Received: 28 July 2023 | Revised: 14 February 2024 | Accepted: 29 February 2024

DOI: 10.1002/acp.4190

RESEARCH ARTICLE

WILEY

Variability in verbal eyewitness confidence

Pia Pennekamp[1,2] | Jamal K. Mansour[3] | Rhiannon J. Batstone[3]

[1]Memory Research Group; Centre for Applied Social Sciences; Division of Psychology, Sociology and Education, Queen Margaret University, Musselburgh, UK

[2]Department of Psychological Science, University of Arkansas, Fayetteville, Arkansas, USA

[3]Department of Psychology, University of Lethbridge, Lethbridge, Alberta, Canada

Correspondence
Pia Pennekamp, Department of Psychological Science, University of Arkansas, Fayetteville, AR, USA.
Email: piap@uark.edu

Abstract

Typically, an eyewitness' verbal confidence is used to judge the reliability of their lineup identification. Across three experiments ($N = 3976$), we examined eyewitnesses' own words confidence in their lineup decision. For identification decisions ($n = 1099$), we identified 781 quantitatively unique responses representing 132 qualitatively unique statements that could be categorized into low, medium, and high confidence. For rejectors ($n = 781$), we identified 599 quantitatively unique responses representing 143 qualitatively unique responses that could be categorized into low, medium, and high confidence. Most participants provided a verbal phrase (e.g., pretty sure) but a significant proportion—34.19% of identifiers and 29.05% of rejectors—provided numbers (e.g., 80%). The present data highlight the variability in how confidence is expressed. The criminal justice system would benefit from guidance for interpreting verbal confidence. We provide a picture of eyewitnesses' verbal confidence as a first step.

KEYWORDS
communication, confidence, eyewitness identification, interpretation, language

1 | INTRODUCTION

Eyewitness evidence is compelling and heavily relied upon in court (e.g., Cutler et al., 1988; Key et al., 2022; Slane & Dodson, 2022). Eyewitness accounts, especially when given with high confidence, are very persuasive (Boyce et al., 2007; Cutler et al., 1990; Key et al., 2022). An eyewitness' confidence in their identification decision can predict accuracy under certain conditions (Wixted & Wells, 2017). However, there is currently no universal standard as to how eyewitness confidence should be obtained, interpreted, or presented. In the United States and Canada, confidence—if collected–is typically collected verbally (in the eyewitness' own words). However, methods to collect confidence also vary across jurisdictions (e.g., Portland police officers are explicitly discouraged from obtaining confidence numerically; Police Bureau Portland, 2023). In the UK, an eyewitness' confidence is as assumed from their decision: 0% confident (no identification) or 100% confident (identification), although, if an eyewitness spontaneously provides a confidence judgment, it is recorded and available as evidence.

Importantly, the collection of confidence in practice differs from research. Researchers typically collect scale judgments (e.g., 0%–100%, with 1% or 10% increments or on a 1-7 Likert-style scale) but, in practice where policies recommend confidence be collected, it tends to be in the eyewitness' own words. Encouragingly, verbal and numeric confidence have been shown to be similarly predictive of identification accuracy (e.g., Mansour, 2020). Verbal confidence statements provide unique diagnostic information (Seale-Carlisle et al., 2022; Steblay & Wells, 2023), but oftentimes, the language used is vague and subject to misinterpretation (Budescu et al., 2009).

The research showing that verbal confidence predicts identification accuracy has relied on a wisdom of the crowd approach to categorizing these verbal judgments so that their ability to predict

accuracy can be compared with other ways of measuring confidence. Wisdom of the crowd approaches improve the accuracy of interpretations because they rely on multiple estimates rather than a single judgment (Budescu & Chen, 2015; Karelitz & Budescu, 2004; Surowiecki, 2004). On average, the modal judgment from a group is more accurate than the judgment of an individual (Larrick et al., 2012). This has been shown to be true for confidence judgments when the common response is the correct response (Litvinova et al., 2020).

Accordingly, Mansour et al. (2020) categorized phrases based on Behrman and Richards' (2005) work. Behrman (2004, as cited in Behrman & Richards, 2005) asked participants to estimate ranges for low, medium, and high confidence on a 0–10 confidence scale. The mean ranges were 0–4, 5–7, and 8–10 for the three confidence categories. Behrman and Richards (2005) then examined 35 verbal confidence statements obtained from real eyewitnesses. Participants rated the verbal confidence statements on a numeric scale from 0 = *No confidence* to 10 = *Absolutely certain*. Based on the mean numeric ratings and the confidence ranges Behrman had obtained, Behrman and Richards coded the 35 statements into low, medium, and high confidence. The categorization by Behrman and Richards involved using the "root" phrases in each eyewitness' confidence judgment. For example, if an eyewitness stated, "I'm very confident, his hair looks the closest," that eyewitness would be considered to have been "very confident" and categorized as highly confident. To date, Behrman and Richards' coding scheme is the only empirically-based coding scheme to interpret eyewitness confidence statements. For phrases not captured by Behrman and Richards' Table 1, Mansour extended Behrman and Richards' coding scheme by having new participants provide 0–10 ratings following Behrman and Richards' method and then categorized those phrases based on the mean rating obtained. To preserve the variability in eyewitness' verbal confidence statements, Smalarz et al. (2021) obtained confidence phrases via free report ("using words" in the verbal condition, or "using percentage terms" in the numeric condition). They then asked others to interpret the full (cf. root) verbal confidence statements on a scale of 0%–100% and reported the average interpretations. Finally, Arndorfer and Charman (2022) had an individual judge interpret all of the full (verbal) confidence phrases that eyewitnesses provided so they could compare them to numeric scale and verbal-numeric scale judgments.

The publications just described differ from each other in terms of two dimensions: First, in terms of whether a root phrase (Behrman & Richards, 2005; Mansour, 2020) versus full confidence phrases (Arndorfer & Charman, 2022; Smalarz et al., 2021) were interpreted. Second, in terms of how many people were responsible for the interpretation: one (Arndorfer & Charman, 2022) or more (Behrman & Richards, 2005; Mansour, 2020; Smalarz et al., 2021). In practice, a full confidence judgment is interpreted by an individual (e.g., the lead detective, a prosecutor, or a judge, a juror) and the specific individual varies across cases (cf. Arndorfer & Charman, 2022 where one rater interpreted all phrases). While prior research suggests verbal confidence is predictive of accuracy, in all cases, the verbal confidence judgments were systematically interpreted (i.e., Arndorfer & Charman, 2022; Behrman & Richards, 2005; Mansour, 2020; Smalarz et al., 2021). Thus, a nonnegligible risk in practice is that an individual's idiosyncratic interpretation may not match that of a different individual interpreter or, more concerningly, the eyewitness' intended interpretation.

The body of literature on how to interpret verbal confidence judgments in relation to eyewitness evidence is limited but already demonstrates considerable variability. Mansour et al. (2020) found the correspondence between participant-eyewitness' numeric confidence and the coded verbal confidence judgments was strong for high confidence (>90% of statements) but not medium (44%–75%) or low confidence (25%–79%). These misinterpretations may reflect the fact that Mansour and other researchers have categorized based on root phrases, which increases the opportunity for misinterpretation because the "full picture" of the eyewitness' confidence is not used. On the other hand, it may have minimized variability because the statements provided fewer cues to consider. A related problem is that people assign different meaning to verbal confidence statements than intended by the eyewitness. For example, eyewitnesses referring to observable features (e.g., distinctive nose) are judged as less accurate than those who do not refer to such features although they are similarly accurate (Dodson & Dobolyi, 2015; Dobolyi & Dodson, 2018). Smalarz et al. (2021) found others (e.g., jurors and judges) underestimated the level of confidence an eyewitness expressed when confidence was provided verbally. Words (e.g., pretty sure) do not convey the same precision as numbers (e.g., 70%).

Nevertheless, people prefer to express confidence verbally (Brun & Teigen, 1988; Budescu et al., 2003; but cf. Kenchel et al., 2021; Smalarz et al., 2021 who found that their participant-eyewitnesses preferred numeric confidence), even while preferring to hear confidence numerically (i.e., the preference paradox; also see Greenspan & Loftus, 2023). Given that the criminal justice system relies on an eyewitness' confidence as an indicator of how reliable their evidence is reliable methods for interpretation are needed. As a first step and to justify this position, we provide a picture of the variability in eyewitness verbal confidence.

2 | METHOD

All studies were approved by the university's research ethics board. All studies were preregistered (Experiment 1: https://osf.io/hvy87/?

	Suspect ID		Filler ID		Rejection	
Lineup	*N*	Proportion	*N*	Proportion	*N*	Proportion
Target-present	489	0.52	276	0.29	179	0.19
Target-absent	N/A	N/A	334	0.36	602	0.64

TABLE 1 Proportion lineup responses in the own words conditions across experiments.

view_only=95fa206d9e184f62b0ab0987fe907019; Experiment 2: https://osf.io/2taec/?view_only=2ff6d3ba9ac74de6be06749645c0073d; Experiment 3: https://osf.io/d5gcr/?view_only=ec3f74f58966451a91e887572109049a).

2.1 | Participants

We report all demographic information collected for each experiment. All experiments recruited at least in part from CloudResearch/Amazon Mechanical Turk (heretofore MTurk).

In Experiment 1, MTurk was used to recruit participants. All workers had to have a HIT approval rate greater than 75% and have had more than 100 HITs approved. The sample after exclusions ($N = 968$) of participants was primarily female (58%; .001% indicated other and .004% did not respond) with a mean age of 38.64 years (SD = 12.89, Range = 18–82). The usable sample ($n = 912$) did not include participants who failed the attention check ($n = 56$). In Experiment 1, there were 125 own-words confidence identifications and 111 own-words confidence rejections.

In Experiment 2, participants were again recruited through MTurk, using the same requirements as in Experiment 1 although they could not have participated in Experiment 1. The sample after exclusions ($N = 981$) of participants was approximately half female (54%; .003% indicated other and .02% did not to respond) with a mean age of 37.73 years (SD = 12.77, Range = 18–80). The usable sample ($n = 862$) did not include participants who failed the attention check ($n = 119$). In Experiment 2, there were 286 own-words confidence identifications and 158 own-words confidence rejections.

In Experiment 3, new participants were recruited via (1) MTurk for pay, (2) a Scottish university's undergraduate psychology participant pool for course credit, (3) the Scottish Universities Pool, and (4) via social media platforms, such as Reddit, Facebook, and Twitter. Those recruited via social media volunteered their time. Participants recruited via MTurk were required to have at least a 75% HIT approval rate and had to have at least 100 HITs approved. Participants recruited via Mechanical Turk had to be from Australia, Canada, the United Kingdom, or the United States. From the total sample ($N = 2027$), we excluded survey previews ($n = 14$) and participants that met one of the exclusion criteria (failed attention checks, did not complete the study, took less than a minute or more than 30 min to complete, duplicate IP addresses, did not consent to participate, or did not provide identification data, $n = 813$). The resultant usable sample ($n = 1200$) included undergraduates who completed the study in exchange for course credit ($n = 134$), people recruited via social media platforms ($n = 64$), and MTurk workers ($n = 1002$). The usable sample comprised 41.08% males, 58% females, 0.42% other, and 0.50% who preferred not to respond. The sample ranged from 17 to 80 years of age ($M = 35.69$, SD = 12.62). In Experiment 3, there were 688 own-words confidence identifications and 512 own-words confidence rejections.

Each experiment had different research aims. Two of the three datasets have been previously published (Mansour, 2020; Mansour et al., 2020). In this short report, we analyze identifiers' ($n = 1099$) and rejectors' ($n = 781$) confidence statements from all three experiments who were asked to provide confidence "in their own words" to paint a picture of the variability in eyewitness' own words confidence judgments.

2.2 | Design

The experiments were all conducted online using Qualtrics (Provo, UT). In all three experiments, participants viewed a mock-crime video followed by a target-present or target-absent lineup. All own words judgments were provided by typing on a computer.

In Experiment 1 (Mansour, 2020), participants viewed a simultaneous lineup and provided confidence either numerically (0%–100%), or in their own words (or explained their decision, chose from a series of confidence phrases) and then provided a scale rating (0%–100%). The experiment was conducted online using Qualtrics.

In Experiment 2 (Mansour, 2020), participants viewed a sequential lineup and provided confidence in their own words followed by the scale rating of 0%–100% (or only via a scale rating). The experiment was conducted online using Qualtrics.

Experiment 3 (Mansour et al., 2020) used simultaneous lineups. A third of participants judged their confidence immediately in their own words and then on a scale of 0%–100%. The remaining participants judged their confidence in their own words and then on a scale after a two-minute delay during which they did a visual search task or generated reasons why their decision may have been incorrect.

Table 1 shows the proportion of lineup responses in the own words conditions across all three experiments.

2.3 | Measures and coding

2.3.1 | Quantitative uniqueness

We coded all confidence statements for quantitative uniqueness in their root phrases. Quantitative uniqueness was operationally defined as being the only one and unlike any other response. To do this, we first recoded all phrases to lower case before running a spellcheck. We removed punctuation (commas, periods, and apostrophes) and spaces in phrases (e.g., "not very confident" to "notveryconfident," "I'm not so sure" to "iamnotsosure"). We treated decimal numbers as proportions. We then created a table to identify unique responses. From the table, we removed all blank responses and duplicates. We conducted an additional formatting check. The first author then manually checked and removed any further duplicates from the table. The rationale for this approach was to quantify the extent to which people use different statements as an illustration of the flexibility of language use when people are asked to express confidence in their own words, which in turn underpins our calls for the development of a systematic approach to verbal confidence.

For the resultant phrase set, we coded unique root phrases for frequency. Accurate, confident, certain, and sure were considered synonyms (following Mansour, 2020) and as well as positive. We considered these stems synonymous according to how synonyms are defined (Murphy, 2013). We felt these stems could be interchanged in

"all contexts without altering the meaning of the text in which they occur" (p. 2, Schreyer, 1976). For example, "I am confident" versus "I am sure" versus "I am positive" versus "I am accurate" versus "I am certain." Figure S7 visually depict these stems and their synonyms according to WordNet. WordNet is a lexical database that includes synonyms and is commonly used in psycholinguistics (e.g., Murphy, 2003). WordNet indicates synonymity among positive, sure, confident, and certain. While accurate did not have connections to positive, sure, confident, and certain (see Figure S7), accurate was used interchangeably with these stems in our data (e.g., "I am quite accurate," "I feel that I am pretty accurate," and "I am about 75% accurate in my choice"). We provide in-depth discussion of the issue of synonymity in the Discussion. In cases where participants provided more than one confidence judgment, e.g., "I am fairly confident I would say 70%," we considered their first judgment, that is, fairly. When participants did not provide any indication of confidence (e.g., only explained their decision, "because I saw his face"), we coded the statement as a "justification only."

2.3.2 | Eyewitness numeric translations of own word judgments

We considered eyewitness' numeric translations of their own words' judgments (0%–100%) for frequently provided phrases.

2.3.3 | Qualitative uniqueness

Qualitative uniqueness was determined using ChatGPT (OpenAI, 2023). ChatGPT is a natural language processing tool (AI technology) that is freely accessible, useful in assessing contextual language, and appropriate for larger qualitative datasets. The rationale for this approach was to provide a qualitative assessment of language use when eyewitnesses are asked to express confidence.

ChatGPT defines qualitative uniqueness as distinctive and specific features or qualities that set a particular language apart from others (in the context of language or linguistics, ChatGPT, 2023). For our analyses, we used ChatGPT GPT-3.5 Model. The ChatGPT account did not have any custom instructions.

2.4 | Identifications

We prompted ChatGPT to categorize qualitatively unique statements into low, medium, and high confidence (November 28, 2023, Prompt: "Sort qualitatively unique phrases into low, medium, high from this set"). Due to our large data set, we separated each categorization analysis between studies (Exp. 1: https://chat.openai.com/share/68c22688-ff84-448c-8571-922ec7c95476, Exp. 2: https://chat.openai.com/share/355daeae-1b4f-463d-9811-a34398c97e15; Exp. 3 part 1: https://chat.openai.com/share/a9e8fdac-16b1-4039-a901-964b062b21e7, part 2: https://chat.openai.com/share/4c864b37-18a5-480f-a93f-e7a0620d9e72). Phrases identified in each category can be found in Table S1b.

To determine which confidence phrases were qualitatively unique for each confidence category across studies, we presented ChatGPT with all qualitatively unique phrases for each confidence category (low, medium, and high). We then asked ChatGPT to identify qualitatively unique phrases (November 28, 2023, Prompt: "Identify all qualitatively unique phrases from this phrase set") and to remove any duplicates (low: https://chat.openai.com/share/7751ca42-98e1-4659-b943-b00e658b2d5b, medium: https://chat.openai.com/share/cd22e72a-7de2-4433-853a-6c0dd41874e5, and high: https://chat.openai.com/share/ed915fc8-8d03-4af9-8312-55cabb5f2d49). Phrases identified in each category can be found in Table S1a.

2.5 | Rejections

For rejections, we again prompted ChatGPT to sort qualitatively unique phrases into low, medium, and high confidence (Prompt: "Sort qualitatively unique phrases into low, medium, and high") for each experiment (Experiment 1: (https://chat.openai.com/share/7742cc78-20a1-430f-ae84-6ce8fdcbf6e8, November 26, 2023), Experiment 2: (https://chat.openai.com/share/454a1ada-be36-413e-a465-fb193ab79ddc, November 28, 2023), and Experiment 3: (https://chat.openai.com/share/0c5f633c-9e79-4d6e-82e2-c4590117e5b0, November 26, 2023). Phrases identified in each category can be found in Table S2b. We again presented ChatGPT with all qualitatively unique phrases for each confidence category (low, medium, and high). We then prompted ChatGPT to identify qualitatively unique phrases (Prompt: "Identify all qualitatively unique phrases from the phrase set") and to remove any duplicates (low: https://chat.openai.com/share/8be4ac50-1de0-4f91-90bd-d529e9fa8b45, November 28, 2023, medium: https://chat.openai.com/share/86cd0159-2d4c-477a-aa4e-e38daf3fd9fe, November 28, 2023, and high: https://chat.openai.com/share/d5792a4e-767f-458e-b149-02abac58d9a1, November 28, 2023). Phrases identified in each category can be found in Table S2a.

2.5.1 | Categorization following Behrman and Richards (2005) coding scheme

We categorized the quantitatively coded root phrases based on their appearance in the coding scheme developed by Behrman and Richards (2005) (Table 1 in Behrman & Richards, 2005; for an explanation see Section 1).

3 | RESULTS

3.1 | Identifications

3.1.1 | Quantitative uniqueness

There were 1082 codable confidence expressions total. Out of those, we identified 781 quantitatively unique responses. We coded all root

phrases for how frequently they occurred (see Table 2). Even though people seem to prefer giving confidence verbally, there is great variation in their responses when asked to provide confidence in their own words. The most common phrase occurred just 7.94% of the time ("pretty sure"). Interestingly, the top four phrases (used approximately 7% of the time), cut across the range of low to high confidence: not very confident, fairly confident, pretty sure, and very confident.

3.1.2 | Modality of confidence statement

When considering the initial confidence expression provided, most identifiers (61.07%) provided a verbal judgment when asked for confidence in their own words. 34.19% used numbers, and 4.74% only justified their decision (e.g., "because I saw his face").

When considering the full statement, most identifiers (59.43%) provided verbal information only, 26.89% provided numeric information only, and 13.68% provided both, verbal and numeric information (e.g., I am 60% confident, but honestly, the more I think about it, the less confident I am).

3.1.3 | Categorization following Behrman and Richards (2005) coding scheme

A subset of the root phrases we coded for frequency included root phrases that appeared in Behrman and Richards (2005) coding scheme. We categorized these 31 root phrases (see Table 2) based on Behrman & Richards' scheme, the only available empirically-developed coding scheme for eyewitness confidence statements. We did not consider statements with root phrases associated with more than one category (e.g., "I think").

From our dataset, there were four root phrases categorized as low confidence (0–4), seven root phrases as medium confidence (5–7) and two as high confidence (8–10) (see Table 3). Our finding suggests that, when using Behrman and Richards' coding scheme, a majority of root phrases eyewitnesses use to express confidence (70.21%, 165 out of 235) were categorized as "medium confidence."

3.1.4 | Qualitative uniqueness for identifiers

ChatGPT reported 132 qualitatively unique phrases for identifiers (see Table S1a). ChatGPT notes that these interpretations are based on ChatGPT's interpretation of uniqueness in phrasing and that it is "challenging to precisely categorize them into low, medium, and high confidence levels as the expressions vary widely and are subjective." Unlike humans, artificial intelligence, such as a language processing model like ChatGPT, uses algorithms to categorize and interpret phrases. Importantly, ChatGPT may categorize phrases differently when presented with (a) the same phrase set, and (b) the same prompts because of its probabilistic algorithm. Consistent interpretation is difficult, likely because of the unique information provided in own words confidence statements (Seale-Carlisle et al., 2022).

3.1.5 | Eyewitnesses' numeric translation of own words confidence phrases

Mansour (2020, Table 4) suggests there is variation between eyewitness-participants in their translations of the same verbal confidence statements to a numeric scale. To provide a picture of the eyewitness' intended meaning of their own word confidence judgment, we examined eyewitness' numeric translations of frequently provided root phrases (10 or more occurrences; Phrases bolded in Table 1). In cases where participants provided more than one confidence judgment, we considered their first judgment. Some phrases contained both, a verbal judgment and a numeric judgment (e.g., "I think I was correct, so I'm feeling 90% confident"), in which cases we again considered their first confidence expression.

Out of the 1082 codable statements, there were 515 cases that used a frequently provided root phrase. All statements that were included for this analysis can be found in Table S3. Figure 1 shows the range of eyewitness' own numeric interpretations for frequently provided root phrases and the distribution of numeric interpretations for each root phrase. Figure 1 illustrates how variable people are in what they mean when they provide a root phrase, even if it is commonly used. Most phrases had peaks, but the distributions tended to extend across about half the scale, indicating considerable interwitness variability.

3.2 | Rejections

3.2.1 | Modality of confidence statement

When considering the initial confidence expression provided, most rejectors (67.11%) provided a verbal judgment when asked for confidence in their own words. 29.05% used numbers, and 3.84% only justified their decision (e.g., "I did not recognize any of the people in the lineup", "none of these," "cannot recall," and "this person").

When considering the full statement, most rejectors (65.72%) provided verbal information only, 23.58% provided numeric information only, and 10.70% provided both, verbal and numeric information (e.g., I am not confident. There was one person I wanted to say yes to, but I wasn't 100% sure).

3.2.2 | Quantitative uniqueness

We coded phrases for quantitative uniqueness by following the same method as for identifications. There were 776 interpretable confidence judgments. Out of those, we identified 599 quantitatively unique responses. Table 4 provides the frequency of occurrence of confidence expressions for each quantitatively unique root phrase. Similarly to identification decisions, there is great variability when confidence is provided in one's own words. The most common root phrases were fairly confident (9.85%), pretty sure (9.18%), very confident (5.84%), and confident (4.84%).

TABLE 2 Frequency of quantitatively unique confidence phrases for identifications.

Phrase	Frequency of occurrence	Case count
Numeric	34.19%	267
Pretty sure[a]	7.94%	62
Not very confident	7.43%	58
Fairly confident[a]	7.30%	57
Very confident[a]	5.00%	39
Justification only (e.g., because I saw his face)	4.74%	37
Confident[a]	3.46%	27
Not sure	2.94%	23
Somewhat confident	2.82%	22
Uninterpretable	2.05%	16
Looks like[a]	1.92%	15
Not 100%[a]	1.54%	12
I think[a]	1.53%	12
Quite confident	1.28%	10
Moderately confident[a]	1.28%	10
Not completely/totally	1.02%	8
A little	0.77%	6
Not real/really	0.77%	6
Reasonably	0.77%	6
Slightly	0.77%	6
Almost positive[a]	0.64%	5
Completely confident	0.64%	5
Remember	0.64%	5
Not at all	0.64%	5
Not that confident	0.51%	4
Relatively[a]	0.51%	4
Extremely	0.38%	3
Half sure	0.38%	3
Nearly	0.38%	3
Partially confident	0.38%	3
Unsure	0.38%	3
Not too confident	0.38%	3
Good	0.26%	2
Mostly	0.26%	2
Not as confident	0.26%	2
Not particularly	0.26%	2
Really confident	0.26%	2
Would not say I was sure/100%	0.26%	2
No confidence	0.26%	2
Not so confident	0.26%	2
A lot	0.13%	1
Familiar	0.13%	1
Resembles[a]	0.13%	1
High level	0.13%	1
Not extremely	0.13%	1
Strongly	0.13%	1

TABLE 2 (Continued)

Phrase	Frequency of occurrence	Case count
Very uncertain	0.13%	1
Real certainty	0.13%	1
Semi confident	0.13%	1
No doubt	0.13%	1
Not that confident	0.13%	1
Kind of	0.13%	1
Low confidence	0.13%	1
Mostly confident	0.13%	1
Not much	0.13%	1
Not super	0.13%	1
Rather confident	0.13%	1
Sort of	0.13%	1
Without a doubt	0.13%	1
More confident	0.13%	1

Note: This table reflects only participants who made an identification and provided confidence in their own words.
[a]Phrases reported as frequently used by real eyewitnesses in Behrman and Richards (2005, Table 1). Bolded statements indicate frequently provided phrases (10 or more occurrences).

3.2.3 | Qualitative uniqueness for rejections

ChatGPT indicates the challenges in interpreting eyewitness confidence are not limited to identification decisions. Across all three experiments, there were 143 qualitatively unique statements for rejectors (See Table S2a).

3.2.4 | Eyewitnesses' numeric translation of own words confidence phrases

Out of the 776 codable statements, there were 395 cases that provided a frequently used confidence phrase (see Table 4). Phrases that were included for this analysis can be found in Table S4. Figure 2 shows the range of eyewitness' own numeric interpretations for frequently provided phrases (i.e., 10 or more occurrences) and the distribution of numeric interpretations for each phrase. The lack of distinct peaks in the distributions indicates rejectors' numeric translations of their own verbal confidence statements are highly variable. Note that some phrases have multimodal distributions, which suggests that, not only do people vary in what they mean, there may be multiple different interpretations that are agreed by minorities of individuals.

4 | DISCUSSION

Our analysis paints a picture of what occurs when eyewitnesses are asked for confidence in their own words. It is probably unsurprising that there is considerable variability in the statements made by eyewitnesses queried about confidence in a way that mirrors typical police practice (i.e., "in your own words"). What is perhaps less obvious is that, when asked to provide numeric translations (i.e., intended meaning), eyewitnesses vary substantially for phrases that are broadly worded the same. This is especially concerning considering that, in practice, eyewitness confidence statements undergo multiple rounds of interpretation (e.g., police officer, judge, juror, and general public).

The variability in interpretation does not seem to be limited to certain phrases. Our data suggests that eyewitness' numeric interpretations of their own-word confidence phrases can span the entirety of the 0–100 scale for phrases like very confident (though a majority of people interpret very confident as 75%+; Range = 0%–100%), fairly confident (though a majority of participants interprets fairly as 50%+; Range = 1%–100%) and not confident (Range = 0%–91%) for identifications, and very confident (though a majority of people interprets very confident at 75%+; Range = 1%–100%), confident (Range = 1%–100%), and pretty confident (Range = 1%–100%) for rejections. This occurs despite the fact that the mean, mode, and median for each of these suggest specific categories (very confident, $M = 91.16$, Mode = 100, Mdn = 94; fairly confident, $M = 72.86$, Mode = 70, Mdn = 74.50; not confident, $M = 36.26$, Mode = 50, Mdn = 30 for identifications; and very confident, $M = 88.16$, Mode = 100, Mdn = 92, confident, $M = 85.08$, Mode = 100, Mdn = 90.5; pretty confident, $M = 78.08$, Mode = 80, Mdn = 80 for rejections).

Consistent with the conclusion that variability is ubiquitous, for identification decisions alone, ChatGPT identified 132 qualitatively unique phrases (i.e., phrases intending to convey different information). It is also worth noting that only 13 phrases (which occurred 235 times out of 781 cases, 30%) matched Behrman and Richards' (2005) coding scheme. Notably, Behrman and Richards selected their phrases because they were "distinct among the total of

Low	Case count	Medium	Case count	High	Case count
Kind of	1	Pretty sure	62	Confident	27
Sort of	1	Looks like	15	Very confident	39
Familiar	1	Almost positive	5		
Resembles	1	Not 100%	12		
		Fairly	57		
		Relatively	4		
		Moderately	10		
Total	4		165		66

TABLE 3 Identifiers' confidence phrases categorized following Behrman and Richards (2005) coding scheme.

Note: We treated confident, sure, certain, positive, accurate as interchangeable. Case counts represent occurrence of phrases in our dataset ($n = 235/781$).

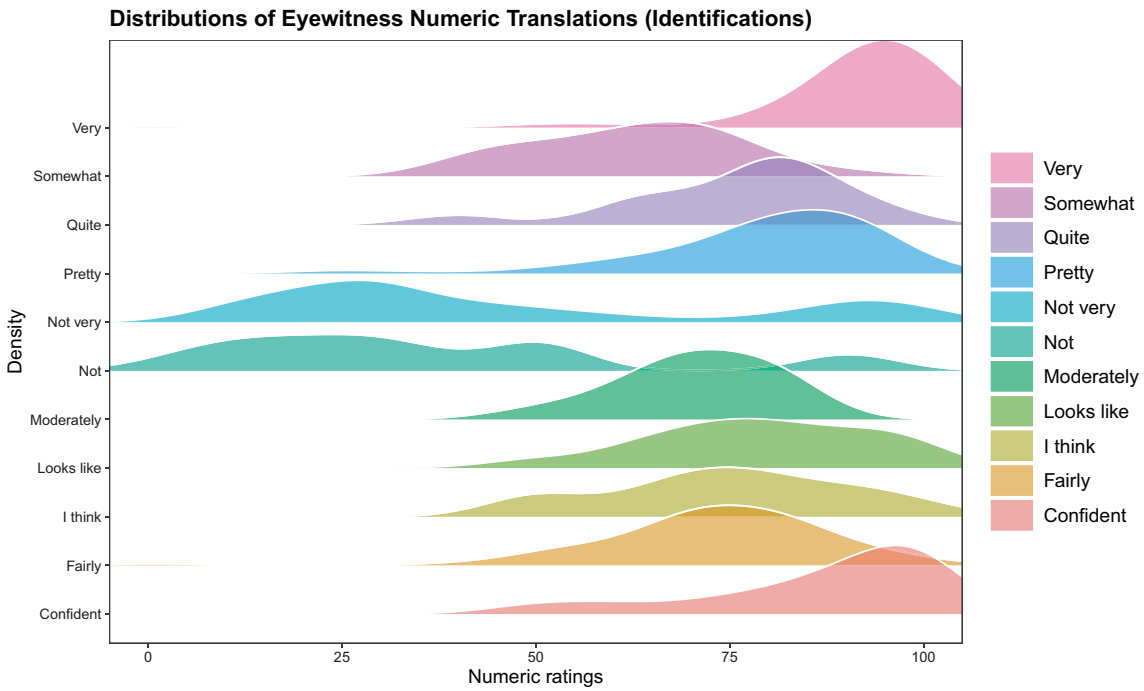

FIGURE 1 Ranges and distributions of identifiers' numeric translations of their own-words confidence judgments.

203 confidence statements" (p. 293). Coding schemes (such as the coding scheme developed by Behrman & Richards, 2005) are unlikely to capture the variety of responses exhibited when eyewitnesses are asked for confidence in their own words. More importantly, given the variability in confidence phrases provided by eyewitnesses, such coding schemes may not accurately capture the intended meaning eyewitnesses assign to commonly used root phrases. The considerable variety in verbal eyewitness confidence highlights the risk of inconsistent interpretation of people's judgments.

Indeed, interpretations of an eyewitness' intended meaning of a verbal confidence statement do not always align with interpretations at the receivers' end. For example, Mansour et al. (2020) asked others to interpret eyewitness' own word confidence statements as low, medium, or high. Approximately 15%–25% of the time, confidence was interpreted differently from how the eyewitness translated it—sometimes higher, sometimes lower (Table 4, Mansour, 2020). An eyewitness may mean 20% confident when she says "pretty sure," but may sometimes be interpreted as 60% confident, other times as 0%. Furthermore, recent work in our own lab (Mansour, 2020; Mansour & Vallano, 2022; Pennekamp & Mansour, 2024) as well as by Greenspan and Loftus (2023) demonstrate that interpretations vary substantially. Even when interpreting frequently provided confidence statements (e.g., completely certain), people vary in their interpretations (Greenspan & Loftus, 2023). Taken together, our and these findings strongly suggest that there will be considerable discrepancies in how eyewitness testimony is interpreted in practice. Given that identifications are likely to make it in front of the courts and triers of fact are asked to make judgments of guilt based on confidence statements, our findings suggest the criminal justice system would benefit from guidance about how to obtain and use verbal confidence statements.

TABLE 4 Frequency of quantitatively unique confidence phrases for rejections.

Phrase	Frequency of occurrence	Case count
Numeric	29.05%	174
Fairly confident[a]	9.85%	59
Pretty sure[a]	9.18%	55
Very confident[a]	5.84%	35
Confident[a]	4.84%	29
Not very confident	3.84%	23
Justification only (e.g., was not there, none of these)	3.84%	23
Somewhat confident	2.67%	16
Not at all	2.34%	14
Not sure	2.17%	13
I think[a]	1.84%	11
Quite confident	1.67%	10
Looks like[a]	1.67%	10
Remember	1.67%	10
Not too confident	1.34%	8
Moderately confident[a]	1.17%	7
Not completely/totally	1.17%	7
Uninterpretable	1.00%	6
Not that confident	1.00%	6
Mostly	0.83%	5
A little	0.67%	4
Almost positive[a]	0.67%	4
Not real/really	0.67%	4
Slightly	0.67%	4
Resembles[a]	0.50%	3
Kind of	0.50%	3
Not so confident	0.50%	3
Rather confident	0.50%	3
Yes	0.50%	3
Maybe	0.50%	3
Extremely	0.50%	3
Completely confident	0.33%	2
Really confident	0.33%	2
Relatively[a]	0.33%	2
Semi confident	0.33%	2
Not totally sure	0.33%	2
Confident but	0.33%	2
More sure	0.33%	2
Familiar	0.17%	1
Not 100%[a]	0.17%	1
Not as confident	0.17%	1
Not extremely	0.17%	1
Reasonably	0.17%	1
Unsure	0.17%	1
Real certainty	0.17%	1
Would not say I was sure/100%	0.17%	1

(Continues)

TABLE 4 (Continued)

Phrase	Frequency of occurrence	Case count
Not super	0.17%	1
Sort of	0.17%	1
Very much so	0.17%	1
Not confident enough	0.17%	1
So so	0.17%	1
Partly	0.17%	1
Guessing	0.17%	1
More likely	0.17%	1
Not quite	0.17%	1
Not overly	0.17%	1
Mild	0.17%	1
Medium	0.17%	1
Confident until	0.17%	1
Not incredibly confident	0.17%	1
Not crazy confident	0.17%	1
Highly	0.17%	1
Less confident	0.17%	1
Decently confident	0.17%	1
A bit	0.17%	1

Note: This table reflects only participants who rejected the lineup and provided confidence in their own words.
[a]Phrases reported as frequently used by real eyewitnesses in Behrman and Richards (2005, Table 1). Bolded statements indicate frequently provided phrases (10 or more occurrences).

To make matters worse, triers of fact are likely to receive phrases in context (sometimes containing more than one statement, such as "I can't be sure, but I think that is him, his hair looks similar and I think he kind of looks like my neighbor, not 100% confident though"). Context can affect interpretation of an eyewitness' confidence judgment (e.g., Cash & Lane, 2017) and triers of fact may not receive a confidence judgment in isolation. For example, phrase use and interpretations may differ depending on the type of crime (e.g., minor offense versus major offense) and depending on expansions (e.g., because I had a good look) or caveats expressed by the witness (e.g., but not 100%). The variability at the interpretative level may thus be even more pronounced in the real world.

Participants in our experiments provided their own words confidence judgments by typing on a computer. To analyze our data quantitatively, we only considered the initial confidence judgment to allow us to systematically code expressions. We analyzed our data qualitatively using ChatGPT to sort phrases by confidence level and qualitative uniqueness. Yet, even ChatGPT had difficulty providing consistent interpretations for some phrases (i.e., different interpretations were given when ChatGPT was prompted a second time). There are also differences in the interpretation of some phrases when comparing to other methods of interpretation (e.g., Behrman & Richards', 2005 coding scheme). For example, while phrases like "pretty sure" are frequently provided by eyewitnesses, they are not clearly interpretable. According to Behrman and Richards' coding scheme, "pretty sure" indicates medium confidence. However, ChatGPT, concluded that "pretty sure" indicates high confidence. The discrepancy in categorization for phrases may impact which identifications are interpreted as highly confident. Though imperfect, it is worth noting that ChatGPT (as other language models) is trained to follow a set of rules to interpret phrases. In practice, however, it is an individual trier of fact who makes the judgment—with no knowledge of how a verbal phrase is typically used or understood. Given the variability in the use of expressions, the variability in their intended meanings, and subjectivity in interpretations at the receiver's end, an eyewitness' verbal confidence judgment may not effectively inform the criminal justice system as a result. The interpretation of that judgment at different stages of the investigation has implications for actions the police take, whether a case is likely to be prosecuted, and the weighing of evidence in court.

A large proportion of participants spontaneously provided numeric judgments when asked for confidence in their own words (34.19% of identifiers, 29.05% of rejectors). There is research to suggest that numeric estimates provide more precision than verbal estimates (e.g., Dhami & Mandel, 2022). But even numeric estimates can be misinterpreted (Grabman & Dodson, 2019; Mansour & Vallano, 2022). The legal system seems reluctant to use numeric estimates (e.g., "quantitative scores might be misunderstood in the courtroom," American Law Institute, 2023, para. 10.06), suggesting that verbal confidence judgments are likely to remain the default in

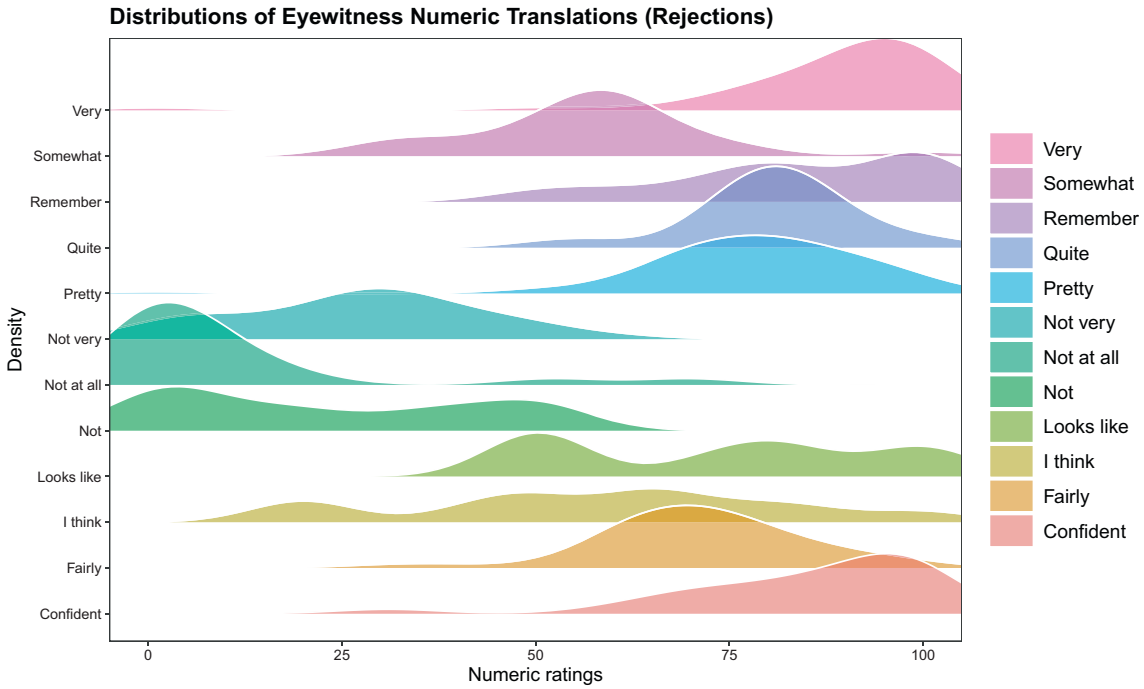

FIGURE 2 Ranges and distributions of eyewitness' numeric translations of own words judgments for rejections.

practice. Even though people (especially researchers) may be aware that verbal confidence statements are variably used and interpreted, we cannot expect the criminal justice system to change the way confidence is obtained and presented without empirical data showing how variable verbal eyewitness' confidence truly is. This article aimed to start that conversation.

While our data paints a picture of the variability in verbal eyewitness confidence, there are limitations to this work. First, in each experiment, we obtained limited demographic information. Data in all three studies were collected online, largely from the United States and the United Kingdom (specifically Scotland). Inferences regarding the generalizability of the variability in verbal eyewitness confidence are thus limited. Encouragingly, there is other work suggesting that there are discrepancies in how verbal probability estimates are understood (e.g., Smalarz et al., 2021 who found that evaluators underestimated verbal confidence in the eyewitness context; see also Brun & Teigen, 1988; Dhami & Mandel, 2022; Wintle et al., 2019).

Second, our work does not provide a conceptual categorization of verbal confidence phrases beyond categorizing phrases into low, medium and high (via ChatGPT). Given the variability in the content of confidence judgments and the variable meaning assigned to phrases, categorizing verbal confidence conceptually is no easy feat. While there are conceptualization methods (e.g., coding schemes) that have provided suggestions for *how* to categorize verbal confidence statements, there are few approaches that have been empirically developed to do so (e.g., Behrman & Richards, 2005; Mansour, 2020) and there are currently no approaches that consider the intended meaning of a phrase. This is important because, in practice, verbal eyewitness confidence is interpreted on an individual case basis. It is the eyewitness expressing confidence and the trier of fact (e.g., police officer, judge, and juror) making the interpretative judgment without any coding scheme or guidance for conceptual categorization. Given that confidence statements can provide useful information about the likely accuracy of eyewitness evidence when coded based on an empirical scheme or the wisdom of the crowd (e.g., Mansour, 2020; Smalarz et al., 2021), we need to provide an empirically-sourced roadmap for verbal confidence to be accurately obtained and interpreted.

Lastly, there is currently no empirical evidence for synonymity between commonly provided root confidence phrases in the context of eyewitness evidence. Does "sure" mean the same as "certain," for example? We treated accurate, confident, certain, and sure as synonyms (because previous research has done so; Mansour, 2020) and additionally included positive as interchangeable. While these phrases may share core meaning, there may be nuances that convey difference in meaning (e.g., near-synonyms). Similarly, other phrases may likely also be considered synonymous. The variability in eyewitness confidence could be even greater (or lesser) if boundaries of meaning are assigned differently. Given the qualitative differences between phrases, it is difficult to judge the degree of interchangeability without knowledge about semantic differences. Future research should assess synonymity between frequently provided eyewitness' confidence phrases to determine the extent to which they are synonymous.

Our data demonstrate eyewitness' use and interpretation of own word confidence varies. Own-word confidence statements are likely to be interpreted inconsistently because, as we demonstrate, of the variety of phrases used and the differences in intended meaning.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Pia Pennekamp https://orcid.org/0000-0001-5826-3020
Jamal K. Mansour https://orcid.org/0000-0001-7162-8493
Rhiannon J. Batstone https://orcid.org/0000-0003-1652-9649

REFERENCES

American Law Institute. (2023, October 17). Principles of law and policing. Retrieved from https://www.policingprinciples.org/chapter-10/10-06-obtaining-and-documenting-eyewitness-confidence-statements/

Arndorfer, A., & Charman, S. D. (2022). Assessing the effect of eyewitness identification confidence assessment method on the confidence-accuracy relationship. *Psychology, Public Policy, and Law*, *28*(3), 414–432. https://doi.org/10.1037/law0000348

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. https://doi.org/10.1007/s10979-005-3617-y

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* (pp. 501–525). Lawrence Erlbaum Associates Publishers.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404. https://doi.org/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. https://doi.org/10.1111/j.1467-9280.2009.02284.x

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. http://www.jstor.org/stable/24550328

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. https://doi.org/10.1002/bdm.440

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

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

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior*, *14*(2), 185-191. https://doi.org/10.1007/BF01062972

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, *26*(6), 514–526. https://doi.org/10.1016/j.tics.2022.03.002

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. https://doi.org/10.1037/xap0000182

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

Grabman, J. H., & Dodson, C. S. (2019). Prior knowledge influences interpretations of eyewitness confidence statements: 'The witness picked the suspect, they must be 100% sure%. *Psychology, Crime & Law*, *25*(1), 50–68. https://doi.org/10.1080/1068316X.2018.1497167

Greenspan, R. L., & Loftus, E. F. (2023). Interpreting eyewitness confidence: Numeric, verbal, and graded verbal scales. *Applied Cognitive Psychology*, *38*, e4151. https://doi.org/10.1002/acp.4151

Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied*, *10*(1), 25–41. https://doi.org/10.1037/1076-898x.10.1.25

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. https://doi.org/10.1002/acp.3870

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*, *29*, 120–141. https://doi.org/10.1080/1068316X.2021.2007912

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In *Social judgment and decision making* (pp. 227–242). Psychology Press.

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (2020). How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision*, *7*(3), 183–211. https://doi.org/10.1037/dec0000119

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. https://doi.org/10.1016/j.jarmac.2020.01.003

Mansour, J. K., Batstone, R. J., & Pennekamp, P. (2020). *The effect of non-social delays on the confidence-accuracy relationship* [Manuscript in preparation]. Division of Psychology, Sociology and Education, Queen Margaret University.

Mansour, J. K., & Vallano, J. P. (2022). *Does "very confident" mean very confident? Perceptions of low, medium, & high eyewitness confidence* [Conference presentation]. American psychology-law society conference, Denver, CO, United States. https://doi.org/10.31234/osf.io/ze427

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Murphy, M. L. (2013). What we talk about when we talk about synonyms: (and what it can tell us about thesauruses). *International Journal of Lexicography*, *26*(3), 279–304. https://doi.org/10.1093/ijl/ect023

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Pennekamp, P., & Mansour, J. K. (2024). Laypeople's Interpretations of "High Confidence". *Psychology, Crime and Law*. https://doi.org/10.1080/1068316X.2024.2329707

Police Bureau Portland. (2023, July 22). Eyewitness identification. Retrieved from https://www.portlandoregon.gov/police/article/780992

Schreyer, R. (1976). *Synonyms in context*. Linguistic Agency University of Duisburg.

Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2022). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*, *151*(6), 1283–1305. https://doi.org/10.1037/xge0001152

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. https://doi.org/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. https://doi.org/10.1037/lhb0000444

Steblay, N. K., & Wells, G. L. (2023). In their own words: Verbalizations of real eyewitnesses during identification lineups.

Psychology, Public Policy, and Law, *29*(3), 272–287. https://doi.org/10.1037/law0000386

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday & Co.

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. https://doi.org/10.1371/journal.pone.0213522

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. https://doi.org/10.1177/1529100616686966

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pennekamp, P., Mansour, J. K., & Batstone, R. J. (2024). Variability in verbal eyewitness confidence. *Applied Cognitive Psychology*, *38*(2), e4190. https://doi.org/10.1002/acp.4190