings further in the General Discussion.

Study 2 (Validation Sample)

We aimed to replicate Study 1's findings with a second sample. If the abbreviated lexicon from Study 1 is reliable, we would expect a second sample to show similar interpretations of phrases and for the abbreviated lexicon to contain the same four phrases. The pre-registration for this study contained the same hypotheses as for Study 1 is available at 10.17605/OSF.IO/5BR6S.

Participants

In the second study (*N* = 32), participants were recruited via SurveyCircle, social media, and word of mouth. The usable sample (*n* = 30) did not include duplicate IP addresses (n = 1), participants who did not provide consent to participate (*n* = 1) or cases where the participants failed 3 of the 3 attention checks (*n* = 0), indicated to have cheated, or did not provide a rating (0-100) for all 11 of the increments (i.e. 0%, 10%, etc.) for a particular phrase. The usable sample of participants identified as primarily female (63.3%), Caucasian/white (50%), British (16.67%), mixed (10%), Asian (6.67%), German (6.67%), Dutch (3.33%), Australian (3.33%), Italian (3.33%), with a mean age of 29.67 years (SD = 9.62, Range = 19-60).

Results

Again, and as hypothesized, all 13 phrases had visually distinct peaks. Visual inspection suggested there were five sets of phrases that overlap. The histograms for eight phrases near the midpoint of the confidence scale were primarily overlapping, potentially indicating synonymity for medium confidence.

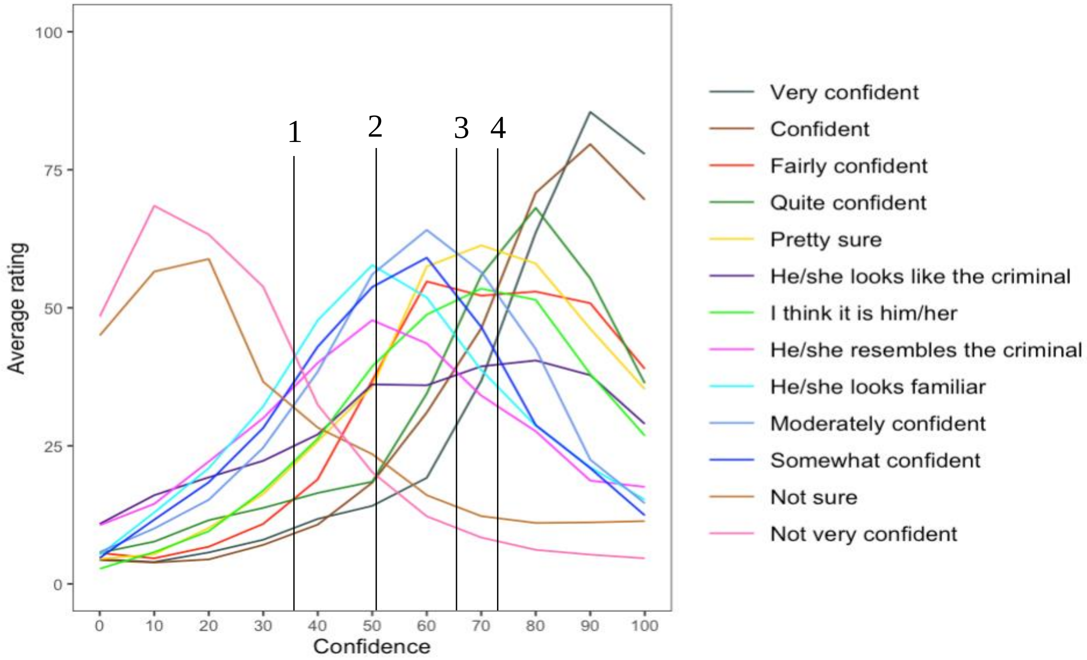
Cut offs

We again determined the cut off points between adjacent phrases by identifying the region of values for which the membership function of a given phrase was higher than all other phrases. Cut off 1 was at 36%, cut off 2 at 52% and cut off 3 at 66% and cut off 4 at 73%.

These cut offs differed slightly from Study 1 in that there were four distinct cut offs, separating the phrases into five sets (instead of four; see Figure 14).

Figure 14

Average Membership Functions for all 13 Phrases in Study 2



Note. The black vertical lines depict the cut offs used to determine synonymity.

1) Determining Peaks

Mean peak values. Again, all 13 phrases had distinct mean peak values. We were thus able to establish a rank order (low to high) for the validation sample like the rank order of mean peak values in Study 1. The rankings were the same as Study 1 except for three phrases representing medium confidence (*I think it was him/her, Somewhat confident, He/she looks like the criminal*). *Somewhat confident* and *He/she looks like the criminal* were ranked lower in Study 2 compared to Study 1 while *I think it was him/her* was ranked higher in Study 2 compared to Study 1 (see Table 7).

Table 7

Rank Ordered Phrases in Study 2 Based on Weighted Means

Phrase	Weighted *M* peak values	Weighted *SD*	Lower bounds	Upper bounds	Range
Not very confident	26.68	22.82†	3.86	49.50	45.64*
Not sure	32.16	27.77	4.39	59.93	55.54
He/she resembles the criminal	52.59	25.69	26.90	78.28	51.38
He/she looks familiar	53.70	23.66	30.04	77.36	47.32
Somewhat confident	54.88	22.88†	32.00	77.76	45.76*
Moderately confident	57.57	22.63†	34.94	80.20	45.26*
He/She looks like criminal	59.04	28.05	30.99	87.09	56.10
I think it is him/her	64.75	22.85†	41.90	87.60	45.70*
Pretty sure	66.36	23.06†	43.30	89.42	46.12*
Fairly confident	68.26	23.05†	45.21	91.31	46.10*
Quite confident	69.03	24.55	44.48	93.58	49.10
Confident	76.80	21.81†	54.99	98.61	43.62*
Very confident	78.15	22.78†	55.37	100	44.63*

Note. *indicates set of eight phrases with the smallest ranges. †indicates set of eight phrases with the smallest standard deviations.

2) Determining Phrase Synonymity

Comparison of overlap for adjacent phrase pairs. We again calculated the shared area under the curve between two adjacent mean membership functions. Overlap was high (>80%) for seven phrase pairings (see Table 8).

Table 8

Overlap Between Adjacent Phrase Pairings in Study 2

Phrase 1	Phrase 2	Overlap (%)
He/she looks familiar	Somewhat confident*†	90.38
Somewhat confident*†	Moderately confident*†	89.41
Pretty sure*†	Fairly confident*†	89.25
I think it is him/her*†	Pretty sure*†	88.58
He/she resembles the criminal	He/she looks familiar	87.38
Not very confident*†	Not sure	84.11
Confident*†	Very confident*†	83.38
Quite confident	Confident*†	79.96
Fairly confident*†	Quite confident	78.62
He/she looks like the criminal	I think it is him/her*†	76.77
Moderately confident*†	He/she looks like the criminal	70.21
Not sure	He/She resembles the criminal	40.53

Note. Overlap represents the shared area under the curves between two adjacent phrases. *indicates phrases included in our set of smallest ranges. †indicates phrases included in our set of smallest standard deviations.

Synonyms. To be considered synonymous, phrases' mean peak values had to be within the same cut off area. When phrases overlapped with one or more phrases, we again chose the "best match" (i.e. highest overlapping phrase match). In Study 2, we found there to be five sets of synonyms in Study 2 (see Table 9).

Table 9

Synonyms in Study 2

Phrase 1	Phrase 2
Moderately confident*†	Somewhat confident*†
Confident*†	Very confident*†
He/she resembles the criminal	He/she looks familiar
Not very confident*†	Not sure
I think it is him/her*†	Pretty sure*†

Note. Synonyms in Study 2. Phrases were adjacent with mean peaks in the same cut off area and had the highest overlap compared to other phrases. *indicates phrases included in our set of smallest ranges. †indicates phrases included in our set of smallest standard deviations.

3) Reducing the Number of Phrases

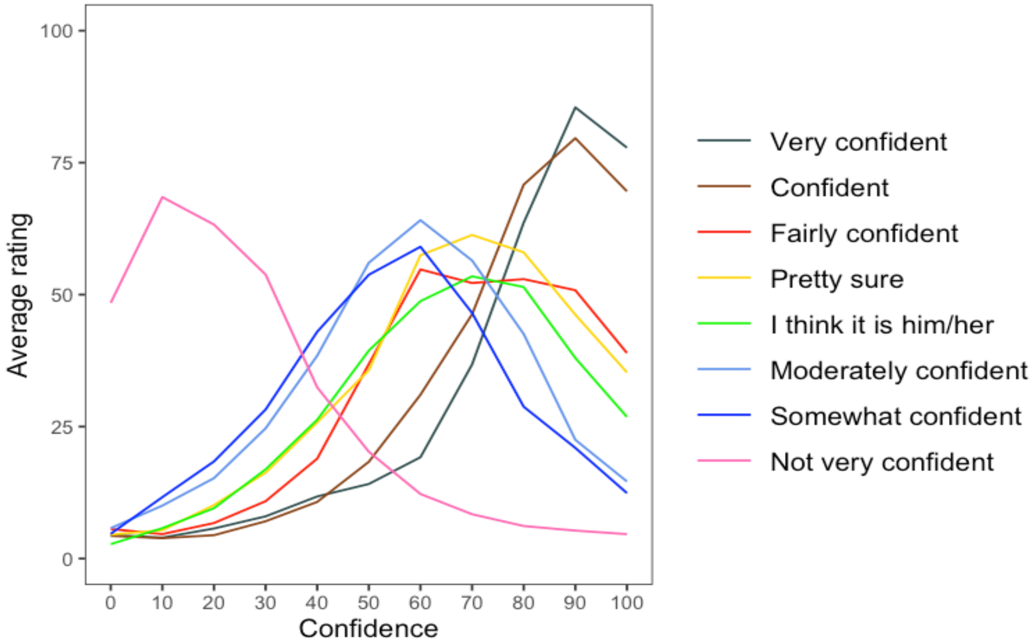
Ranges. We again identified the eight phrases with the narrowest ranges—and therefore the easiest to interpret. The selected phrases are starred in Table 7.

Standard deviations. We again identified the eight phrases with the smallest standard deviations. The eight phrases with the lowest standard deviations and narrowest ranges are identified with a cross in Table 7.

Based on the narrowest ranges and smallest weighted standard deviations, the following eight phrases had the least variation in their interpretation (see Figure 15).

Figure 15

Eight Phrases Meeting our Criteria for Inclusion in the Lexicon



Note. Phrases with membership functions that were part of the set of 8 narrow membership functions and 8 smallest standard deviations.

Compared to Study 1, phrases included in the set of eight narrowest standard deviations were identical except for *pretty sure*. *Pretty sure* was included in the set of eight narrowest standard deviations in Study 2, but not Study 1. Instead, Study 1 included *quite confident* in the set of eight narrowest standard deviations while Study 2 did not.

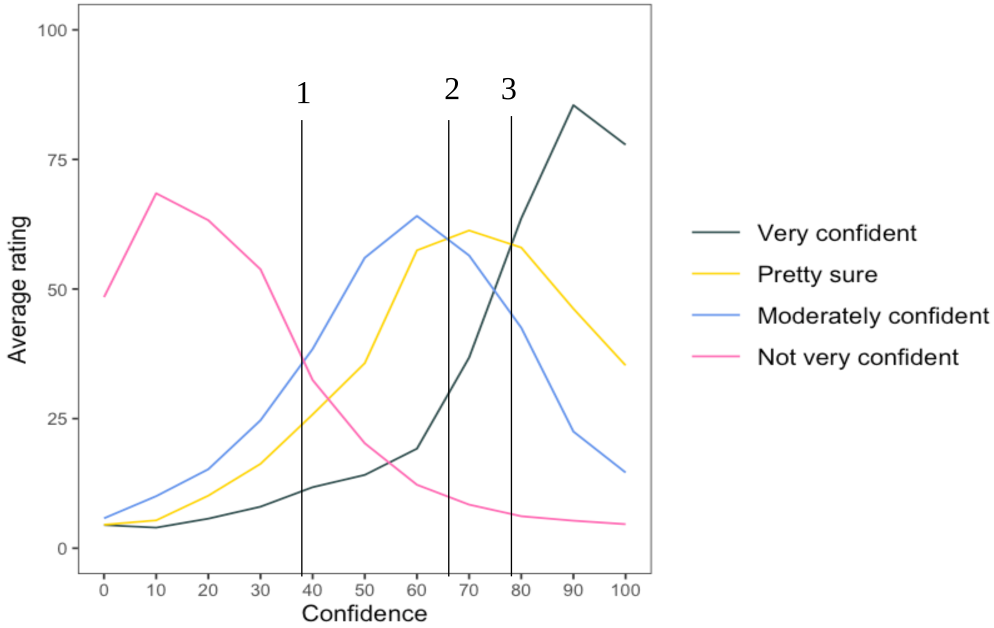
4) Abbreviating the Lexicon

As hypothesized and consistent with Study 1, four phrases met the criteria for our abbreviated lexicon (i.e. had distinct peaks and spanned 0-100%, and had a sufficiently narrow range). The phrases that met these criteria were *not very confident, moderately confident*, *pretty sure* and *confident.* We again determined cut offs by identifying the crossover between one membership function into the next.

Contrary to our expectations, the set was not identical to Study 1. Study 1 included *not very confident, somewhat confident*, *quite confident* and *very confident*, whereas as Figure 16 illustrates, the Study 2 abbreviated lexicon comprised *not very confident, moderately confident*, *pretty sure* and *very confident*.

Figure 16

Abbreviated Lexicon From Study 2



Note. The four phrases with membership functions for phrases that had a) unique peaks, b) narrow membership function, c) small standard deviations. Black lines represent cut offs.

Synonyms in the abbreviated lexicon. In the abbreviated lexicon produced by Study 2, the resultant synonyms were *somewhat confident* (synonymous with *moderately confident*), *I think it is him/her* (synonymous with *pretty sure*) and *confident* (synonymous with *very confident*).

Discussion

We aimed to replicate our findings from Study 1 with a second sample (Study 2). We expected the second sample to show similar interpretations of phrases as Study 1 and for the abbreviated lexicon to contain the same four phrases. This hypothesis was largely supported. The abbreviated lexicon in Study 2 also consisted of four phrases. The phrases differed in part to those in the abbreviated lexicon in Study 1. While *not very confident* and *very confident* were included in the abbreviated lexicons in both studies, *moderately confident* and *pretty sure* were included in Study 2 instead of *somewhat confident* and *quite confident*. However, *moderately confident* and *pretty sure* covered distinct numeric probability ranges, similar to the distinct numeric probability ranges covered by *somewhat confident* and *quite confident* in the abbreviated lexicon derived in Study 1. Like Study 1, *somewhat confident* was synonymous with *moderately confident* in Study 2. We will further discuss these findings in the General Discussion.

Producing the Final lexicon

To establish the final cut offs for the lexicon, we compared cut offs in Study 1 to cut offs in Study 2 (see Table 10). For the final lexicon, cut offs were rounded to the nearest percentile between the two cut off points from each abbreviated lexicon (e.g., cut off 1 in abbreviated lexicon Study 1: 45; cut off 1 in Study 2: 38, cut off 1 in final lexicon: 40). While cut offs were consistent across studies (i.e. three cut offs in each abbreviated lexicon), we decided to round to the nearest percentile for simplicity.

Table 10

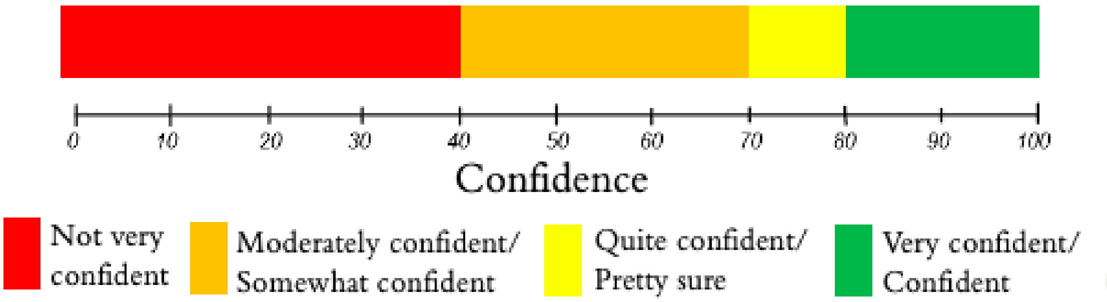
Cut offs between Abbreviated Lexicons in Study 1 and Study 2

	Cut off 1	Cut off 2	Cut off 3
Study 1	45	72	83
Study 2	38	66	78
Final Lexicon	40	70	80

In practice (National Intelligence Council and Defense Intelligence), lexicons range between seven and six categories. Dhami (2018) found that the average analysts' lexicon size is more than the "seven and six-category lexicons" currently used in practice. Given that individuals become eyewitnesses by accident, and thus do not have any judgement-specific expertise, we attempted to produce a lexicon that is as simple as possible—that is, an abbreviated lexicon. Ho et al.'s (2015) abbreviated lexicon included the four probability phrases that were most frequently used in IPCC reports. We had thus hypothesized that our lexicon would contain between four to eight phrases. Based on our abbreviated lexicon criteria and the semantic overlap between phrases, we were able to construct a final lexicon consisting of four phrases, including three synonyms. Core phrases are listed above their respective synonyms (see Figure 17).

Figure 17

Final Lexicon



Note. Final confidence lexicon including cut offs and synonyms.

Evidence-based lexicons in other studies (Wintle et al., 2019; Ho et al., 2015) present probabilities in form of a bar graph (visual scale), including numbers and verbal probability phrases. For our final lexicon (see Figure 17), we decided to present visual, numerical and verbal probability information in addition to a colour-schemed scale. We chose to use red (at the low end), orange, (lower medium confidence), yellow (upper medium confidence) and green (high confidence), similar to a traffic light system. We chose those colors and their ordering because they are likely to be familiar to most people and those colors are used in many countries in the same ways.

We suggest that eyewitnesses can use the final lexicon however they see fit. They could circle a number, a phrase, or simply draw a line. We intended for this lexicon to be as accessible as possible in its use in hopes that it could be used by anyone and everyone (including children and the elderly). After an eyewitness makes their decision, the marked-up lexicon could then be given to a police officer, judge, jury or anybody else that is asked to interpret eyewitness confidence. Our hope is that by utilizing a variety of modes (verbal, numeric, and graphic) and deriving the cut offs and associations between phrases and numbers empirically, the likelihood of miscommunication between parties will be minimized.

Comparing Study 1 and Study 2

Across two studies, we were able to construct an empirically derived lexicon to communicate and interpret eyewitness confidence. In both studies, all 13 phrases had distinct membership functions. All 13 membership functions had distinct probability peaks, allowing us to establish a rank order in each study based on the range of probabilities each membership function represented.

On average, phrases were rated as having higher weighted mean peak values in Study 2 compared to Study 1. The rank order in Study 2 largely replicated our rank order of mean peaks in Study 1. That is, the lowest and highest end of the scale contained the same phrases as Study 1. While all phrases representative of the medium scale spectrum were rated as such across both samples, rank order differed for three phrases in that category. Phrases representing medium confidence (such as *I think that is him/her;*

He/She looks like the criminal; *Somewhat confident*) may not hold one precise meaning across individuals. This is consistent with previous work, suggesting that individuals can translate high confidence consistently, but not low or medium confidence (Mansour, 2020). In our studies, low confidence may have been consistently interpreted because of the limited number of verbal phrases we found that represented the lower end of the scale (*not very confident, not sure*). In practice, individuals may use a larger variety of phrases to represent low confidence, making it harder for others to interpret. Yet, the 13 phrases we selected were the most frequently used ones across 3976 verbal confidence statements from participants who identified someone from a lineup in experiments conducted in our laboratory. Thus, it may alternatively be that people are less likely to make an identification when they have low confidence, and therefore language more broadly used to express low confidence is less strongly associated with eyewitness identifications. Consequently, eyewitnesses may use a more limited selection of phrases. Indeed, participants in two experiments translated their verbal confidence judgements onto a 0-100% scale and only 79% (Experiment 1; 69% in Experiment 2) of them translated their confidence to a value of below 50% (cf. 94%; 95% above 90% or highly confident, and 75%; 63% in between or medium confident; Table 4, Mansour, 2020).

The set of eight narrowest ranges in Study 1 was broadly replicated in Study 2. In both studies, the set of eight narrowest ranges contained the seven phrases that were the same (*Not very confident, I think it is him/her*, *Moderately confident, somewhat confident, fairly confident, Confident, Very confident*). Unlike Study 1, the selection in Study 2 included *pretty sure* as having a narrow span. *Quite confident,* one of the phrases included in the selection of eight of narrow spans in Study 1 and our final lexicon, did not have a narrow span in Study 2 and was thus not included in the set of eight. Study 2 also replicated the set of eight lowest standard deviations in Study 1. In both studies, the set of eight lowest standard deviations contained the same seven phrases (*Not very confident, I think it is him/her, Moderately confident, Somewhat confident, Fairly confident, Confident, Very confident*). Unlike Study 1, Study 2 included *pretty sure* in the selection of eight lowest standard deviations. *Quite confident* was also not included in the set of eight lowest standard deviations for Study 2.

In sum, we did not replicate our findings for *quite confident* in Study 2. While *quite confident* was included in the abbreviated lexicon for Study 1, we were unable to replicate this finding in Study 2. In Study 2, *quite confident* was not included in our set of eight for lowest standard deviations, nor our set of eight for narrow spans. Instead, we found in Study 2 that *pretty sure* covered a similar area on the scale. We speculate about possible reasons for this difference in the General Discussion.

We had hypothesized that the membership functions for some phrases would almost fully overlap, indicating interchangeability (synonyms). Specifically, we expected *somewhat confident* and *moderately confident*, and *quite confident* and *pretty sure* to overlap. This hypothesis was partially supported in both samples. In both studies, *somewhat confident* and *moderately confident* were rated to be synonymous. However, in both studies *moderately confident* was rated to be easier to interpret than *somewhat confident* (i.e. smaller standard deviation and narrower range).

Quite confident and *pretty sure* were not rated as synonymous by either sample. However, in each abbreviated lexicon, *quite confident* and *pretty sure* seemed to represent a distinct range of values (66-78%). Even though these phrases were not rated as synonymous within each study, between studies these phrases represented a similar range of values (see Appendix 4, Figure 25). As you will see below, we thus included *moderately confident*/*somewhat confident,* and *quite confident*/*pretty sure* as synonyms in our final lexicon. Similarly, *confident* and *very confident* were rated as synonymous in both studies with *very confident* having the higher peak. We thus included *confident* and *very confident* as synonyms in our final lexicon. While we replicated synonymity between both studies for four phrase pairs (*moderately confident/somewhat confident; confident/very confident; he/she resembles the criminal/ he/she looks familiar; not very confident/not sure*), interpretations for all other phrases varied between studies. While people seem to interpret some phrases consistently, others do not seem to have clear meaning across people. While eyewitnesses frequently use phrases such as *fairly confident* or *I think that is him/her* to express confidence, others may not be able to interpret these phrases reliably (e.g., interpretation of phrases in Behrman & Richards on a 10-point confidence scale). Given this inconsistency in interpretations, it further negates the use of arbitrarily selected ranges in combined formats. Behrman and Richards asked a group of participants to estimate the ranges for low, medium and high confidence levels on a 10-point scale. “The mean ranges were 0-4, 5-7 and 8-10 for the three confidence levels” (Behrman & Richards, 2005, p. 284). However, our data suggest that individuals’ interpretation of commonly used confidence phrases is manifold. That is, on a scale of 0-100, participants interpret verbal confidence statements using (at least) four categories.

As hypothesized, phrases indicative of extreme levels of confidence on the upper end of the scale (e.g., *very confident*) had more narrow membership functions overall than those indicative of middle-range confidence (e.g., *pretty sure*). This hypothesis was only partially supported for extreme levels of confidence on the lower end of the scale (e.g., *not very confident*). Only in Study 2 did *not very confident* have a narrower membership function compared to phrases representing middle-range confidence. Our findings indicate that extreme levels of confidence on the upper end of the scale have a more commonly understood and concise meaning across individuals. This finding converges with previous work. People prefer to communicate uncertainty verbally in part because verbal statements are "fuzzy" and vague (Mandel & Irwin, 2021) and tend not to use extreme phrases (such as 0% confident/very unlikely or 100% confident/Very likely) when talking about their confidence in a decision (Wintle, Fraser, Wills, Nicholson & Fidler, 2019). It may be less appealing for people to communicate uncertainty using extreme indication based on the conciseness of interpretation for extreme ends of the scale. These findings further call into question why eyewitness identification evidence is treated as being one or the other when its informational value suggests confidence is much more nuanced.

Despite the criminal justice system's preference for eyewitnesses to give confidence in their own words, verbal probabilities are judged to be not as clear as numeric estimates when communicating degrees of probability (Collins & Mandel, 2019). Overall, phrases in the "medium" confidence area had much broader density functions, leading to more overlap between membership functions compared to phrases in the "lower" and "upper" end of the scale. These findings suggest that individuals interpret some probability statements more easily than others. For example, *not very confident* and *very confident* were interpreted as spanning a distinct and specific range across both studies. However, it seems that individuals have difficulty interpreting "medium" confidence. It may be that "medium" confidence statements do not hold one precise meaning across individuals. But note also that a majority of our phrases were judged as indicating medium confidence (8 out of 13 phrases). This in and of itself shows that eyewitnesses are not (or not perceived as being) one or the other–0% or 100% confident. Considering that a singular verbal confidence statement is likely to be interpreted by multiple triers of fact (police officer, jurors, judge), the risk for misperception of eyewitness confidence may be even more pronounced in practice.

We hypothesized that the final lexicon would include at least some of the statements reported in Behrman and Richards' (2005) Table 1 as commonly used by eyewitnesses. This hypothesis was supported. In total, 9 out of the 13 phrases we examined were included in Behrman and Richards' Table 1. Of these, four—*Moderately confident*, *pretty sure, confident* and *very confident*—appeared in our final lexicon.

We expected the terms included in the final lexicon to be similar to those used in climate science lexicons (i.e. four phrases spanning the scale; very unlikely, unlikely, likely, very likely in IPCC reports, Ho et al., 2015). However, unlike the climate science lexicons, our lexicon included the phrases *quite confident* and *pretty sure* to represent the upper medium area of the scale (cf. likely or confident). This suggests that a confidence lexicon for eyewitnesses differs from those used in climate science in its practical application. For example, the phrase *quite confident* may be more natural to be used by the general public (including eyewitnesses) but is not as frequently used by experts when communicating uncertainty. It may also be the case that individuals interpreting uncertainty in climate science attach different meanings to the phrases *quite* *confident* and *pretty sure* given the associated consequences or outcome severity (e.g., an earthquake versus changing tides). This suggests that probability lexicons vary to some extent based on discipline and user characteristics, which is consistent with research demonstrating the importance of context for the use of language expressing certainty (Brun & Teigen, 1981; Cash & Lane, 2017; Dodson & Dobolyi, 2015). Future research should explore the underlying differences when communicating uncertainty in varied settings.

Finally, we hypothesized that the validation sample data (Study 2) would result in a similar lexicon (i.e. similar or the same number of terms and including nearly all the same or the same terms; similar or the same synonyms in Study 1 as in Study 2). This hypothesis was partially supported. Both abbreviated lexicons contained four phrases that spanned 0-100% with distinct peaks across the span. However, our abbreviated lexicon in Study 2 differed from our abbreviated lexicon in Study 1 in two ways. First, our Study 1 abbreviated lexicon contained 2 phrases (*quite confident; somewhat confident*) that were not included in our Study 2 abbreviated lexicon. Second, adjacency varied for six phrases, five of which were generally rated as reflecting medium confidence. There are possible explanations for the differences in our findings. In Study 1, participants were recruited via CloudResearch Amazon Mechanical Turk (English-speaking countries only). In Study 2, our participants were recruited via word of mouth and social media. It may be that interpretations of phrases vary based on user characteristics (e.g., native language, culture). Future research should test the generalizability of the lexicon across diverse populations and explore possible underlying mechanisms for determining synonymity between phrases.

General Discussion

It was possible to select a subset of phrases that had distinct membership functions spanning a full 0-100% scale. Based on previous work (Dhami, 2018; Ho et al, 2015; Renooij & Witteman, 1999), we expected to be able to abbreviate the selected phrases (between four to eight) given their distinct membership functions and interchangeability. We were able to abbreviate the lexicon to four phrases (with three synonyms). Our research suggests that people have stable interpretations of verbal confidence statements. While individuals interpret probability estimates similarly on a 0-100% scale, the boundary conditions for phrases to be interpreted as synonymous appears more complex. Nevertheless, this finding is promising. Individuals seem to interpret and quantify verbal confidence statements on a scale from 0-100% using four phrases.

A standardized lexicon approach in applied settings might facilitate a better understanding of the degree of certainty an eyewitness intends to express and create common ground to limit the variability in interpretation within and between those who receive and interpret the eyewitness' expression. Overall, this systematic approach to interpreting eyewitness confidence may address some of the other shortcomings of verbal, numeric, and other scale judgements previously tested in research.

We intended the format of the lexicon to be as flexible and accessible as possible to accommodate preferences for giving and receiving confidence. The lexicon could be given to eyewitnesses to make their decision. Eyewitnesses could report one of the phrases, a number, a range, or simply draw a line or circle on the lexicon to represent their confidence. Their response, together with the lexicon, could then be given to the trier of fact or police officer so that they can make a more informed decision about the eyewitness' confidence level than if they were simply provided a single statement or number. That is, the lexicon by virtue of including verbal, numeric, and graphical information may better contextualize the eyewitness' judgement allowing for more consistency in how they eyewitness' certainty is interpreted. Furthermore, the lexicon could be modified for children and other vulnerable populations by using child-friendly phrases or by removing the words and numbers altogether. Additionally, these options would eliminate some of the downfalls of combined formats (e.g., need for numeracy skill and cognitive abilities, also cf. Mandel & Irwin, 2021) without compromising the precision numbers provide. One of the shortcomings of combined formats (e.g., numerically-bound linguistic schemes) is the forced categorization of language. A lexicon foregoes the issue by allowing individuals to respond however they may choose to respond. Previous research indicates that people do not reliably interpret numeric ranges with their associated verbal terms (Budescu et al., 2009; 2012; 2014). Attaching arbitrary numbers to phrases oversimplifies the complexity of language, its natural use in context and subsequent interpretation. Importantly, a subjective categorization of verbal confidence statements is unlikely to be acceptable in practice (e.g., "putting words in the eyewitness' mouth" as cited in Mansour, 2020). Our lexicon differs from numeric-bound linguistic schemes as it provides empirically-established boundaries for verbal confidence phrases in a multi-mode format (visual, verbal, numeric). Giving eyewitnesses the freedom to choose may alleviate some of the hesitancy that combined formats bring about.

The application of a lexicon in practice depends on practitioners' and eyewitnesses' willingness to use it. Law enforcement agencies involved in the investigation of suspects may not believe a lexicon outperforms current procedures (i.e. administration by a police officer). There are further caveats to the obtainment of confidence via lexicon in practice: What if an eyewitness identifies someone with anything other than "100% confidence"? Documenting eyewitness confidence in a standardized way can limit the variability in interpretations and variability in presentation of eyewitness confidence statements later on. As of now, there is no universal standard of how eyewitness' confidence should be obtained (if at all; Fitzgerald, Rubinova & Juncu, 2021). While best practice recommendations (Wells et al., 2020; Wixted & Wells, 2017) emphasize confidence to be obtained immediately following a lineup decision, even that initial statement undergoes interpretation (e.g., by a police officer). We do not know how accurately police officers (and other triers of fact) interpret eyewitness' verbal confidence statements. However, we do know what happens when verbal statements are not accurately interpreted (e.g., Ronald Cotton). *I think that is him* (Jennifer Caninno-Thompson, as cited in Weir, 2016) should be interpreted as 61-65% confident, not 100%.

Current procedures in the United Kingdom treat eyewitness confidence as 100% (identification) or 0% (no identification). This binary treatment of eyewitness confidence fails to distinguish between the true outcomes of a lineup decision. On a lineup test, a positive outcome refers to either a filler identification (false alarm, i.e. false positive) or a suspect identification (hit, i.e. true positive). However, a negative outcome can refer to a correct rejection (i.e. true negative), a false rejection (i.e. miss, false negative) or "I do not know”. A compelling reason for the use of the lexicon (or a standardized record of confidence) is the provision of further information about the true outcomes on a lineup test. For example, recording a confidence statement as “it is definitely one of the foils, 100% sure” (false alarm) rather than “0% confidence” (no ID) provides a more accurate representation of the evidence the eyewitness provided. From a practitioners’ perspective, recording confidence in the true outcomes on a lineup test has the potential to benefit investigations (e.g., to pause investigation on a suspect if the eyewitness says “he is not there” with high confidence).

While we were able to empirically develop a lexicon for eyewitness confidence, there are limitations to this research. We followed Ho et al.’s methodological approach but our sample sizes were relatively small—as were theirs. There is currently no standard as to what constitutes a sufficient sample size for this type of work because no effect sizes are calculated or significance tests conducted. Additionally, the majority of both our samples were Caucasian and English-speaking. However, even though our samples differed on some characteristics (e.g., location), the lexicons and interpretations of phrases were very similar.

While we were able to span the entirety of the scale with four phrases, there was not one phrase best representing the 40% mark. It may be that our full phrase set (all 13 phrases) did not include a phrase that individuals reliably interpret as “40% confident”. It may be that there is no phrase used by eyewitnesses to convey “40% confidence” (or that eyewitnesses are never “40% confident”). To test this possibility, future researchers could ask participants which phrases they would use to represent “40%” specifically.

Previous work indicates that context influences interpretation of probability phrases (Cash & Lane, 2017; Dodson & Dobolyi, 2015). In this experiment, we presented our phrases with limited context (“statements made by eyewitnesses”). Presenting phrases without context may not be feasible in practical settings. That is, interpreters may receive a detailed confidence statement (sometimes containing more than one statement, such as “I can’t be sure but I think that is him, his nose looks familiar and he kind of looks like my cousin”), or may hear specific circumstances of the crime (e.g., eyewitness was the victim, crime was personal, suspect is familiar). In addition, individuals assign different meanings to phrases when the consequences of events vary (Mosteller & Yout, 1982). One individual may consider *pretty sure* to be sufficient evidence for a conviction of a petty theft, but not when the outcome is life in prison. Future research should test the extent to which interpretations of phrases are stable when phrases are presented in different contexts (for example, using vignettes), including varying consequences that more closely match the real-world context for which the lexicon is intended.

Lastly, though we compared confidence phrases for semantic meaning by using a systematic approach (shared area under the curve, peaks in the same cut off area), there was variation in synonymity between samples. Our approach to analyse synonymity between phrases needs replication. For example, a more stringent measure of synonymity (e.g., direct comparison between phrases in isolation) would provide greater information about the usefulness of our approach. We chose to employ a quantitative approach to produce objective data. However, it may be useful to employ other methods to produce a lexicon and determine synonymity. Qualitative methodologies, such as the nominal group technique, may provide alternative approaches to replicate our quantitative lexicon approach.

In summary, our research highlights the potential of empirical interpretations for probability judgements in the eyewitness area. Our innovative, empirically based approach provides common ground for eyewitnesses and triers of fact when asked to provide and interpret verbal statements of confidence. Across both samples, four phrases (and three synonyms) reliably spanned a 0-100% scale. This lexicon provides initial evidence for boundaries of phrases interpreted as low, medium, and high confidence. Validation of the lexicon is needed to judge its practical performance, but this tool has the potential to reduce the extent to which a game of telephone ensues when an eyewitness expresses confidence in a decision and that confidence is used to make decisions at different levels of the criminal justice system.

Abstract

Triers of fact must interpret verbal eyewitness confidence statements, but we do not know what phrases are interpreted as high, medium and low confidence. We (i.e. myself and supervisor) aimed to validate the interpretation tool developed in the previous study (i.e. the lexicon) by testing the replicability of rank order for the 13 commonly used verbal confidence phrases used to develop the lexicon. Participants rank ordered phrases from lowest to highest level of confidence expressed. The interpretations of the phrases were stable when the phrases were ranked as indicating low (not very confident; not sure) and high confidence (very confident; confident). Our results suggest that individuals have stable rank orders for some medium confidence phrases (such as quite confident; fairly confident; moderately confident), but not for others (e.g., he/she looks like the criminal). This research replicates previous work and provides an empirically-sourced rank order for eyewitness verbal confidence statements.

Chapter 3

Rank Order of Verbal Confidence Statements

Given that verbal confidence statements are commonly used in practice and generally preferred, we need to improve the ability of the criminal justice system to ensure that others interpret eyewitness confidence in the way the eyewitness intended. While we (i.e. myself and supervisor) believe a lexicon can be a useful tool to minimize miscommunication of eyewitness confidence, there are other ways to improve the communication and interpretation of confidence statements. A first step to limiting the variability in interpretations of verbal expressions of uncertainty could be the determination of a defined rank-order of verbal confidence phrases (Renooij & Witteman, 1999).

People are less variable in assigning numeric estimates to expressions in an ordered list than to expressions in a random list (Hamm, 1991) and show consistency in their rank-ordering for verbal expressions of uncertainty (Budescu & Wallsten, 1985; Newman, 1967). Rank ordering of verbal probability phrases also seems to be consistent over time (Kong, Barnett, Mosteller, & Youtz, 1986). Given that individuals have relatively stable lexicons for probability phrases, rank orders (or in other words, an ordinal scale) can provide a way to reliably translate meanings of verbal probabilities between the sender and receiver (Budescu et al., 1988; Mandel & Irwin, 2021).

We do not know the extent to which these findings generalize to eyewitness' statements of confidence. Our prior work suggests individuals can quantify verbal confidence statements on a scale of 0-100% using four phrases (see Chapter 2).

Establishing a rank order could provide a practical and straightforward-to-implement alternative to the use of a lexicon in applied settings. Rank orders circumvent assigning numeric (i.e. predetermined) meanings to verbal statements without compromising preference to communicate confidence verbally. Use of rank orders may alleviate some of the hesitancy when people are asked to use numbers or combined scales to express their confidence with the potential to minimize variability in interpretation.

Numeric rank orders are easy to elicit, universally understood and come with a pre-determined order: 80% is always higher than 25%, for example. In research, rank orders are commonly determined by calculating means for each linguistic probability term to then establish rank orders numerically (e.g., Chapter 2; Mansour, 2020; Kenchel, et al., 2021; MacLeod & Pietravalle, 2017). Numbers can offer precision but translating verbal terms into numbers is unnatural for individuals and may not translate to practical settings, especially for eyewitnesses (given that the criminal justice system currently relies on verbal communication of eyewitness evidence). Renooji and Witteman (1999) proposed the assignment of rank numbers instead of means (i.e. assigning numbers, not distances) to establish an ordinal scale for expressions of uncertainty. This approach is considerably more nuanced than assigning numeric estimates as it allows for intuitive arrangement of verbal probabilities. In the present study, we adopted this approach for the eyewitness area and in relation to the 13 phrases we explored in the previous study.

Our lexicon provides initial evidence that individuals can quantify eyewitness confidence statements using four phrases on a scale of 0-100%: One phrase represents “low” confidence (*Not very confident*), two phrases represent “medium” confidence (*Moderately confident* at the lower end, *Quite confident* at the upper end) and one phrase represents “high” confidence (*Very confident*) (see Chapter 2). We were able to establish a rank order using mean peak values of membership functions. However, these findings need replication. It would be useful to see if the rankings replicate when solicited in a different way.

There is no work available that has specifically tried to establish a rank order for eyewitness’ verbal confidence statements by asking individuals to rank phrases in comparison to one another. While it may seem intuitive to consider certain phrases (e.g., *very confident, moderately confident, not very confident*) to represent high, medium and low confidence, there is great variability when individuals are asked to interpret verbal confidence statements in practice (Mansour, 2020), especially when phrases represent medium confidence (e.g., Chapter 2). Such variability in understandings for verbal statements of uncertainty can negatively affect decision-making (Dhami & Wallsten, 2005; Ligertwood & Edmond, 2012; McQuiston-Surrett & Saks, 2007). Establishing an empirically sourced rank order thus offers a first step to minimize variability in interpretations of eyewitness confidence statements. Rather than relying on subjective intuitions, an empirically-established rank order can provide information about understandings of “high”, “medium” and “low” confidence phrases when asked to give and interpret eyewitness confidence.

Behrman & Richards (2005) provided initial evidence of rank ordering of eyewitness confidence judgements. They asked participants (*N* = 84) to rate 35 verbal confidence phrases used by real eyewitnesses on a 10-point confidence scale. A second group of individuals (*N* = 40) estimated the ranges for low, medium and high confidence levels on a 0-10 confidence scale. “The mean ranges were 0-4, 5-7 and 8-10 for the three confidence levels. The confidence phrases were assigned to one of the three confidence levels on the basis of their confidence ratings" (Behrman & Richards, 2005, p. 284). Behrman and Richards categorized phrases based on their ratings but we do not know whether their presentation organized the phrases in any particular order. Mansour (2020) provided further evidence for a rank order for own words confidence judgements. Participants (*N* = 36) rated statements that were not a part of Behrman and Richards' (2005) set on a scale of 0 (no confidence) to 10 (absolute certainty)—the same scale Behrman and Richards used. Mansour then calculated mean ratings for each phrase and reported them in Supplementary Table 2 in an ordered fashion (lowest to highest based on mean ratings).

The aim of the current study was to establish a stable rank order for the 13 most frequently used verbal confidence statements by participant-eyewitnesses (and specifically the four phrases used in our previously developed lexicon, Chapter 2).

If understandings of phrases are stable, the rank order of phrases (compared to rank orders in the previous study) should replicate. Specifically, we expected high confidence (*very confident and confident*) and low confidence (*not very confident, not sure*) to be rated as such (*very confident* as 1st and *confident* as 2nd, i.e. as highest in the rank order; *not very confident* as 13th, and *not sure* as 12th, i.e. as lowest in the rank order). Based on our prior work, we expected *somewhat confident* to be rated as 7th, and *moderately confident* to be rated as 8th (i.e. to be ranked in the middle of the phrases). Our hypotheses were pre-registered on the Open Science Framework: osf.io/dbncz. The study was approved by the university's research ethics board.

Method

Participants

Participants were adults with sufficient visual capacity to view a computer screen. Participants (*N* = 85) were recruited via QMU's SONA system for course credit. The usable sample (*n* = 49) did not include duplicate IP addresses, cases where participants failed the attention check, indicated to have technical difficulties, did not provide a ranking for each phrase, took too long (> one hour), did not understand the task or completed the task randomly (*n* = 36) (i.e. participants that did not rank *very confident* in 1st to 11th position, and/or did not rank *not very confident* in 5th to 13th position). The usable sample of participants identified as primarily female (83.67%), male (12.24%), and other (4.08%). Participants identified as white (73.47%), Scottish (8.16%), British (2.04%), Arab (2.04 %), African (2.04%), Indian (2.04%) or preferred not to answer (10.20%) with a mean age of 21.19 years (SD = 7.14, Range = 17-51).

Design

This study used a within-subjects design with a single factor: confidence phrases, of which there were 13.

Materials

The study was programmed on Qualtrics.

Phrases

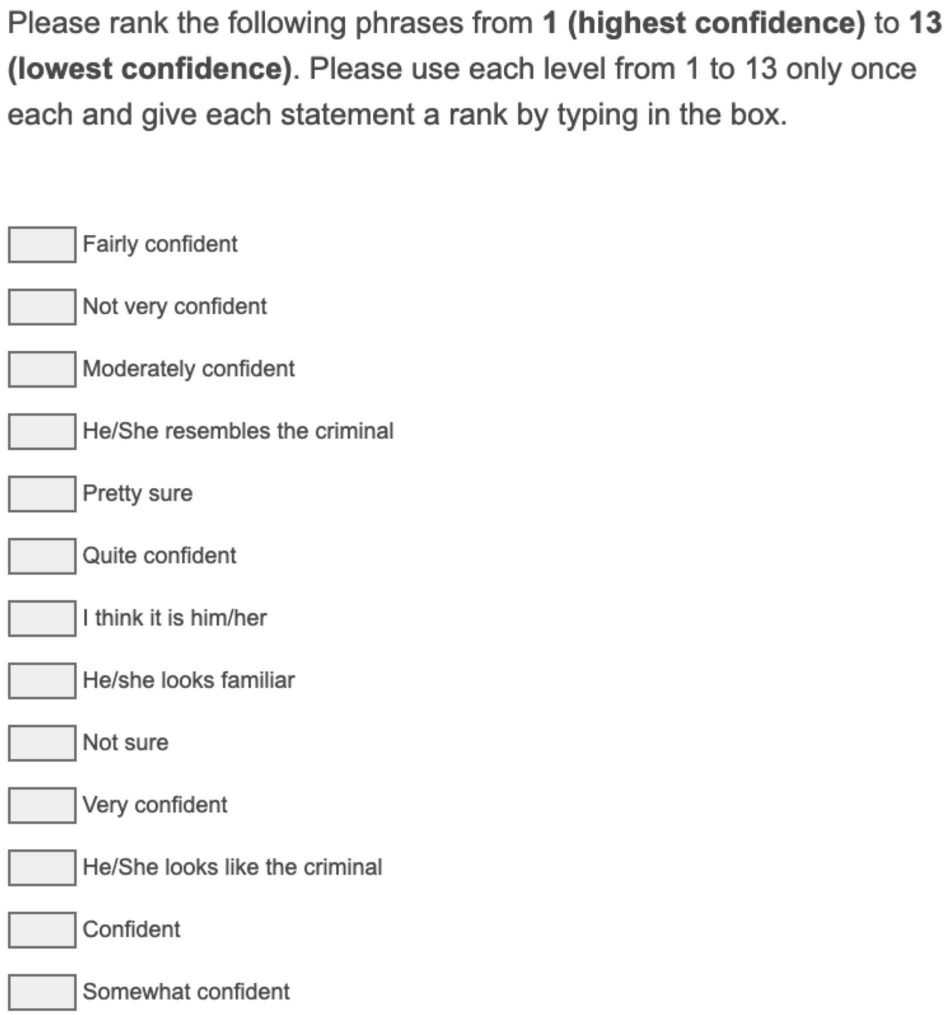
The phrases participants were presented with were most frequently provided by participants in our prior research (Chapter 2; Mansour, 2020; see Table 3, p.48). Phrases were presented without context (e.g., pretty sure, "statements made by real eyewitnesses"). That is, phrases were presented without provision of the full, original statement obtained from eyewitness-participants (e.g., “I think I got the right guy, I’m fairly confident”).

Ranking task

Participants were presented with all 13 phrases and asked to rank order them from 13 to one by assigning rank numbers, with one denoting the highest level of confidence. Each participant assigned a rank to each of the verbal phrases presented. Participants could only assign each rank number once. All phrases were presented at once. The order of presentation of phrases was randomized (see Figure 18).

Figure 18

Ranking Task



Please rank the following phrases from **1 (highest confidence)** to **13 (lowest confidence)**. Please use each level from 1 to 13 only once each and give each statement a rank by typing in the box.

□ Fairly confident

□ Not very confident

□ Moderately confident

□ He/She resembles the criminal

□ Pretty sure

□ Quite confident

□ I think it is him/her

□ He/she looks familiar

□ Not sure

□ Very confident

□ He/She looks like the criminal

□ Confident

□ Somewhat confident

Attention checks

The study included one attention check (e.g., “We asked you to evaluate judgements reported by who?” Multiple choice answer: Eyewitnesses). Participants that failed this attention check were excluded from analyses. Participants were also asked questions pertaining to task comprehension (“Did you understand how to do this task?”), cheating (“Did you cheat in any way? That is, when doing the study did you do anything to make it easier to answer the questions. This will not affect your credit/reimbursement”) and technical difficulties (“Did you have any technical difficulties?”). Data from trials where a participant responded that they experienced technical difficulties, indicated they cheated, or indicated they did not understand the task were excluded from analyses, as per our pre-registration.

Procedure

After viewing the information sheet and providing informed consent (see Appendix 1), participants were asked to complete the study on a desktop. After providing demographic information, participants were instructed that they would be seeing 13 statements that eyewitnesses provided when asked for confidence in their own words. Participants were next presented with all 13 phrases and asked to assign rank numbers to each of the 13 phrases. After participants assigned a rank to each of the phrases, they were presented with the attention check and questions pertaining to data quality. Finally, the participants were debriefed, thanked for their participation, and granted credit.

Results

Table 111 displays the descriptive statistics for each phrase. High confidence phrases (*very confident* and *confident*) and low confidence phrases (*not very confident*, *not sure*) were ranked as expected (*very confident* as 1st and *confident* as 2nd, i.e. as highest in the rank order; *not very confident* as 13th, and *not sure* as 12th, i.e. as lowest in the rank order) across each measure of central tendency (with median ranking indicating *not sure* to also be ranked 13th) (see Table 11).

Moderately confident, one of the core phrases in our lexicon representing medium confidence (40-70%), was ranked as 7th in the present study compared to 8th in Study 1 and Study 2 (Chapter 2). Median and mode ranking indicate *moderately confident* to be most frequently ranked as 6th.

We did not replicate the rank order for *somewhat confident*. While *somewhat confident* was ranked 7th in Study 1 (Chapter 2), *somewhat confident* was ranked 9th in Study 2 (Chapter 2) and in our present study (see Table 10). Median rankings indicated *somewhat confident* to be ranked 8th. Mode rankings indicated *somewhat confident* to be ranked 7th. Median and mode rankings also indicate that some phrases were ranked similarly (e.g., *quite confident* and *fairly confident*, median ranking 5th for both). Standard deviations for *very confident, confident, not sure* and *not very confident* were lower than all other phrases.

We compared this rank order to the rank orders previously established in our lexicon studies (Study 1 and Study 2 in Chapter 2). As you can see, the rank order established via assignment of rank numbers almost perfectly replicated our rank orders in Study 1 and Study 2 (Chapter 2). Table 11 displays the rankings obtained in the current study with those from Study 1 and Study 2 (Chapter 2). The four top ranked statements (*very confident, confident, and quite confident, fairly confident*) and the two lowest ranked statements (*not very confident, not sure*) were in the same order across all three studies. The remaining phrases varied in their ranking order between the three studies (see Table 12).

Table 11

Descriptives of all 13 Phrases in Study 3

Phrase	*Mdn*	*M*	*Mo*	*SD*	*n*	*Min*	*Max*
Very confident	1	1.51	1	1.36	37	1	9
Confident	2	2.82	2	1.81	25	1	12
Quite confident	5	5.37	5	2.22	9	2	11
Fairly confident	5	5.47	3	2.61	11	1	11
He/she looks like the criminal	6	5.82	3	3.18	10	1	12
I think it is him/her	6	6.27	3	2.93	6	1	11
Moderately confident	6	6.94	6	2.54	11	3	13
He/She resembles the criminal	9	7.71	5	3.03	9	1	13
Somewhat confident	8	7.71	7	2.19	9	3	11
Pretty sure	8	7.73	7	2.18	9	2	11
He/She looks familiar	10	9.24	11	2.18	13	1	13
Not sure	13	12.18	12	1.03	23	7	13
Not very confident	13	12.22	13	1.21	25	7	13

Note. Lower rank numbers indicate a higher level of confidence. The phrases are arranged by mean rank from highest to lowest.

Table 12

Rank Order in Study 3 Compared to Study 1 and Study 2 in Chapter 2

Rank order	Study 1	Study 2
Very confident	**Very confident**	**Very confident**
Confident	Confident	Confident
Quite confident	**Quite confident**	**Quite confident**
Fairly confident	Fairly confident	Fairly confident
He/she looks like the criminal	Pretty sure	Pretty sure
I think it is him/her	He/She looks like the criminal	I think it is him/her
Moderately confident	Somewhat confident	He/She looks like criminal
He/She resembles the criminal	**Moderately confident**	**Moderately confident**
Somewhat confident	I think it is him/her	Somewhat confident
Pretty sure	He/She looks familiar	He/She looks familiar
He/She looks familiar	He/She resembles the criminal	He/she resembles the criminal
Not sure	Not sure	Not sure
Not very confident	**Not very confident**	**Not very confident**

Note. Rank order in the present study compared to Study 1 and Study 2 in Chapter 2. Lexicon phrases (see Chapter 2) are shown in bold.

Discussion

We aimed to establish a rank order for the 13 most frequently used verbal confidence statements by eyewitnesses (and specifically the four phrases used in our previously developed lexicon, Chapter 2). We compared this rank order to the rank orders previously established in our lexicon studies (Study 1 and Study 2 in Chapter 2).

We expected to replicate the rank orders found in our studies 1 and 2. Specifically, we expected high confidence (*very confident* and *confident*) and low confidence (*not very confident*, *not sure*) to be ranked as such (*very confident* as 1st and *confident* as 2nd, i.e. as highest in the rank order; *not very confident* as 13th, and *not sure* as 12th, i.e. as lowest in the rank order). This hypothesis was supported. For those phrases, the rank order established via assignment of rank numbers almost perfectly replicated our rank orders in Study 1 and Study 2 (Chapter 2). On the high end of the rank order, *very confident* was ranked as 1st (i.e. highest level of confidence) in all three studies. *Confident* was ranked as 2nd highest level of confidence in all three studies. On the low end of the rank order, *not very confident* was ranked 13th (i.e. lowest level of confidence) in all three studies. *Not sure* was ranked 12th (i.e. second lowest) in the rank order. Standard deviations for *very confident, confident, not sure* and *not very confident* were lower than all other phrases, indicating less variability in interpretation. These findings suggest that individuals can reliably rank and interpret high-confidence and low-confidence phrases.

Based on our prior work, we expected *somewhat confident* to be ranked as 7th, and *moderately confident* to be ranked as 8th (i.e. to be ranked midpoint of the rank order). We found partial support for this hypothesis. *Moderately confident,* one of the core phrases in our lexicon representing medium confidence (40-70%), was ranked as 7th in the present study compared to 8th in Study 1 and Study 2 (Chapter 2). We did not replicate the rank order for *somewhat confident*. While *somewhat confident* was ranked 7th in Study 1 (Chapter 2), *somewhat confident* was ranked 9th in Study 2 (Chapter 2) and in our present study. Thus, although the ranking was not perfectly stable, it was still quite consistent. Certainly, if we consider whether the phrases were considered low, medium, or high confidence, that categorization would be supported across all three studies. Indeed, mode rankings indicate *somewhat confident* to be most frequently ranked 7th, median rankings indicate *somewhat confident* to be ranked 8th.

We also replicated the rank order for *quite confident* and *fairly confident* across all three studies. *Quite confident*, a core phrase in our lexicon representing upper medium confidence at 70-80%, was ranked 3rd in Study 1, Study 2, and in the present study. *Fairly confident* was ranked 4th in all three studies. This finding is encouraging: People seem to have stable rank orders for some medium confidence phrases (such as *quite confident, moderately confident* and *fairly confident*). Our results suggest that individuals reliably rank and interpret *moderately confident* to represent a midpoint of the rank order and *quite confident* to represent the upper end of medium confidence. The replication of reliable interpretation for *moderately confident* and *quite confident* provide support for the use of these two phrases to represent "medium confidence" in our lexicon.

However, no other phrases (all of which represent medium confidence) were ranked consistently across all three studies. There are possible explanations for the differences in rankings of these phrases. While some phrases (*moderately confident, fairly confident*, *quite confident*) were reliably interpreted, it may be that individuals have difficulty rank ordering and interpreting "medium confidence" phrases more broadly. Phrases like *he/she looks like the criminal* and *pretty sure* may not hold one precise meaning across individuals. This result is in line with previous findings, indicating that individuals do not reliably interpret medium confidence (Mansour, 2020; Kenchel, et al., 2021; Chapter 2). However, it may also be that individuals use certain phrases interchangeably. For example, *he/she looks like the criminal* and *I think it is him/her* may be considered to hold similar meanings (i.e. may be synonyms). In our present study, participants were only able to assign each rank number once. It may be that people use (and rank) some of the phrases interchangeably when given the choice. Either way, our findings emphasize the interpretive difficulties when people are presented with “medium confidence phrases” (e.g., *he/she looks like the criminal*). Given that verbal confidence statements undergo multiple levels of interpretation in practice (e.g., police officer, judge, juror, general public), these difficulties are likely even more complex in practice. For example, a police officer may interpret *he/she looks like the criminal* to indicate “high confidence” when the eyewitness intended to communicate “medium confidence”. A judge or juror may interpret *he/she looks like the criminal* as “low confidence”. It is important to provide people with guidelines to minimize this variability in interpretation.

Previous work asked participants to assign scale ratings to confidence phrases to establish rank orders (via mean ratings; Mansour, 2020; Kenchel et al., 2021). As Rennoij & Witteman (1999) note, mean ranks are a more accurate representation of the data than means: Participants assigned rank numbers, “not distances between expressions” (p. 180). We also calculated the median and mode ranks for all phrases. The rank orders of some phrases (e.g., *He/she resembles the criminal; somewhat confident*) differed in median and mode ranks compared to mean ranks (both phrases had mean ranks of 7.71 but differed in medians and modes). Median and mode ranks supported our hypotheses: *Very confident* and *confident* were most frequently ranked as 1st and 2nd by most participants (*n* = 37; *n* = 25). Similarly, low confidence (*not very confident; not sure*) were ranked as 13th and 12th most frequently by a majority of participants (*n* = 25; *n* = 23). The ranking differed for medium confidence phrases for mode and median ranks: *Moderately confident* was ranked 6th (mode and median). Median rankings indicated *somewhat confident* to be ranked 8th and mode rankings indicated *somewhat confident* to be ranked 7th. Median and mode ranks may provide further information when individuals are asked to rank order verbal probability phrases. For example, it may be more useful to consider most frequently assigned rankings rather than the average ranking for a certain phrase. Median and mode ranks suggest assignment of the same rank number to some phrases (e.g., median rankings for *fairly confident* and *quite confident*, both ranked 5th), potentially indicating synonymity between phrases.

While we were able to establish and replicate a rank order for frequently used verbal confidence phrases, there are limitations to our study. In our study, we presented the phrases with limited context (i.e. "these were statements made by eyewitnesses"). However, in the real world, eyewitnesses are unlikely to encounter confidence phrases without context. For example, eyewitnesses may rank order phrases differently when outcome severity varies. If the outcome is severe (e.g., life in prison), *somewhat confident* may be interpreted differently compared to an outcome that is not as severe (e.g., community service). While previous work suggests that rank order is not affected by context (Renooij & Witteman, 1999), future research should replicate the rank order when phrases are presented in context (e.g., with vignettes). Additionally, future research should test the stability of rank order when user characteristics vary. For example, does the rank order for confidence phrases replicate when a police officer or a judge is asked to assign rank numbers?

Lastly, in our study we presented all phrases at once. Participants' judgement and interpretation processes may have been influenced by the simultaneous presentation of all phrases. It may be that individuals rank order phrases differently when asked to make judgements sequentially. However, the procedure used in Study 1 and Study 2 (Chapter

2) was sequential and although the task was not to rank phrases, those results suggest that simultaneous versus sequential presentation is unlikely to impact rankings. We are not aware of any other studies that have examined the extent to which the method of presentation for eyewitness confidence statements affects subsequent rank orders. More research is needed.

This research highlights the potential of empirical interpretations for probability judgements in the eyewitness area. We currently do not have guidelines on how to interpret eyewitness confidence statements. If triers of fact are unwilling to utilize a tool (such as a lexicon), we may still be able to reduce misinterpretation by providing a rank order of common confidence phrases. A rank order can provide an initial framework to guide eyewitnesses and triers of fact without compromising the ability to communicate confidence verbally (as done in practice and generally preferred by eyewitnesses). Similarly, an empirically-established rank order can help discern the weight that is given to an eyewitness' testimony. A rank order can provide common ground for eyewitnesses and triers of fact when asked to provide and interpret verbal statements of confidence.

People can reliably interpret phrases indicating high confidence and low confidence. However, people also seem to have stable rank orders for some medium confidence phrases (such as *quite confident, moderately confident* and *fairly confident*). A rank order provides evidence for understandings of phrases interpreted as low, medium, and high confidence.

Abstract

Triers of fact must interpret verbal eyewitness confidence statements, but we do not know the extent to which such verbal phrases are used interchangeably (i.e. are synonyms). We (i.e. myself and supervisor) aimed to validate an interpretation tool (i.e. lexicon; Chapter 2) by determining the sameness of or difference between 13 commonly used verbal confidence statements. Participants made pairwise comparisons between all 13 phrases on a visual scale (Completely the same; Completely different). Similarity was highest (>75%) for one phrase pairing representing high confidence (*Very confident/ Confident*), high confidence (*Very confident/ Confident*), two phrase pairings representing medium confidence (*Pretty sure/ Fairly confident; Quite confident/ Fairly confident*) and one phrase pair representing low confidence (*Not very confident/ Not sure*). People consistently interpret verbal confidence phrases representing low and high confidence, but only some medium confidence phrases. This research provides evidence for synonymity between verbal confidence statements and can inform how triers of fact interpret verbal confidence judgements to reduce the potential for misinterpretation of an eyewitness' level of confidence.

Similarity of Verbal Confidence Statements

A single eyewitness often uses multiple phrases when expressing a particular level of confidence (e.g., 65% confident, *moderately confident* in one instance, *somewhat confident* in another). Expressing a given level of confidence using two or more phrases suggests individuals use some phrases interchangeably. Interchangeability implies intra-individual overlap in meaning–that is, for a particular eyewitness, *moderately* and *somewhat* may represent the same probability construct. However, we do not know if meanings of phrases are consistently shared *across* individuals. For example, do most people consider *moderately confident* and *somewhat confident* to be interchangeable?

Synonyms are defined as different word forms that share the same meaning (Clark & Clark, 1977; Searle, 1969). Early work proposed "synonyms to be (…) words that can substitute for one another in sentences without changing meaning" (Herrman, 1978, p. 491; Ogden & Richards, 1923). Synonymity is difficult to measure due to the complexity and variability of language. Herrman (1978) suggests that the similarity in meaning between two words can be rated on a "Likert-type scale" (p. 495). However, numerically bound linguistic probability schemes (such as Likert-type scales) are not accurately interpreted by individuals (Budescu et al., 2009; 2012; 2014). That is, individuals do not reliably associate terms with assigned categories. Assessing similarity between probability phrases using a categorical Likert-type scale may thus not be appropriate.

As an alternative, Renooij and Witteman (1999) assessed similarity of verbal probabilities by "scoring pairs of expressions on a 10 cm line" (p. 182). (Dis)similarity of phrases was scored in millimetres. The approach of assessing similarity on a visual scale as outlined by Renooij and Witteman is significantly more nuanced as it allows for similarity to be rated on a continuum without forcing the categorization of individual language use. Moreover, visual scales limit variability in interpretation when individuals are asked to translate between modes of probability information (e.g., verbal probability expressions into numeric estimates, cf. Mansour, 2020). We (i.e. myself and supervisor) thus sought to test synonymity (i.e. (dis)similarity) between eyewitness verbal confidence phrases using a visual scale.

Renooij and Witteman (1999) presented as anchors the expressions “exact same” and “completely different”. Some work suggests that even though two phrases may be considered synonymous, they are not “the **exact** same” (“exactness is highly idiosyncratic”, Herrmann, 1978, p. 494). Rather, similarity-in-meaning indicates the associative overlap between words (Cofer, 1957; Rubenstein & Goodenough, 1965, but cf. Herrmann 1978 for different kinds of meaning). While the use of a visual analogue scale (as used in Renooij and Witteman) to assess similarity-in-meaning provides considerable advantages, one can potentially avoid the problems associated with the phrase “exact” by “completely the same” and “completely different” as anchors to display sameness versus its opposite.

There is evidence to suggest that individuals agree on the extent to which phrases are similar (Rubenstein & Goodenough, 1965). Ho et al. (2015) provides initial evidence for synonymity implied in lexicons in practice (NIC and DI). Ho et al. obtained membership functions (for an explanation of membership functions, see p.40 of this document) and visually compared the membership functions of two phrases to determine if “the items in each of these pairs are, for all practical purposes, indistinguishable and thus can be treated as synonyms” (p. 49). Ho et al. notes that synonymity in practice is implied and suggests this inference is appropriate.

We followed Ho et al.’s (2015) approach to determine synonymity in the development of our lexicon. However, rather than just comparing membership functions visually, we sought to determine the shared area under the curves of two membership functions (i.e. phrases). Two adjacent phrases were deemed synonymous if the shared area under the curve was considerably larger than other pairings for a given phrase, and if the mean peaks of both phrases fell in the same cut-off area. Even though we were able to determine synonymity in our lexicon between adjacent phrases by using this approach, there were considerable differences between our two studies (see p. 74 of this document). A more stringent measure of synonymity (e.g., direct comparison between phrases in isolation) would provide greater information about the usefulness of our approach and allow us to replicate our findings of synonymity. In the current study, we aimed to do just that. We asked participants to rate the sameness of or difference between eyewitness’ verbal confidence statements. We tried to address the extent to which the 13 commonly-used expressions of eyewitness confidence (i.e. the 13 own-words confidence phrases used in studies 1 and 2, Chapter 2) are interchangeable. In other words, when asked to make pairwise judgements of (dis)similarity, to what extent do individuals judge expression of confidence as synonyms/antonyms?

Based on our findings in Study 1 and Study 2 (Chapter 2), we expected *not very confident* and *not sure* to be judged similarly (i.e. be synonymous). We expected *he/she resembles the criminal* and *he/she looks familiar* to be judged similarly. We expected *moderately confident* and *somewhat confident* to be judged similarly. We expected *very confident* and *confident* to be judged similarly (see Table 13). We expected *not very confident* and *very confident* (and their respective synonyms, i.e. *not sure* and *confident*) to be judged as extremely different (i.e. be antonyms). We pre-registered our hypotheses on the Open Science Framework: osf.io/dbncz. The study was approved by the university's research ethics board.

Table 13

Hypotheses for Synonymity in Study 4

Phrase 1	Phrase 2
Not very confident*	Not sure
He/she resembles the criminal	He/she looks familiar
Moderately confident*	Somewhat confident
Very confident*	Confident*

Note. *indicates core phrases used in the lexicon.

Method

Participants

Participants were adults with sufficient visual capacity to view a computer screen. Participants (N = 36) were recruited via the university's SONA system for course credit. The usable sample (n = 27) did not include duplicate IP addresses, cases where the participants indicated to have technical difficulties, took longer than an hour to complete the study, did not pass the attention check, indicated they did not understand the task, or indicated that they cheated (n = 9). The usable sample of participants identified as primarily female (85.19%; male: 14.81%). Participants identified as European (85.19%), Asian (11.11%), or preferred not to answer (3.70%) with a mean age of 21.19 years (*SD* = 7.14, *Range* = 17-48).

Design

This study used a within-subjects design with a single factor: confidence phrases, of which there were 13.

Materials

The study was programmed on Qualtrics.

Phrases

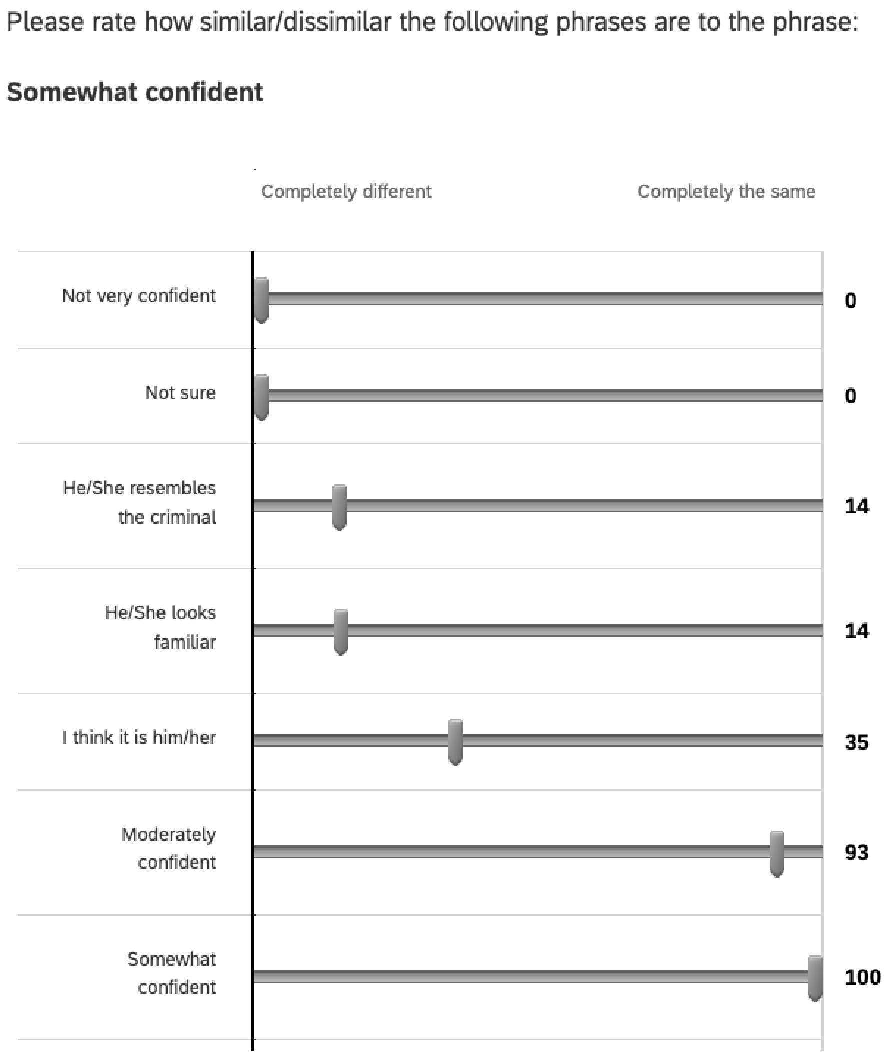
The phrases presented were most frequently provided by participants in our prior research (Mansour, 2020; see Table 3, p.48).

Similarity ratings

In this study, we asked participants for pairwise comparisons to judge similarity amongst all possible pairs of 13 verbal confidence statements. Thus, for each of the 13 expressions, there were 13 pairs to compare. Each participant rated the (dis)similarity of 13 pairs of phrases (169 judgements per participant) on a scale from 0 (Completely different) to 100 (Completely the same) but the numbers were not shown to participants. Rather, they simply saw a line with a slider that they could move between the two anchors (which were presented; see Figure 19). Participants were presented with one phrase at a time and asked to compare that phrase to all 13 phrases. The order in which each phrase was presented was randomized.

Figure 19

Similarity Rating Task



Please rate how similar/dissimilar the following phrases are to the phrase:

Somewhat confident

Not very confident	0
Not sure	0
He/She resembles the criminal	14
He/She looks familiar	14
I think it is him/her	35
Moderately confident	93
Somewhat confident	100

Note. Similarity rating task. Example of ratings given by one participant. All 13 phrases were presented on the left-hand side. Phrase presentation for phrase to rate (e.g., somewhat confident) was randomized.

Attention checks

The study included one attention check (e.g., “We asked you to evaluate judgements reported by who?” Multiple choice answer: Eyewitnesses). Participants that failed this attention check were excluded from analyses.

Procedure

After viewing the information sheet and providing informed consent (Appendix 1), participants provided demographic information. Participants were then instructed that they would be seeing statements that eyewitnesses provided when asked for confidence in their own words. Participants were informed that we are interested in their own interpretation of these statements on a scale from “Completely different” to “Completely the same”. Participants completed one practice trial and were shown a visual example (i.e. a figure of possible practice ratings another participant may have given) before beginning the similarity ratings task. Upon completing the similarity ratings, participants completed the attention check, then were asked questions about task comprehension, prior participation in eyewitness studies, and technical difficulties. Finally, the participants were debriefed, thanked for their participation, and granted credit.

Results

We calculated mean ratings of similarity for each phrase pairing, and these are shown in Table 14. Figure 20 shows direct comparison of similarity ratings to all phrases for the four phrases from our lexicon. Figure 21 presents a visual of the (dis)similarity between all phrases.

We expected to find *not very confident* and *not sure* to be judged similarly (i.e. to be synonymous). This hypothesis was supported. *Not very confident* was rated to be most similar to *not sure* (and vice versa). We also expected *very confident* and *confident* to be judged similarly. We again found support for this hypothesis (see Table 14). *Very confident* was rated to be most similar to *confident* (see Figure 20). Standard deviations were smallest for the comparison of these two phrases. Figure 21 shows that these two phrases are clearly interpreted as synonyms across the entire sample because they received the highest similarity ratings (yellow).

We hypothesized that *he/she resembles the criminal* and *he/she looks familiar* would be judged similarly. We found partial support for this hypothesis. *He/she looks familiar* was rated to be most similar to *he/she resembles the criminal* (see Table 14). However, *he/she resembles the criminal* was rated to be most similar to *he/she looks like the criminal*. Figure 21 indicates that the interpretations of these phrases are fuzzier compared to phrases at the upper (*very confident*) and lower (*not very confident*) end of the scale.

We expected *moderately confident* and *somewhat confident* to be judged similarly. We found partial support for this hypothesis. *Moderately confident* was rated to be most similar to *somewhat confident*, but *somewhat confident* was rated to be most similar to *pretty sure* (see Table 14). Figure 26 (see Appendix 5) indicates that *somewhat confident* was considered very similar to *moderately confident* (second highest match).

We expected *not very confident* and *very confident* (and their respective synonyms, *not sure* and *confident*) to be judged as extremely different (i.e. be antonyms). This hypothesis was partially supported. As you can see by looking at the standard deviations, *not very confident* was clearly judged to be antonymous to *very confident* and vice versa (see Table 14). However, *very confident* was also judged as antonymous to *not sure*, and *not very confident* was also judged as antonymous to *confident*. Surprisingly, *not very confident*, *not sure* and *very confident* were the only phrases judged to indicate clear antonymy compared to all other phrases in the sample (see Figure 21).

Table 14

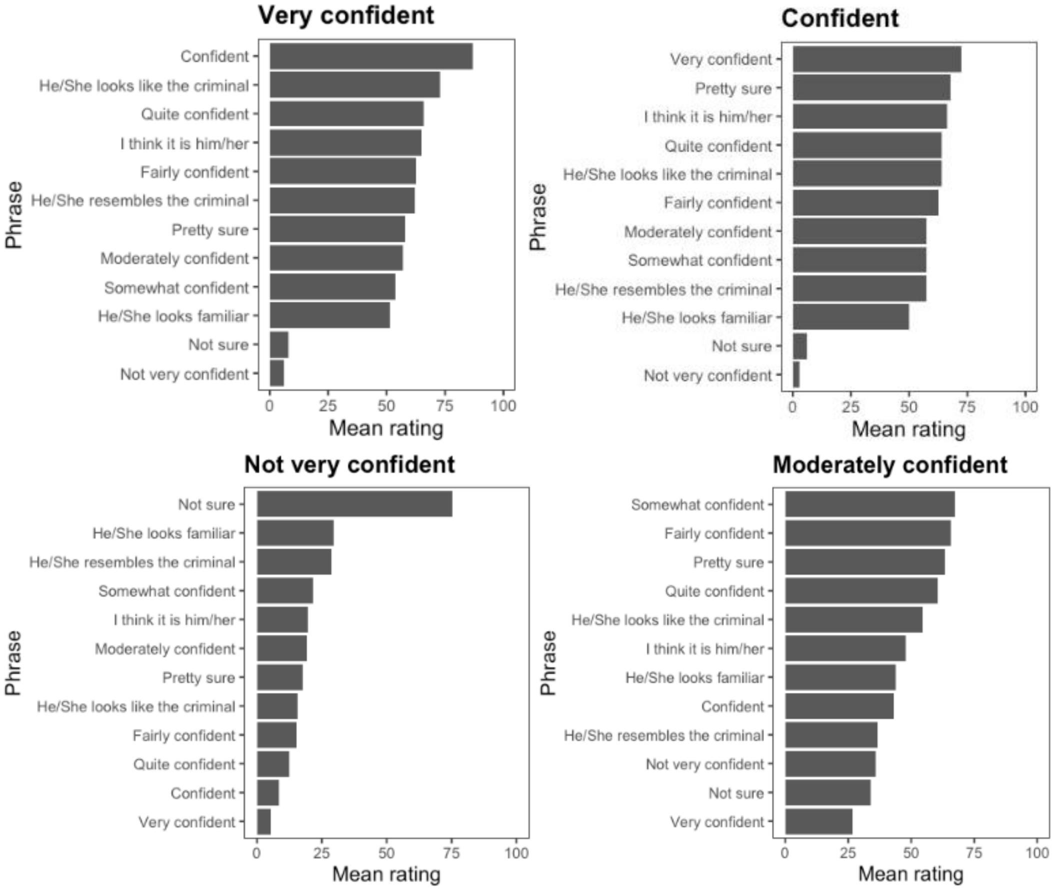
Similarity Ratings for Synomyns and Antonyms

Phrase	Synonym	*M (SD)*	Antonym	*M (SD)*
Not very confident	Not sure	75.33 (24.31)	Very confident	5.37 (15.22)
Not sure	Not very confident	79.59 (23.55)	Very confident	4.52 (14.80)
He/she resembles the criminal	He/She looks like the criminal	72.44 (28.52)	Not very confident	22.00 (26.20)
He/she looks familiar	He/She resembles the criminal	64.44 (26.59)	Very confident	23.81 (30.73)
Somewhat confident	Pretty sure	65.04 (23.44)	Very confident	26.04 (27.63)
Moderately confident	Somewhat confident	67.19 (25.96)	Not very confident	36.00 (31.98)
He/She looks like the criminal	He/She resembles the criminal	64.44 (21.85)	Not sure	37.15 (27.72)
I think it is him/her	He/She looks like the criminal	68.74 (32.08)	Not very confident	33.96 (29.89)
Pretty sure	Fairly confident	76.52 (23.59)	Not sure	26.63 (19.36)
Fairly confident	Quite confident	72.07 (21.83)	Not very confident	18.11 (20.54)
Quite confident	Fairly confident	76.48 (22.90)	Not very confident	17.44 (23.19)
Confident	Very confident	72.59 (24.47)	Not very confident	3.04 (5.60)
Very confident	Confident	86.74 (16.42)	Not very confident	6.19 (19.67)

Note. Mean similarity ratings for each phrase's synonym (i.e. phrase pairing with the highest similarity rating) and antonym (i.e. phrase pairing with the lowest similarity rating).

Figure 20

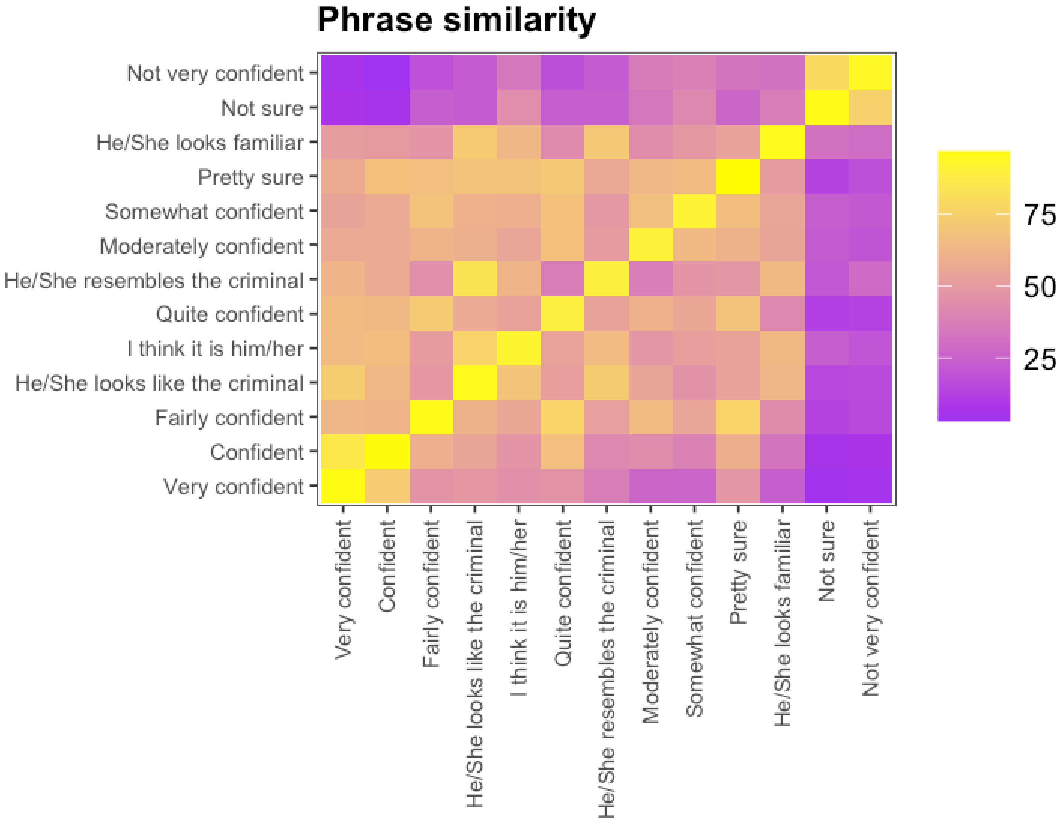
Mean Ratings for Pairings for the Four Phrases From Our Lexicon



Note. Similarity ratings for each phrase compared to all other phrases. Phrases rated to be most similar (i.e. synonyms) in comparison to each phrase are presented at the top of each graph, phrases rated to be most dissimilar (i.e. antonyms) are presented at the bottom of each graph.

Figure 21

Phrase Similarity Across All Phrase Pairings



Note. Mean ratings for similarity across all phrase pairs. Yellow indicates high similarity (i.e. yellow = 100%, completely the same), purple indicates low similarity (i.e. purple = 0%, completely opposite). Phrases at the low end (not very confident, not sure) and at the high end (very confident, confident) are clearly interpreted as similar. Interpretations for phrases spanning medium confidence are not clear.

Discussion

We asked participants to rate the sameness of or difference between eyewitness' verbal confidence statements. We tried to address the extent to which eyewitness' probability expressions (13 own-words confidence phrases) commonly used by eyewitness-participants in prior research (Chapter 2; Mansour, 2020) are interchangeable. That is, we asked participants to make pairwise judgements of (dis)similarity to determine the extent to which individuals judge expression of probability as synonyms/antonyms.

We expected to find *not very confident* and *not sure* to be judged similarly (i.e. to be synonymous). This hypothesis was supported. *Not very confident* was rated to be most similar to *not sure* (and vice versa). We also expected *very confident* and *confident* to be judged similarly. We found support for this hypothesis. *Very confident* was rated to be most similar to *confident*. This finding is in line with our previous work, suggesting that people reliably interpret certain low-confidence phrases (*not very confident, not sure*) and high confidence phrases (*very confident, confident*) in the same way (Chapter 2).

We hypothesized that *he/she resembles the criminal* and *he/she looks familiar* would be judged similarly. We found partial support for this hypothesis. *He/she looks familiar* was rated to be most similar to *he/she resembles the criminal*. However, *he/she resembles the criminal* was rated to be most similar to *he/she looks like the criminal*. This finding suggests that phrases like *he/she resembles the criminal* do not hold one precise meaning across people, replicating previous findings that medium confidence is not reliably interpreted (Mansour, 2020; Kenchel, et al., 2021; Chapter 2). However, people often use phrases that are fuzzy and vague in their meaning when asked to express uncertainty in a decision (Mandel & Irwin, 2021; Wintle et al., 2019), further highlighting the need for standardized approaches to the interpretation and presentation of difficult-to-interpret confidence phrases. Indeed, when people use such phrases, they should give triers of fact particular pause: across four studies we have shown that it is these phrases where misinterpretation of eyewitness confidence is most likely.

We expected *moderately confident* and *somewhat confident* to be judged similarly. We found partial support for this hypothesis. *Moderately confident* was rated to be most similar to *somewhat confident*, but *somewhat confident* was rated to be most similar to *pretty sure*. *Moderately confident* and *somewhat confident* were presented as synonyms in our previously established final lexicon (Chapter 2) but people consistently seem to interpret *moderately confident* more reliably than *somewhat confident* (Chapter 2; also cf. Table 9, p. 85). It may be that *moderately confident* holds more concise meaning across groups of people than *somewhat confident*. For example, there may be cultural differences that influence the differences in interpretation for the phrase *somewhat confident.* British people (majority in Study 2, Chapter 2 and the present study) seem to be interpret *somewhat confident* less consistently compared to North Americans (CloudResearch Amazon Mechnical Turk sample, Study 1, Chapter 2). It may be that the phrase *somewhat* is more frequently used in daily language exchanges in North America compared to the United Kingdom (e.g., Algeo, 1986; Dunkerley & Robinson, 2002. Future research should investigate the extent of culture on interpretations of eyewitness confidence phrases.

Lastly, we expected *not very confident* and *very confident* (and their respective synonyms, *not sure* and *confident*) to be judged as extremely different (i.e. be antonyms). This hypothesis was partially supported. *Not very confident* was judged to be antonymous to *very confident* and vice versa. However, *very confident* was also judged as antonymous to *not sure*, and *not very confident* was judged as antonymous to *confident*. Given that people seem to be able to clearly interpret *not very confident* and *very confident*, the use of these phrases may be more intuitive and comfortable than the use of other phrases. It may be that people rate antonymy based on the accessibility of interpretive processes (i.e. the ease with which associative meaning can be assigned to a phrase) when comparing phrase pairs. Information that is easy to process is believed to be learned well (ease-of-processing; Kornell, Rhodes, Castel, & Tauber, 2011), and thus maybe easier to interpret. It may be easier for people to compare to phrases that hold clear meaning (e.g., *not very confident*) to rate (dis)similarity. It could also be that people frequently use phrases like *not very confident* or *very confident* in their daily life and are thus simply more familiar with them (e.g., availability heuristic, Gabrielcik, & Fazio, 1984). In sum, making comparisons using phrases that hold concise meaning across individuals may be less cognitively demanding than using phrases that do not hold concise meaning. Renooij & Witteman (1999) also found that pairwise similarity judgements "forced interpretation of the expressions toward the endpoints of the scale" (p. 184). Previous work suggests that the public interprets expressions of probability in a regressive manner (i.e. they underestimate high probabilities and overestimate low probabilities; Budescu, Broomell & Por, 2009). Given that participants in our study rated *not very confident* and *very confident* to be antonyms to a majority of phrases (11 out of 13, 84.62%), we may infer that people naturally interpreted *not very confident* as the lowest end point, and *very confident* as the highest end point, compared to all other phrases, in our present study. This finding again replicates interpretation of these phrases in our lexicon (Chapter 2) and our previously established rank order.

Overall, we partially replicated previous findings indicating synonymity between phrases using a membership functions approach (Chapter 2). We replicated synonymity via pairwise comparisons for four phrase pairings (compared to Study 1, Chapter 2). The phrase pairings (*moderately confident/somewhat confident*, *very confident/confident*, *not very confident/not sure* and *fairly confident/quite confident*) were deemed synonymous in Study 1 (Chapter 2) and in the present study. *Moderately confident* and *quite confident* were both presented as core phrases in our lexicon representing medium confidence (final lexicon, Chapter 2). The findings support our notion that people seem to have shared understandings for certain phrases and seem to be able to interpret these phrases reliably.

While we replicated synonymity for four phrase pairings from Study 1 (Chapter 2), we did not replicate synonymity between other phrase pairings (e.g., *pretty sure*). It may be that phrases like *pretty sure* are not clearly interpreted or not judged to be synonymous with any other phrase. *Pretty sure* was not judged to be synonymous with any of our core lexicon phrases in either study (but represented a distinct range of values). This finding highlights the need for replication and calls into question the inclusion of *pretty sure* as a synonym in our final lexicon.

Previous research suggests that inter-personal agreement is high for similarity ratings between phrases (Rubenstein & Goodenough, 1965) but we do not know the extent to which this applies to phrases used to express confidence. Different methods of eliciting comparisons of similarity (e.g., membership functions versus direct phrase comparisons) may depend on different underlying cognitive mechanisms. For example, Windschitl and Wells (1996) theorized that probability estimates derived from deliberative, rule-based reasoning differ from those that do not require deliberation. That is, rule-based probability estimates are likely to be intuitively conveyed using numbers (e.g., 65% chance of precipitation) while associative judgements that do not necessitate deliberation may be better assessed using verbal probability estimates. It may be that membership functions (i.e. rating how well a number represents a phrase) rely on rule-based reasoning, while direct comparisons of (dis)similarity between phrases may depend on associative judgements. Further research should assess the underlying cognitive mechanisms when individuals are asked to determine synonymity using different methods.

While our findings provide further indication of synonymity between phrases used to express eyewitness confidence, there are limitations to our study. In Renooij & Witteman (1999), participants “made each judgement on a separate sheet of paper” (p. 182; i.e. presented phrase pairs sequentially). In our study, participants were presented with all phrases at once. It may be that our simultaneous presentation of phrases influenced the comparison process (i.e. led to relative judgements) between phrases. Future research should test ratings of (dis)similarity when phrase pairs are presented in isolation.

We recruited our participants via the university’s SONA recruitment system. Our participants were students completing the study for credit. 12 participants in our entire sample (*N* = 36) indicated they had participated in an eyewitness study before. It may be that judgements of synonymity of phrases are influenced by prior experience with eyewitness confidence phrases or by other types of relevant experiences. It may also be the case that individuals interpreting similarity in different settings attach different meanings to phrases based on the associated consequences or outcome severity. For example, a student completing a SONA study for credit may interpret *pretty sure* and *somewhat confident* differently than an intelligence analyst evaluating the likelihood of a terrorist attack. This suggests that similarity between phrases may vary to some extent based on context and user characteristics. Future research should explore the interchangeability of phrases when communicating uncertainty in varied settings.

In this study, we presented phrases in isolation (e.g., “pretty sure” instead of “I am pretty sure that is the person I saw”). However, previous work suggests “synonyms to be (…) words that can substitute for one another in sentences without changing meaning” (Herrman, 1978, p. 491; Ogden & Richards, 1923). Presenting phrases in isolation may limit misinterpretation of confidence statements. For example, hearing a piece of evidence in conjuction with other evidence (e.g., confessions) changes how evidence is interpreted (Hasel & Kassin, 2009). While eyewitnesses are continued to be asked to provide confidence in “their own words” in practice, presenting such statements in isolation may minimize the influences of social cues on interpretation of evidence. Given that triers of fact are asked to interpret confidence statements in context, future research should address the extent to which phrases are treated as interchangeable when presented in sentences.

Our research demonstrates the extent to which confidence expressions commonly used by eyewitness-participants in prior research are interchangeable. Determining the sameness of and difference between such phrases can improve triers of facts’ ability to establish common ground with eyewitnesses. Individuals share understandings of meaning for phrases representing low confidence (*Not very confident/ Not sure*) and high confidence (*Very confident/ Confident*). While our findings suggest that individuals use some phrases representing medium confidence interchangeably (*Pretty sure/ Fairly confident; Quite confident/ Fairly confident*), the extent to which individuals interpret other phrases representing medium confidence as interchangeable is less clear. In sum, individuals can consistently interpret verbal confidence phrases representing low and high confidence, but only interpret some medium confidence phrases reliably. Our research provides further evidence for synonymity between verbal confidence statements, particularly on the extreme ends of the confidence scale (i.e. low and high).

General Discussion

The overarching goal of our project was to develop and test reliable ways to obtain, interpret, and present confidence statements with an eye to minimizing the extent to which a game of telephone occurs between eyewitnesses and triers of fact. We first tested the effect of order on accuracy when obtaining both, verbal and numeric confidence statements. Results suggest eyewitnesses should provide only one confidence statement. Given that verbal confidence statements are commonly used in practice and generally preferred, this result encouraged us to aim to improve the ability of the criminal justice system to ensure that others interpret verbal expressions of eyewitness confidence in the way the eyewitness intended. To do so, we drew on advancements in the broader decision science literature to use a method that has proven effective in the fields of climate science and intelligence. Specifically, we developed a lexicon (i.e. translation tool) comprising four phrases (including three synonyms). Importantly, the phrases that comprise the lexicon have empirically-derived numeric meanings and the way we designed the lexicon to be used (i.e. its graphical appearance) is intended to make it easy to use for a variety of populations. But of course, all tools require validation and in this thesis, we began the lengthy process of validating the produced lexicon.

First, we replicated the rank order of the phrases in our lexicon and the synonymity of four phrase pairs. This research provides further evidence for synonymity between verbal confidence statements. Our findings suggest that interpretations of phrases are consistent across individuals for low (*not very confident*) and high confidence (*very confident; confident*), and for some phrases representing medium confidence (such as *quite confident; moderately confident*).

Ultimately, the legal system should seek to eliminate confounding factors, such as the potential for systematic misinterpretation of eyewitness confidence. A translation tool, such as our lexicon, could provide alternative approaches to minimize misinterpretation of eyewitness confidence statements. Our data shows that the development of such a lexicon is possible. Importantly, our lexicon provides a standardized documentation tool for the recording of initial confidence statements. Right now, there is no systematic data documenting how confident eyewitnesses were in their initial identification in wrongfully convicted cases (other than retrospective accounts given at trial, Garrett, 2011, p. 64). Using standardized methods to record eyewitness confidence statements may prevent subjective interpretation of eyewitness confidence (e.g., documentation by a police officer). Recording an initial confidence statement is a first step to documenting the statement judges and jurors should rely upon. By the time Ronald Cotton's case made it to court, Jennifer Thompson-Cannino was "certain" she identified the right person (Weir, 2016). Due to its malleable nature, confidence statements at trial, after a delay or provided retrospectively are not reliable. Had her initial statement been accurately obtained, interpreted and presented ("I think that is him"), jurors may not have wrongfully convicted Ronald Cotton.

While our findings highlight the potential of empirical interpretations for confidence judgements in the eyewitness area, there are limitations to our research. We do not know the extent to which alternative approaches perform in practice (but cf. Ho et al., 2015, for a comparison of lexicons in practice compared to evidence-based lexicons). Encouragingly, Ho et al. suggests that empirically-developed lexicons outperform those currently used by practitioners.

To test the practical utility of the lexicon, it is imperative to assess the performance of the lexicon on a lineup test (in the laboratory and in practice) in comparison to current methods to obtain eyewitness confidence (e.g., “in your own words” versus numeric). Other work suggests that confidence ratings obtained using a scale predict accuracy, irrespective of the type of scale presented (Dobolyi & Dodson, 2016). Does the lexicon lead to better calibration between confidence and accuracy compared to verbal and numeric approaches? If calibration is better (or equal) for the lexicon compared to verbal or numeric approaches, it may underline the value of obtaining confidence judgements via the lexicon in practice. That is, a more reliable tool will give law enforcement more information regarding the direction of their investigation.

Secondly, future research should test the extent to which intra- and inter-personal translation of confidence varies when individuals are asked to provide confidence in their own words versus numerically versus via lexicon. Does the lexicon lead to more consistent translation of confidence? There is variation in the translation of verbal to numeric confidence statements (Mansour, 2020). These discrepancies are even more pronounced when others are asked to interpret verbal eyewitness confidence statements (Mansour, 2020; Behrman & Richards, 2005; Smalarz et al., 2021). If the lexicon minimizes miscommunication of (verbal) confidence statements, there is reason to suggest the use of a lexicon to present eyewitness confidence statements in court settings. Additionally, we do not know the extent to which jurors’ and judges’ perception differ when eyewitness evidence is provided and presented via the lexicon. Do jurors and judges interpret confidence statements presented via the lexicon more accurately (as compared to verbal and numeric methods), or perceive them to be more accurate? Future research should address the extent to which a standardized translation tool minimizes misinterpretation and influences perceptions of jurors and judges.

Third, it may be that the extent to which people make decisions differs when mode of probability information varies (for example, when probability information is presented numerically or verbally, cf. Renooij & Witteman, Exp. 4). Do people make decisions with a similar level of confidence when the mode of probability information presented varies? For example, do people make similar decisions when probability information (e.g., confidence statements) is presented verbally, numerically, or via the lexicon? It would be useful to test the extent to which people make decisions with a level of confidence on varying decision situations when information is presented verbally, numerically or via the lexicon. Specifically, future research should test the extent to which people make decisions when outcome severity and context varies (Cash & Lane, 2017; Dodson & Dobolyi, 2015, Mosteller & Yout, 1982). Are interpretations of lexicon phrases consistent across different levels of outcome severity and context? If interpretations of the lexicon phrases vary, it would suggest that lexicons are context specific. It may mean that the confidence lexicon needs adaptation to accommodate a wider array of varying situations, or it may mean that lexicons should be developed and used for specific contexts and/or for a particular levels of outcome severity. If interpretations of lexicon phrases do not vary, it would suggest that the interpretations of these phrases are stable even when outcome severity and context changes.

Even though our research provides initial boundaries for verbal confidence phrases, we do not know what constitutes “sufficient confidence” for an identification to be deemed reliable. Agreeably, initial identifications made with low confidence, regardless of testing conditions, should be seen as highly prone to error (Wixted & Wells, 2017; Berkowitz, Garrett, Fenn, & Loftus, 2020). But what if an identification is made with 65%? 75%? “Quite confident”? 78.5%? We do not know what constitutes a sufficient threshold for reliability of eyewitness evidence. While research may not be able to address this question in its entirety, there are possible research avenues that could shed some light. For example, future research may investigate what people perceive to constitute “sufficient evidence” or reasonable doubt for eyewitness’ confidence. To what extent are such thresholds shared across people? What, if anything, malleableizes such thresholds? Future directions should explore the underlying mechanisms of the beliefs for sufficiency of evidence.

Berkowitz et al. (2020) state “it may not be possible to assure that a lineup is fairly constructed so that the suspect does not stand out, or that the eyewitness does not assume that police are presenting the lineup because they caught the culprit” (p. 10-15). We do not know the extent to which alternative methods to obtain confidence, such as our lexicon, perform under pristine and non-pristine conditions. It is worth investigating factors that might compromise the diagnosticity of confidence statements, including our lexicon. What happens when one (or more) identification procedures are not pristine (or, not pristine enough)? How pristine do conditions need to be? Is it more important for some procedures than others to be pristine? And does the use of a lexicon protect confidence in any way? For example, it may be that the effect of a failure to warn the eyewitness that the culprit may or may not be present in the lineup could be mitigated to some small extent by a lexicon because it highlights that uncertainty is an option. We do not know the full extent to which differing procedures affect the confidence-accuracy relationship.

Lastly, eliciting confidence statements using a lexicon may influence the reasoning processes underlying judgements of probability (e.g., Wintle et al., 2019). Semmler et al.'s constant likelihood ratio model suggests that individuals are aware of factors that influence their memory (and subsequent confidence judgements). From a theoretical standpoint, we do not know if such a constant likelihood ratio model also works for confidence. Moreover, we do not know what factors eyewitnesses believe affect their confidence. Are people aware of factors that influence confidence judgements but not memory (e.g., post-identification feedback)? To what extent do these factors influence memory if people do not expect it will affect their memory (or have not considered it)? Berkowitz et al. (2020) propose eyewitnesses' initial confidence statements "could artificially" be bolstered by pre-lineup experiences (e.g., seeing a mugshot in the newspaper and identifying the person from a lineup after seeing the mugshot in the newspaper). They note that real-world contamination of eyewitness confidence is complex as it may occur prior to the lineup decision (Gronlund & Benjamin, 2018, Berkowitz et al., 2020). We do not know if people are aware of such factors and whether they adjust their confidence. Alternatively, if individuals were made aware of such factors, would they adjust their confidence accordingly? Future research should address the influences of pre-lineup experiences on initial confidence statements.

Conclusion

Over the last 40 years, research has advanced our understanding of eyewitness evidence and its reliability in eyewitness identification. Due to the nature of the criminal justice system, triers of fact continue and will continue to rely on eyewitness confidence as an indicator of reliability. One of the main caveats to the assessment of eyewitness confidence as evidence is its fundamentally subjective interpretation. Triers of fact must interpret the intended level of confidence expressed by eyewitnesses to make decisions about the accuracy of eyewitnesses' identifications of suspects.

Undoubtedly, the increased risk for misinterpretation holds more weight in real-world settings than it does in the laboratory. However, not obtaining confidence at all (as common in policing practice around the world, e.g., Fitzgerald, Rubinova, & Juncu, 2021) is incongruous with fair administration of justice. Eyewitness confidence statements offer additional information about the accuracy of identifications and therefore could affect the outcome of a case. Failing to obtain and/or provide such evidence infringes on the process of criminal justice, especially when that evidence is heavily relied upon in court (e.g., Cutler, Penrod, & Stuve, 1988; Cutler & Penrod, 1995; Key et al., 2022; Slane & Dodson, 2022). The criminal justice system accepts that other forms of evidence are based on probabilities (e.g., DNA evidence, trace evidence, fingerprints, ballistic reports). Yet, eyewitness evidence is being treated as binary (identification versus no identification) when eyewitness confidence suggests it to be much more nuanced. Why should eyewitness evidence be treated differently than other types of forensic evidence?

Eyewitness confidence should be treated for what it is: A probability judgement about a memory-based decision. Not every identification is made with a 100% confidence and even if an eyewitness asserts a “100% confidence”, there is still a chance they may be wrong (Stretch & Wixted, 1998; Giacona, Lampinen & Anastasi, 2021; Wixted et al., 2018, p. 344; Dodson, 2020, p. 37). Identifications accompanied by verbal confidence statements that are evidently prone to be misinterpreted (e.g., “I think it is him”, Jennifer Thompson-Cannino as cited in Weir, 2016, p. 40) and/or known to be interpreted as “low confidence” (e.g., “not sure”) further call into question why identifications are presented as 100% (i.e. identification made, versus 0% no identification made), even when the eyewitness was initially uncertain in their identification (e.g., 34 cases of mistaken ID in which eyewitnesses testified to their initial uncertainty at trial, Garrett, 2011). There is currently no standard as to how verbal confidence statements are obtained (Garrett, 2011, p. 64; Berkowitz, Garrett, Fenn, & Loftus, 2020) and there is no standard as to how such verbal confidence statements are presented (cf. Wells, Kovera, Douglass, Brewer, Meissner, & Wixted, 2020 for recommendations, e.g., videorecording procedures). To make eyewitness evidence more reliable, the criminal justice system needs standardized, empirically-developed (and – ideally – easily implemented) tools that limit subjectivity in the assessment of evidence. We cannot expect the criminal justice system to adopt recommended practices without evidence that proposed approaches are in fact superior (e.g., numeric compared to verbal). But importantly, we also need to consider practicalities and that is what the lexicon is about—accommodating preferences and the need for good procedures. At the forefront of moving towards evidence-based practices for the obtainment, interpretation and presentation of eyewitness evidence stands the development of methods that provide the highest likelihood of a correct outcome—like a valid and easy-to-use eyewitness confidence lexicon.

Our research highlights the potential of empirical interpretations for probability judgements in the eyewitness area. Understandings of verbal confidence statements are shared and quantifiable across individuals. Our approaches to obtain, interpret and present eyewitness confidence statements can provide common ground for eyewitnesses and triers of fact when asked to provide and interpret verbal statements of confidence.

References

Algeo, J. (1986). The Two Streams: British and American English. *Journal of English Linguistics, 19*(2), 269–284. doi: 10.1177/007542428601900208

Ayala, N. T., Smith, A. M., & Ying, R. C. (2022). The rule-out procedure: Increasing the potential for police investigators to detect suspect innocence from eyewitness lineup procedures. *Journal of Applied Research in Memory and Cognition.* Advance online publication. doi: 10.1037/mac0000018

Ben-Shachar et al., (2020). Effectsize: Estimation of Effect Size Indices and Standardized Parameters. Journal of Open Source Software, 5(56), 2815. doi: 10.21105/joss.02815

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis, *Law and Human Behavior, 29*, 279-301. doi: 10.1007/s10979-005-3617-y

Berkowitz, S. R., Garrett, B. L., Fenn, K. M., & Loftus, E. F. (2020). Convicting with confidence? Why we should not over-rely on eyewitness confidence. *Memory.* Advance online publication. doi: 10.1080/09658211.2020.1849308

Boyce, M., Beaudry, J., & Lindsay, R. C. L. (2007). Belief of eyewitness identification evidence. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, 2,* 501-525. Lawrence Erlbaum Associates Publishers.

Bradfield, A. L., Wells, G. L., & Olson, E. A. (2002). The damaging effect of confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology, 87*(1), 112–120. doi: 10.1037/0021-9010.87.1.112

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness Identification Accuracy and Response Latency. *Law and Human Behavior, 30*(1), 31–50. doi: 10.1007/s10979-006-9002-7

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8*(1), 44–56. doi: 10.1037/1076-898X.8.1.44

Brewer, N., Vagadia, A. N., Hope, L., & Gabbert, F. (2018). Interviewing witnesses: Eliciting coarse-grain information. *Law and Human Behavior*, *42*(5), 458–471. doi:10.1037/lhb0000294

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*(1), 11–30. doi: 10.1037/1076-898X.12.1.11

Brewin, C. R., Andrews, B., & Mickes, L. (2020). Regaining Consensus on the Reliability of Memory. *Current Directions in Psychological Science, 29*(2), 121–125. doi: 10.1177/0963721419898122

Brigham, J. C., & Bothwell, R. K. (1983). The ability of prospective jurors to estimate the accuracy of eyewitness identifications. *Law and Human Behavior, 7*(1), 19–30. doi: 10.1007/BF01045284

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both?. *Organizational Behavior and Human Decision Processes, 41*, 390-404. doi: 10.1016/0749-5978(88)90036-2

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change. *Psychological Science*, *20*(3), 299–308. doi: 10.1111/j.1467-9280.2009.02284.x

Budescu, D. V., Karelitz, T. M., & Wallsten, T. S. (2003). Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making, 16*(3), 159– 180. doi: 10.1002/bdm.440

Budescu, D.V., Por, H. H. & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, 113, 181–200 (2012). doi: 10.1007/s10584-011-0330-3

Budescu, D., Por, HH., Broomell, S., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change,* **4,** 508–512. doi: 10.1038/nclimate2194

Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes*, *46*, 240-263. doi: 10.1016/0749-5978(90)90031-4

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391–405. doi: 10.1016/0749-5978(85)90007-X

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 281–294. doi: 10.1037/0096-1523.14.2.281

Bull Kovera, & M., Evelo A. J. (2020). Improving Eyewitness-Identification Evidence Through Double-Blind Lineup Administration. *Current Directions in Psychological Science*, *29*(6), 563-568. doi:10.1177/0963721420969366

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition, 6*(1), 82–92. doi: 10.1037/h0101806

Cash, D. K., & Lane, S. M. (2017). Context influences interpretation of eyewitness confidence statements. *Law and Human Behavior, 41*(2), 180–190. doi: 10.1037/lhb0000216

Clark, D.A. Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology,* 9, 203–235 (1990). doi: 10.1007/BF02686861

Clark, H. H., & Clark, E. V. (1977). Psychology and language.

Cofer, C. N. (1957). Associative commonality and rated similarity of certain words from Haagen's list. *Psychological Reports*, *3*(3), 603-606.

Collins, R. N., & Mandel, D. R. (2019). Cultivating credibility with probability words and numbers. Judgement and Decision Making, 14(6), 683-695. doi: 10.31234/osf.io/gkduw

Cumming, G., Williams, J. & Fiona Fidler, F. (2004). Replication and Researchers' Understanding of Confidence Intervals and Standard Error Bars, *Understanding Statistics, 3*(4), 299-311, doi: 10.1207/s15328031us0304_5

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology and the law*. Cambridge University Press.

Cutler, B.L., Penrod, S.D. & Martens, T.K. (1987). The reliability of eyewitness identification. *Law and Human Behavior, 11*, 233–258. doi: 10.1007/BF01044644

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12,* 41-55. doi: 10.1007/BF01064273

McQuiston-Surrett, D., & Saks, M. J. (2007). Communicating opinion evidence in the forensic identification sciences: Accuracy and impact. *Hastings Law Journal, 59*, 1159

Available at: https://repository.uchastings.edu/hastings_law_journal/vol59/iss5/7

Deffenbacher, K. A., & Loftus, E. F. (1982). Do jurors share a common understanding concerning eyewitness behavior? *Law and Human Behavior, 6*(1), 15–30. doi: 10.1007/BF01049310

Devlin, Lord P. (1976*). Report to the Secretary of State for the Home Department on the Departmental Committee on Evidence of Identification in Criminal Cases.* London: HMSO.

Dhami, Mandeep K. (2018). Towards an Evidence-Based Approach to Communicating Uncertainty in Intelligence Analysis. *Intelligence and National Security, 33*(2): 257–272. doi: 10.1080/02684527.2017.1394252.

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33*(6), 1057–1068. doi: 10.3758/BF03193213

Dodson, C. S. (2020). Distinguishing between reliable and unreliable eyewitnesses. *Judicature, 104(1),* 37-40.

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior, 39*(3), 266–280. doi: 10.1037/lhb0000120

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and Eyewitness Identifications: The Cross-Race Effect, Decision Time and Accuracy. *Applied Cognitive Psychology, 30,* 113-125. doi: 10.1002/acp.3178

Dobolyi, D. G., & Dodson, C. S. (2018). Actual vs. perceived eyewitness accuracy and confidence and the featural justification effect. *Journal of Experimental Psychology: Applied*, *24*(4), 543-563. doi: 10.1037/xap0000182

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior, 39,* 266-280. doi: 10.1037/lhb0000120

Dunkerley, K. J., & Robinson, W. P. (2002). Similarities and Differences in Perceptions and Evaluations of the Communication Styles of American and British

Mangers. *Journal of Language and Social Psychology*, *21*(4), 393–409. doi: 10.1177/026192702237956

Dunning, D. & Stern, L. (1994). Distinguishing Accurate From Inaccurate Eyewitness Identifications via Inquiries About Decision Processes. *Journal of Personality and Social Psychology*. *67*. 818-835. doi: 10.1037/0022-3514.67.5.818.

Fitzgerald, R. J., Rubínová, E., & Juncu, S. (2021). Eyewitness identification around the world. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks* (pp. 294-322). Taylor and Francis. doi: 10.4324/9781003138105-16

Gabrielcik, A., & Fazio, R., H. (1984). Priming and Frequency Estimation: A Strict Test of the Availability Heuristic. *Personality and Social Psychology Bulletin*. *10*(1):85-89. doi: 10.1177/0146167284101009

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press.

Giacona, A. M., Lampinen, J. M., & Anastasi, J. S. (2021). Estimator variables can matter even for high-confidence lineup identifications made under pristine conditions. *Law and Human Behavior, 45*(3), 256–270. doi: 10.1037/lhb0000381

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition, 8*(2), 233–243. doi: 10.1037/h0101835

Greenspan, R. L., & Loftus, E. F. (2020). Eyewitness confidence malleability: Misinformation as post-identification feedback. *Law and Human Behavior, 44*(3), 194–208. doi: 10.1037/lhb0000369

Hamm, R. M. (1991) Selection of verbal probabilities: a solution for some problems of verbal probability expression. Organizational Behavior and Human Decision Processes, 48, 193–223.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS one, 7,* e45457. doi: 10.1371/journal.pone.0045457

Harris, A. J. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1571–1578. doi: 10.1037/a0024195

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science, 20*(1), 122–126. doi: 10.1111/j.1467-9280.2008.02262.x

Herrmann, D. J. (1978). An old problem for the new psychosemantics: Synonymity. *Psychological Bulletin, 85*(3), 490–512. doi: 10.1037/0033-2909.85.3.490

Ho, E. H., Budescu, D. V., Dhami, M. K., Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioural Science and Policy, 1*(2), 53-66. doi: 10.1353/bsp.2015.0015

Innocence project. (n.d.) *Innocence project: Causes and remedies of wrongful convictions*. Retrieved 9th June 2022 from http://www.innocencecanada.com/causes.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1304-1316. doi: 10.1037/0278-7393.22.5.1304

Kebbell, M. R., & Milne, R. (1998). Police officers' perceptions of eyewitness performance in forensic investigations. *The Journal of Social Psychology, 138*(3), 323–330. doi: 10.1080/00224549809600384

Kenney, R. M. (1981). Between never and always. *The New England journal of medicine*, *305*(18), 1097–1098.

Kenchel, J. M., Greenspan, R. L., Reisberg, D., & Dodson, C. S. (2021). "In your own words, how certain are you?" Post-identification feedback distorts verbal and numeric expressions of eyewitness confidence. *Applied Cognitive Psychology, 35*(6), 1405– 1417. doi: 10.1002/acp.3870

Kenchel, J., Reisberg, D., & Dodson, C. S. (2017). "*In your own words, how certain are you?" Post- identification feedback powerfully distorts verbal expressions of witness confidence.* Paper at the American Psychology-Law Society. Seattle, U.S.A.

Key, N. K., Neuschatz, J. S., Gronlund, S. D., Deloach, D., Wetmore, S. A., McAdoo, R. M., & McCollum, D. (2022). High eyewitness confidence is always compelling:

that's a problem. Psychology, Crime & Law. Doi: 10.1080/1068316X.2021.2007912

Kong, A., Barnett, G. O., Mosteller, F., & Youtz, C. (1986). How medical professionals evaluate expressions of probability. *The New England Journal of Medicine, 315*(12), 740–744. doi: 10.1056/NEJM198609183151206

Kornell, N., Rhodes, M., G., Castel, A., D., & Tauber S., K. (2011). The Ease-of-Processing Heuristic and the Stability Bias: Dissociating Memory, Memory Beliefs, and Memory Judgements. *Psychological Science, 22*(6):787-794. doi: 10.1177/0956797611407929

Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior, 43*(5), 436–454. doi: 10.1037/lhb0000343

Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence-accuracy relationship in memory for people. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol. 2. Memory for people* (pp. 377–425). Lawrence Erlbaum Associates Publishers.

Ligertwood, A., & Edmond, G. (2012). Expressing evaluative forensic science opinions in a court of law, *Law, Probability and* Risk, *11*(4), 289–302. doi: 10.1093/lpr/mgs016

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9,* 215–218. doi: 10.1111/1467-9280.00041

Lockamyeir, R. F., Carlson, C. A., Jones, A. R., Carlson, M. A., & Weatherford, D. R. (2020). The effect of viewing distance on empirical discriminability and the confidence–accuracy relationship for eyewitness identification. *Applied Cognitive Psychology.* Advance online publication. doi: 10.1002/acp.3683

Loftus E.F. (1981). Reconstructive Memory Processes in Eyewitness Testimony. In: Sales B.D. (eds) The Trial Process. *Perspectives in Law & Psychology, 2*. Springer, Boston, MA.

MacLeod, A. and Pietravalle, S. (2017). Communicating risk: variability of interpreting qualitative terms. *EPPO Bull, 47*: 57-68. doi: 10.1111/epp.12367

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66*(4), 482–489. doi: 10.1037/0021-9010.66.4.482

Malpass, R. S, Tredoux, C. G. & McQuiston-Surrett, D. E. (2007). Lineup construction and lineup fairness. in R. C. L. Lindsay, D. F. Ross, J. D. Read & M. P. Toglia (Eds.), The Handbook of Eyewitness Psychology (Vol. II): Memory for People. Lawrence Erlbaum & Associates.

Mandel D. R. (2015) Accuracy of Intelligence Forecasts From the Intelligence Consumer's Perspective. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1):111-120. doi:10.1177/2372732215602907

Mandel, D. R., Wallsten, T. S., & Budescu, D. V. (2021). Numerically bounded linguistic probability schemes are unlikely to communicate uncertainty effectively. *Earth's Future,* 9, e2020EF001526. doi: 10.1029/2020EF001526

Mandel, D. R., & Iwrin, D. (2021). Facilitating sender-receiver agreement in communicated probabilities: Is it best to use words, numbers or both?. *Judgement and Decision Making, 16*(2), 363-393. doi: 10.1371/journal.pone.0248424

Mansour, J. K. (2020). The confidence-accuracy relationship using scale versus other methods of assessing confidence. *Journal of Applied Research in Memory and Cognition, 9*(2), 215–231. doi: 10.1016/j.jarmac.2020.01.003.

Mansour, J. K. & Vallano, J. P. (2022, March 17-19). *Does "very confident" mean very confident? Perceptions of low, medium, & high eyewitness confidence.* [Conference presentation]. American Psychology-Law Society Conference, Denver, CO, United States.

Merz, J. F., Druzdzel, M. J., & Mazur, D. J. (1991). Verbal Expressions of Probability in Informed Consent Litigation. *Medical Decision Making*, *11*(4), 273–281. doi: 10.1177/0272989X9101100405

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*(2), 93–102. doi: 10.1016/j.jarmac.2015.01.003

Newman, S. E., & Williams, C. M. (1976). Response speed for easy- and hard-to-pronounce trigrams. Journal of Verbal Learning and Verbal Behavior, 6(4). 661-667. doi: 10.1016/S0022-5371(67)80032-X

Ogden, C. K., & Richards, I. A. (1923). The meaning of meaning: A study of the influence of thought and of the science of symbolism.

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. Law and Human Behavior, 42, 1-12. doi: 10.1037/lhb0000272

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55–71. doi: 10.1037/a0031602

Pennekamp, P., Batstone, R. J., & Mansour, J. K. (2019, December 9-10). *Eyewitness Identification Confidence: Requesting, articulating, and apperceiving* [Poster presentation]. Scottish Institute for Policing Research Conference, Edinburgh, U.K.

Pennekamp, P., & Mansour, J. K. (2021, August 24). *Does order matter? Numeric and verbal eyewitness confidence* [Poster presentation]. European Association for Psychology and Law Conference (Virtual Edition). United Kingdom.

Pennekamp, P., & Mansour, J. K. (2022, March 17-19). *Confidence Lexicon: An evidence-based approach for interpreting eyewitness confidence* [Conference presentation]. American Psychology-Law Society Conference, Denver, CO, United States.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1*(4), 817–845. doi: 10.1037/1076-8971.1.4.817

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627-633.

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*(4), 337-347. doi: 10.1007/s10979-009-9192-x

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in Using Eyewitness Confidence to Diagnose the Accuracy of an Individual Identification Decision. *Psychology, Public Policy, and Law*. *25*(3), 147-165. doi: 10.1037/law0000203

Searle, J. R. (1969). How to derive 'ought' from 'is'. *The is-ought question*. 120-134. Palgrave Macmillan, London.

Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. (2020, November 3). The language of accurate and inaccurate eyewitnesses. doi: 10.31234/osf.io/dhzk2

Semmler, C., Brewer, N., & Douglass, A. B. (2011). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), Conviction of the innocent: Lessons from psychological research. 185-209. Washington, DC: APA Books.

Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of Postidentification Feedback on Eyewitness Identification and Nonidentification Confidence. *Journal of Applied Psychology, 89(*2), 334–346. doi: 10.1037/0021-9010.89.2.334

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*(3), 400–415. doi: 10.1037/xap0000157

Shaw, J. S. III, Appio, L. M., Zerr, T. K., & Pontoski, K. E. (2007). Public eyewitness confidence can be influenced by the presence of other witnesses. *Law and Human Behavior, 31*(6), 629–652. doi: 10.1007/s10979-006-9080-6

Shaw, J. S. III, & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior, 20*(6), 629–653. doi: 10.1007/BF01499235

Slane, C. R., & Dodson, C. S. (2022). Eyewitness confidence and mock juror decisions of guilt: A meta-analytic review. *Law and Human Behavior, 46*(1), 45–66. doi: 10.1037/lhb0000481

Smalarz, L., Yang, Y., & Wells, G. L. (2021). Eyewitnesses' free-report verbal confidence statements are diagnostic of accuracy. *Law and Human Behavior, 45*(2), 138–151. doi: 10.1037/lhb0000444

Smith, A. M., Smalarz, L., Ditchfield, R., & Ayala, N. T. (2021). Evaluating the claim that high confidence implies high accuracy in eyewitness identification. *Psychology, Public Policy, and Law, 27*(4), 479–491. doi: 10.1037/law0000324

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing Between Investigator Discriminability and Eyewitness Discriminability: A Method for Creating Full Receiver Operating Characteristic Curves of Lineup Identification Performance. *Perspectives on psychological science: a journal of the Association for Psychological Science, 15*(3), 589–607. doi: 10.1177/1745691620902426

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78*(1), 22–33. doi: 10.1037/0021-9010.78.1.22

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315–327. doi: 10.1037/0033-2909.118.3.315

Steblay, N. M. (1997). Social Influence in Eyewitness Recall: A Meta-Analytic Review of Lineup Instruction Effects. Law and Human Behavior, 21. 283-297. doi: 10.1023/A:1024890732059

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379–1396. doi: 10.1037/0278-7393.24.6.1379

Tekin, E., Lin, W. & Roediger, H. L. (2018). The relationship between confidence and accuracy with verbal and verbal + numeric confidence scales. *Cognitive Research, 3*(41). doi: 10.1186/s41235-018-0134-3.

Vredeveldt, A., & Sauer, J. D. (2015). Effects of eye-closure on confidence-accuracy relations in eyewitness testimony. *Journal of Applied Research in Memory and Cognition*, *4*(1), 51–58. doi: 10.1016/j.jarmac.2014.12.006

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgements. *Management Science, 39*, 176-190. doi: 10.1287/mnsc.39.2.176

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*, 348-365. doi: 10.1037/0096-3445.115.4.348

Weir, K., (2016). Mistaken Identity: Is eyewitness identification more reliable than we think? *Monitor on Psychology, 47*(2)

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*(12), 1546–1557. doi: 10.1037/0022-3514.36.12.1546

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*(3), 360–376. doi: 10.1037/0021-9010.83.3.360

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification-feedback effect be moderated? *Psychological Science, 10*(2), 138–144. doi: 10.1111/1467-9280.00121

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*(1), 3–36. doi: 10.1037/lhb0000359

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603-647. doi: 10.1023/A:1025750605807

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. doi: 10.1037/1076-898X.2.4.343

Wintle, B.C., Fraser, H., Wills, B.C., Nicholson, A.E., & Fidler, F. (2019). Verbal probabilities: *Very likely* to be *somewhat* more confusing than numbers. *PLOS ONE 14*(4): e0213522. doi: 10.1371/journal.pone.0213522

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262–276. doi: 10.1037/a0035940

Wixted, J.T., Mickes, L., Dunn, J. C., & Wells, W. (2015). Estimating the reliability of eyewitness identifications from police lineups. *PNAS, 113*(2). 304-309. doi: 10.1073/pnas.1516814112

Wixted, J. T., Mickes, L., Brewin, C. R. & Andrews, B. (2021). Doing right by the eyewitness evidence: a response to Berkowitz et al., *Memory, 30*(1), 73-74, doi: 10.1080/09658211.2021.1940206

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*(6), 515–526. doi: 10.1037/a0039510

Wixted, J.T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2015). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences. 113*(2). 304-309. doi: 10.1073/pnas.1516814112

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10-65. doi: 10.1177/1529100616686966

Zimmer, A. C. (1983). Verbal versus numerical processing of subjective probabilities. In R. W. Scholz (Ed.), Decision making under uncertainty (pp. 159-182). Amsterdam: North-Holland.

Appendix 1

Information sheet

Queen Margaret University
EDINBURGH

My name is Pia Pennekamp and I am a PhD Candidate in Psychology at Queen Margaret University in Edinburgh. The purpose of this research is to understand eyewitness identification.

Everyone is welcome to participate in this study as long as they can see a standard computer screen; it is ok if you need glasses or another aid to help you do so, though.

By consenting to this study, you agree to watch a video of an event. You will then be asked questions about this event. The researcher is not aware of any risks associated with answering these questions and no personally-identifying information will be collected.

The whole procedure should take no longer than 15 minutes. You are free to withdraw from the study at any stage and you do not have to give a reason. The results may be published and/or presented in academic settings (such as conferences or classes).

If you would like to contact an independent person, who knows about this project but is not involved in it, you are welcome to contact Dr Olivia Sagan. Her contact details are given below.

Name of researcher: Pia Pennekamp
Address: PhD Candidate, Psychology, Sociology, & Education
Queen Margaret University
Edinburgh, UK EH21 6UU
Email / Telephone: ppennekamp@qmu.ac.uk / 0131 474 0000

Independent adviser: Olivia Sagan
Address: Head of Division, Psychology, Sociology & Education
Queen Margaret University
Edinburgh, UK EH21 6UU
Email / Telephone: osagan@qmu.ac.uk / 0131 474 0000

If you have read and understood the information presented, you have no questions, and you wish to consent to participate, please click the "I consent to participate" button.

By clicking the "I consent to participate" button, you are indicating that you:

- understand that you are under no obligation to take part in this study
- understand that you have the right to withdraw from this study at any stage without giving any reason
- agree to participate in this study.

Appendix 2

Figure 22

Mock-crime Videos

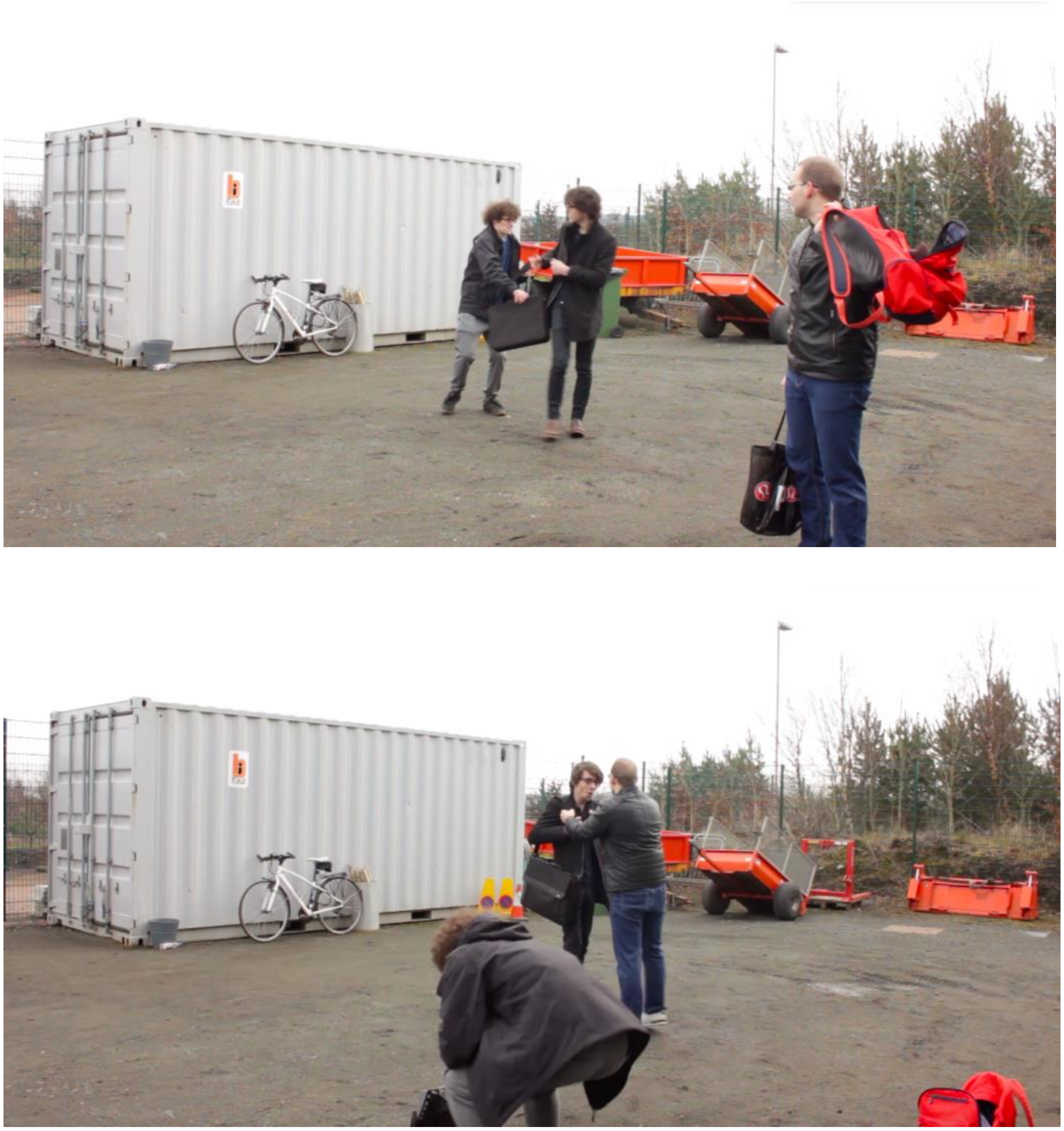
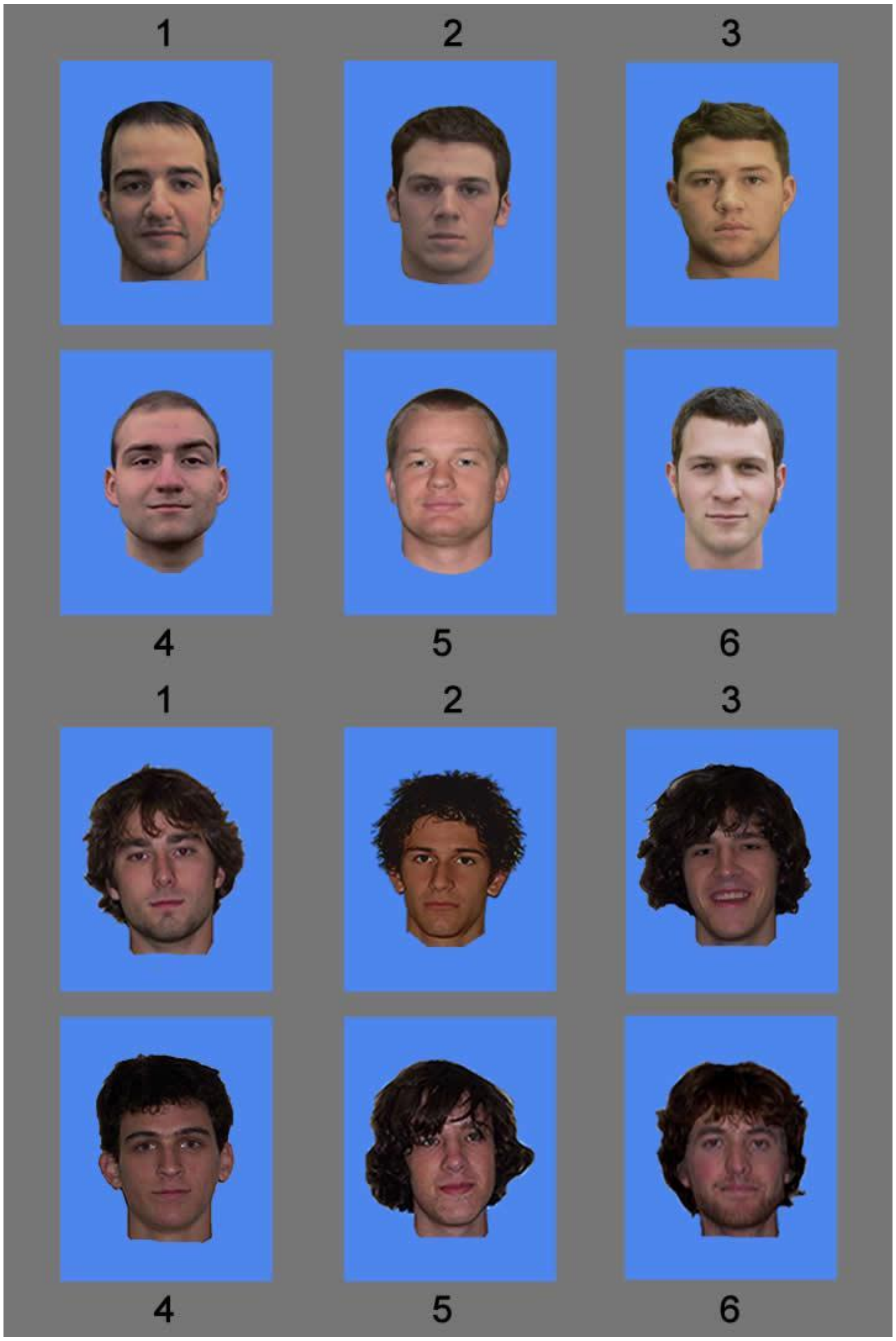


Figure 23

Presentation of Lineups

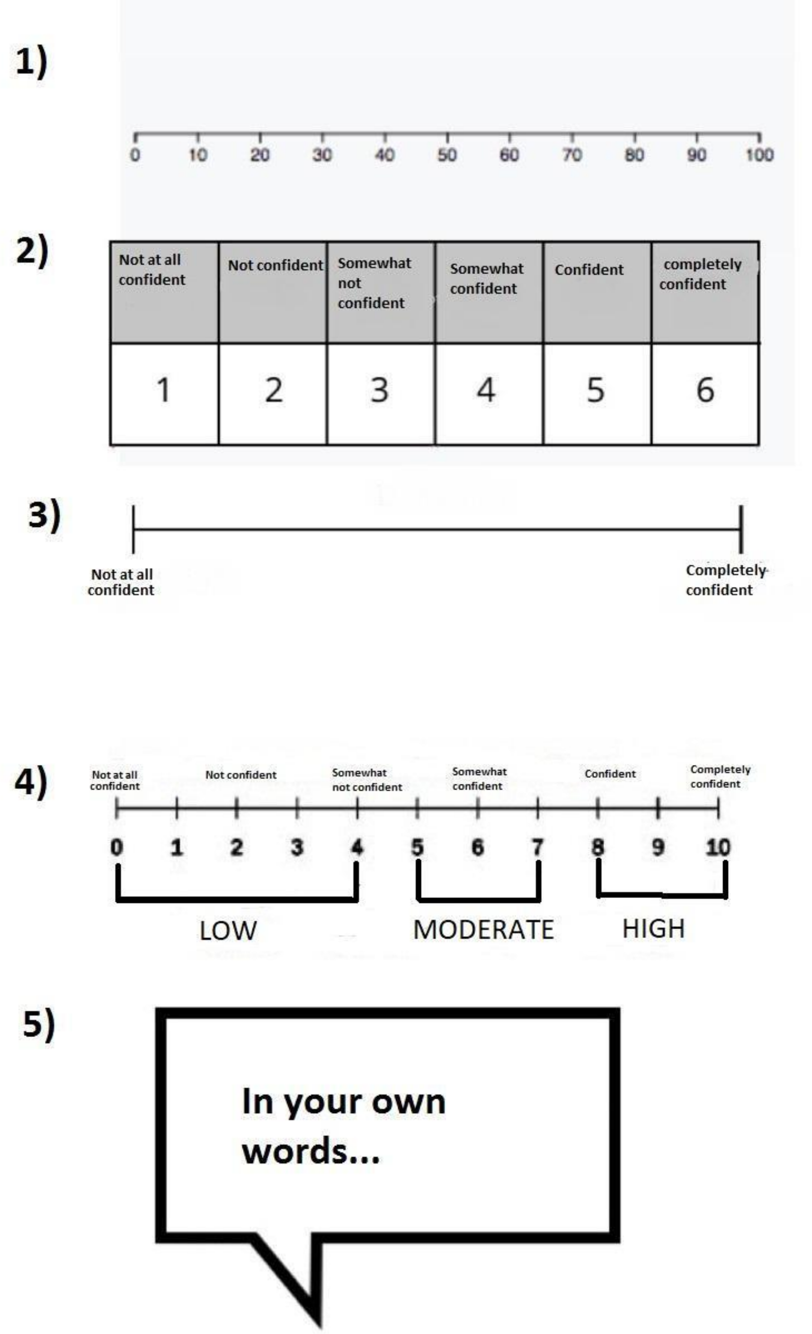


Appendix 3

Figure 24

Scale Preferences

In your opinion, which is the best way to ask someone for their confidence?

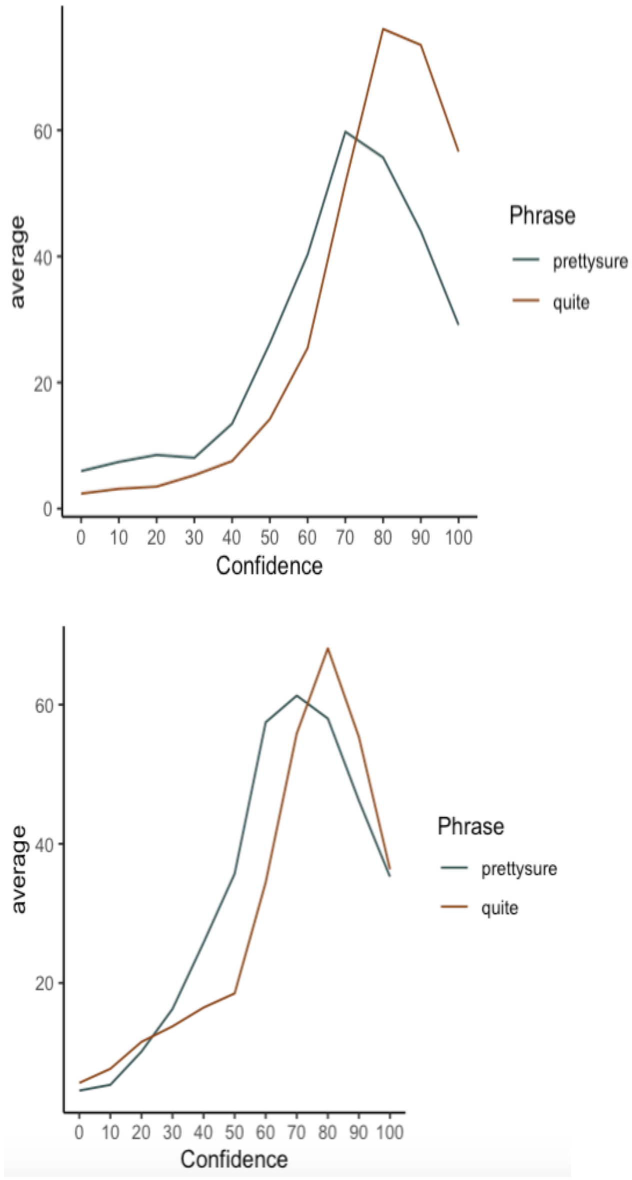


Not at all confident	Not confident	Somewhat not confident	Somewhat confident	Confident	completely confident
1	2	3	4	5	6

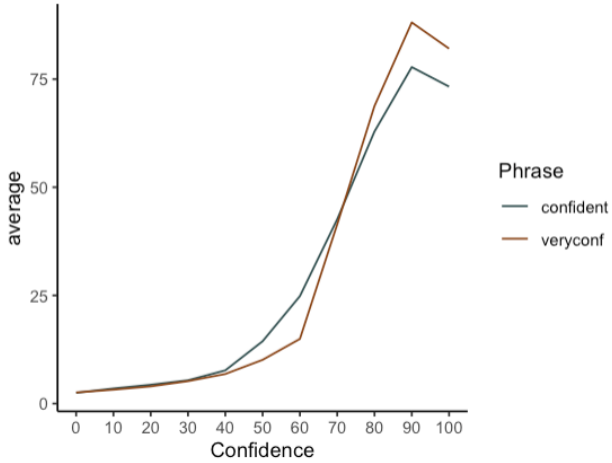
Appendix 4

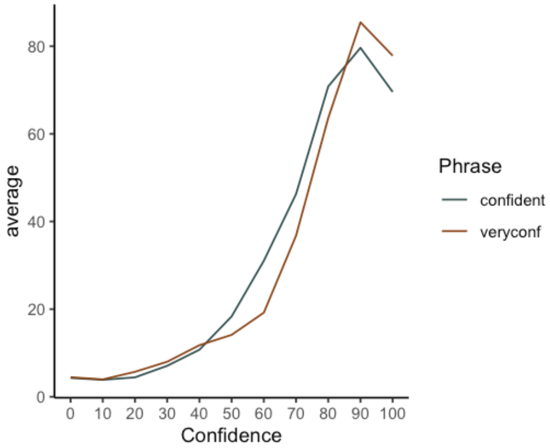
Figure 25

Comparison of "Quite confident" and "Pretty sure" in Study 1 and Study 2



Comparison of "Confident" and "Very confident" in Study 1 and Study 2

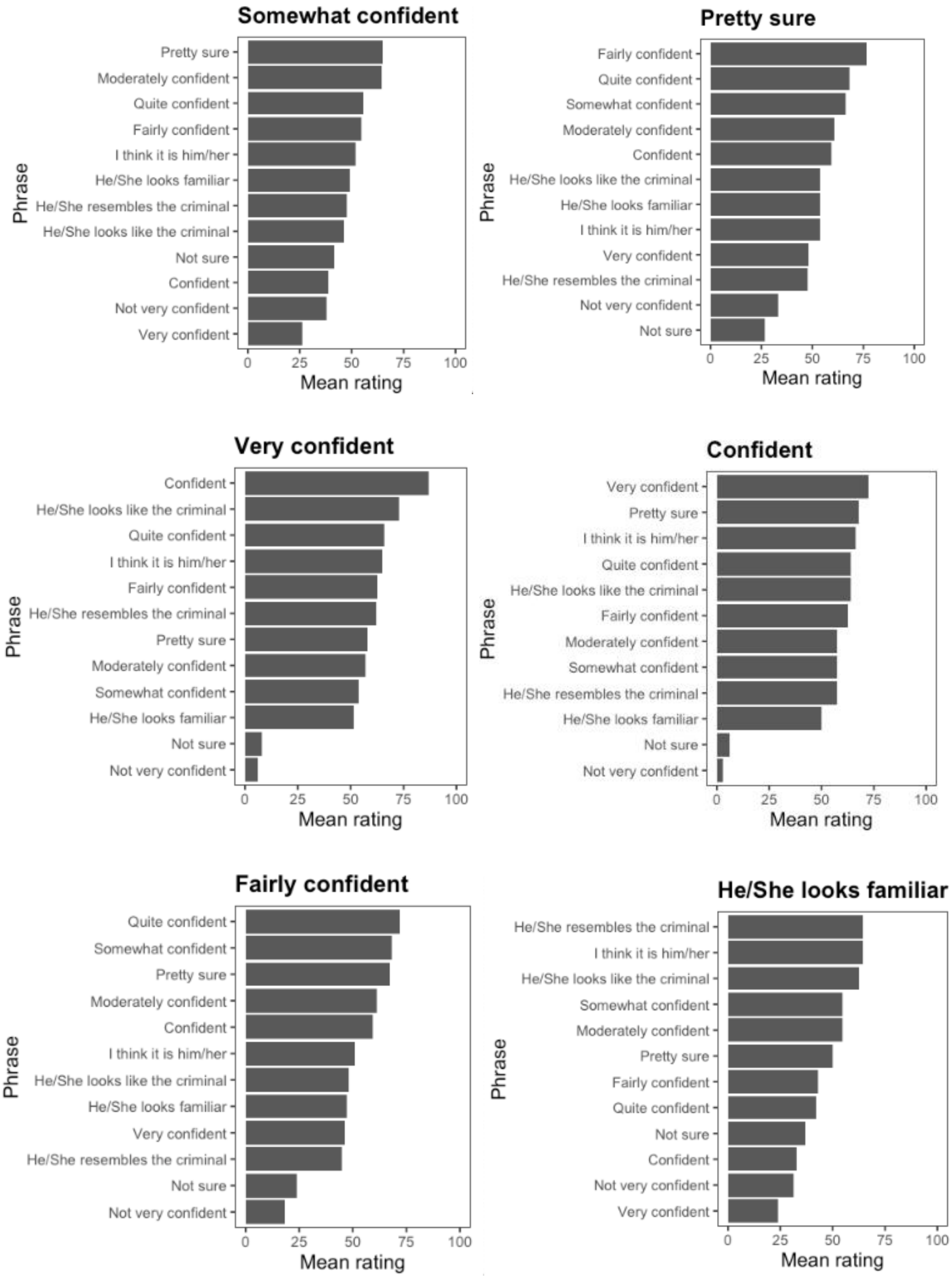

average
0
25
50
75
Confidence
0
10
20
30
40
50
60
70
80
90
100
Phrase
confident
veryconf

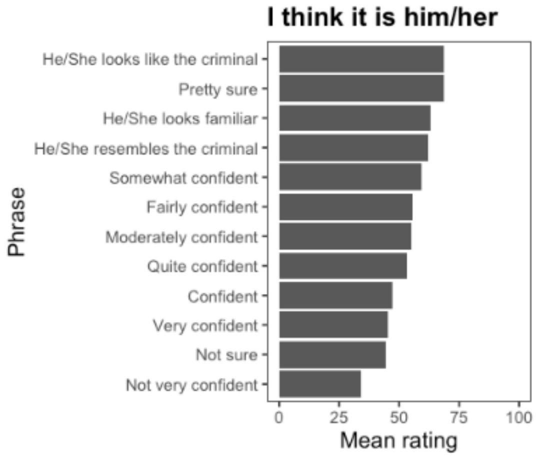

average
0
20
40
60
80
Confidence
0
10
20
30
40
50
60
70
80
90
100
Phrase
confident
veryconf

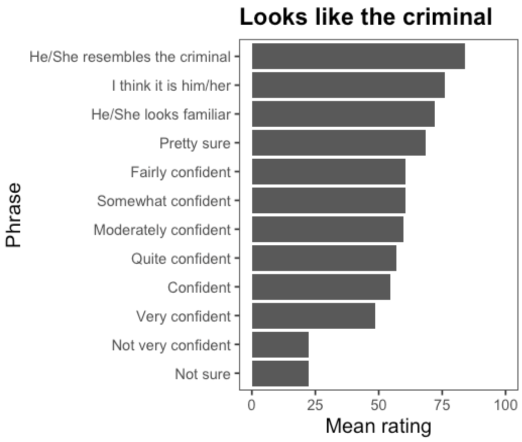
Appendix 5

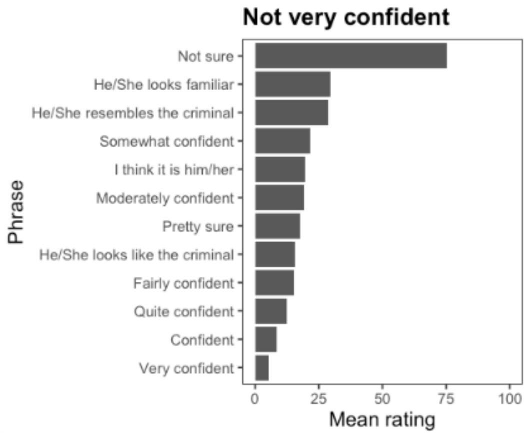
Figure 26

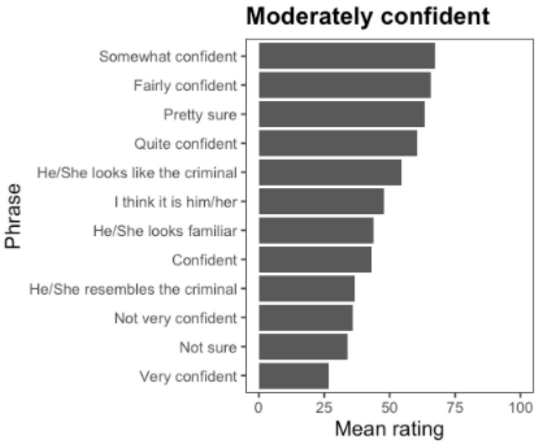
Mean Ratings for all Phrase Pairings.

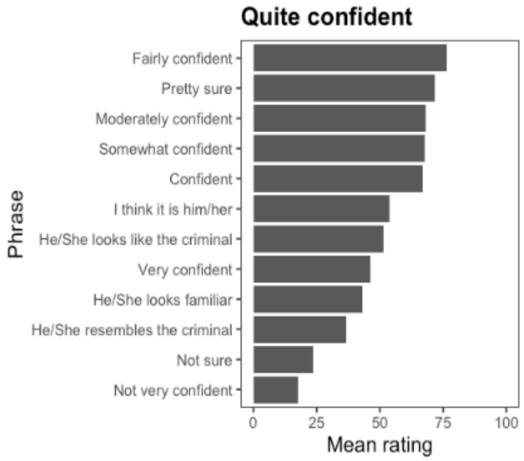


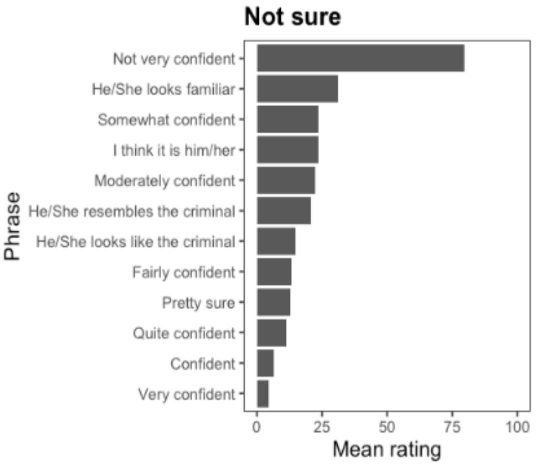

I think it is him/her
He/She looks like the criminal
Pretty sure
He/She looks familiar
He/She resembles the criminal
Somewhat confident
Fairly confident
Moderately confident
Quite confident
Confident
Very confident
Not sure
Not very confident
Phrase
0
25
50
75
100
Mean rating

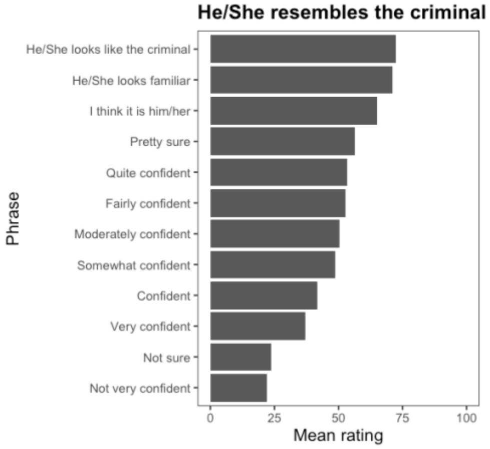

Looks like the criminal
He/She resembles the criminal
I think it is him/her
He/She looks familiar
Pretty sure
Fairly confident
Somewhat confident
Moderately confident
Quite confident
Confident
Very confident
Not very confident
Not sure
Phrase
0
25
50
75
100
Mean rating


Not very confident
Not sure
He/She looks familiar
He/She resembles the criminal
Somewhat confident
I think it is him/her
Moderately confident
Pretty sure
He/She looks like the criminal
Fairly confident
Quite confident
Confident
Very confident
Phrase
0
25
50
75
100
Mean rating


Moderately confident
Somewhat confident
Fairly confident
Pretty sure
Quite confident
He/She looks like the criminal
I think it is him/her
He/She looks familiar
Confident
He/She resembles the criminal
Not very confident
Not sure
Very confident
Phrase
0
25
50
75
100
Mean rating


Quite confident
Fairly confident
Pretty sure
Moderately confident
Somewhat confident
Confident
I think it is him/her
He/She looks like the criminal
Very confident
He/She looks familiar
He/She resembles the criminal
Not sure
Not very confident
Phrase
0
25
50
75
100
Mean rating


Not sure
Not very confident
He/She looks familiar
Somewhat confident
I think it is him/her
Moderately confident
He/She resembles the criminal
He/She looks like the criminal
Fairly confident
Pretty sure
Quite confident
Confident
Very confident
Phrase
0
25
50
75
100
Mean rating



Note. Similarity ratings for each phrase compared to all other phrases. Phrases rated to be most similar (i.e. synonyms) in comparison to each phrase are presented at the top of each graph, phrases rated to be most dissimilar (i.e. antonyms) are presented at the bottom of each graph.