International Journal of Language & Law

Kendro, JLL 15 (2025): 86–110
www.languageandlaw.eu

Lexical Diversity Predicts Juror Perceptions of Written Eyewitness Statements

*Kelly Kendro**

Abstract

While many studies investigate how eyewitness identities such as gender, race, and dialect might affect evaluations of the eyewitness's credibility, lexical dimensions of these statements are often overlooked. Properties of language production such as lexical diversity can reveal information about a person's language status or other identities, serving as indirect proxies for those attributes when explicit information is unavailable. Previous research has found that a higher number of total (*volume*) and unique (*abundance*) words, lower type-token ratio (*variety-repetition*), and higher proportions of "vivid" descriptors are linked to veracity (e.g., Colwell et al., 2007; Morgan et al., 2013, 2015). However, it is unclear how more complex dimensions of lexical diversity might be related to perceived credibility. In this study, mock jurors ($n = 64$) evaluated written eyewitness statements for perceived accuracy, credibility, deceptiveness, eloquence, and prestige. The statements, evaluated across six dimensions of lexical diversity (*volume*, *abundance*, *variety-repetition*, *evenness*, *disparity*, and *dispersion*), were written by L1 and L2 English speakers. Quantitative analyses indicated that *volume*, *abundance*, *variety-repetition*, *evenness*, and *disparity* predicted perceived integrity (*accuracy*, *credibility*, and *deceptiveness*), while *abundance*, *variety-repetition*, *evenness*, and *disparity* predicted perceived status (*eloquence* and *prestige*). When controlling for lexical diversity, L1 statements were nevertheless rated more positively than L2 statements. These results highlight how juror perceptions may be biased even absent explicit demographic information. Furthermore, they identify the lexical aspects of a text on which jurors may rely when evaluating the integrity of a witness, with implications for the study of human error in veracity judgments.

Keywords

eyewitness testimony, credibility, lexical diversity, linguistic bias

Submitted: 1 October 2025, accepted: 27 March 2026, published online: 4 May 2026

**Kelly Kendro*: Northern Arizona University, kelly.kendro@nau.edu. The author would like to thank Scott Jarvis, two anonymous reviewers, and attendees of the 2024 AAAL conference for their feedback on earlier versions of this work.

DOI: 10.14762/jll.2026.86

1. Introduction

Eyewitness testimony is often introduced as evidence in court trials. In countries with jury trials, a panel of jurors is expected to evaluate this testimony and use that evaluation to ultimately inform their decision regarding a defendant's guilt or liability. Many previous studies have investigated the accuracy of eyewitness accounts (e.g., Buckhout, 1974; Wells & Turtle, 1987; Kassin et al., 1989; Loftus, 1996, 2019; O'Neill Shermer et al., 2011), and others have probed how jurors' perceptions about the eyewitness's race, accent, and ethnic origin (e.g., Frumkin, 2007; Lev-Ari & Keysar, 2010; Frumkin & Stone, 2020; Frumkin & Thompson, 2020) might be reflected in their evaluation of testimony.

Human raters largely fail to correctly gauge whether a given statement is true or false (Bond et al., 1985), leading to questions about what might help characterize truthfulness in a given text. While researchers have studied the composition of witness statements themselves in relation to the veracity of those statements (e.g., Morgan et al., 2013), dimensions of the text that comprise its lexical diversity, such as the length of the text, the number of unique words, and the relationships between those words, have not yet been considered in the context of juror perceptions. Lexical diversity is known to be linked to demographic variables such as whether a person is using their first language (L1) or second language (L2); jurors may therefore unconsciously use clues from the lexical diversity of a statement to make decisions about the eyewitness's identities and their credibility.

The current study asked participants eligible to serve on juries in the United States to assess written eyewitness statements for their credibility and related attributes. The statements were assessed along six dimensions of lexical diversity (Jarvis, 2013b), which is known to be affected by language proficiency and status, to determine whether any are predictive of the participants' perceptions of the statements. Additionally, this study attempts to evaluate "vividness" (Bell & Loftus, 1985, 1988; Colwell et al., 2007; Reyes et al., 1980) in the statements to consider whether this lexical aspect is related to perceptions of credibility. Finally, the study examines whether the mock jurors rate statements written by L2 English speakers more harshly than L1 English speakers, even when that language status is not explicitly revealed.

2. Background

2.1. Language in the Courtroom

Though the Sixth Amendment of the U.S. Constitution guarantees the right to a trial by an impartial jury, human jurors have human biases. Linguistic bias is particularly pervasive, and linguists have examined how these biases affect the judicial process from the beginning of a police interrogation (e.g., Pavlenko et al., 2019) to the ultimate verdict

(e.g., Rickford & King, 2016). People using their L2 in the courtroom are uniquely vulnerable participants; as Powell (2008) notes, courtrooms are often expected to be a monolingual setting that does not reflect a richly multilingual society. For those who are not dominant in or comfortable speaking the language in which court is conducted, the courtroom context can lead to miscommunication and lack of comprehension, archaic legalisms notwithstanding. Leung (2008) maintains that legal interpreting is not only a necessary step towards equality in the courtroom, but a fundamental linguistic right that all people must be afforded. However, even when support personnel such as translators or interpreters are not necessary, the way an individual uses language may still lead to bias or discrimination in the courtroom.

Sociolinguistic aspects of language production such as accent and dialect can have a tremendous impact on courtroom outcomes, even if the person testifying is doing so in their L1. Minoritized varieties of a language, especially those tied to another identity such as African American English (AAE), can affect how language is understood and perceived in the courtroom. Jones et al. (2019) found that when court reporters encountered AAE testimony, their accuracy declined compared to mainstream US English, demonstrating particular difficulty in correctly transcribing sentences and phrases. These results call into question the validity of subsequent appeals and related legal proceedings that are unknowingly using inaccurate transcriptions. In the context of juror perceptions, Rickford and King (2016) detail how linguistic bias against Rachel Jeantal, a Black witness testifying at the 2013 trial in which a man was tried for killing her friend Trayvon Martin, ultimately led jurors to acquit the defendant. Jeantel speaks AAE and used this variety when responding to questions during her testimony. Interviews with jurors after the trial's conclusion made it clear that Jeantal's use of AAE led some jurors to dismiss her testimony, with one admitting that she discounted Jeantel as "not credible" due to *how* she spoke rather than her testimony itself (Rickford & King, 2016: 971). Beyond the United States, Drobnjak (2024) reports varying perceptions of eyewitness credibility in Slovenian courts using audio recordings of eyewitness statements in Slovene. Drobnjak notes that Slovenia has low levels of income inequality such that socioeconomic status is largely naturally controlled. The stimuli featured statements read by L1 speakers of regional varieties of Slovenian, and participants were asked to give binary responses to prompts about the witnesses' credibility, disturbingness, eloquence, naturalness, pleasantness, respectability, trustworthiness, and understandability, among other attributes. Dialect did indeed significantly predict the participants' ratings, underscoring how jurors can be biased against intrinsic, immutable characteristics of a witness.

Studies have also investigated the relationship between perceptions of witness identities and how jurors rate their credibility. In a study employing video stimuli, Frumkin (2007) investigated perceptions of spoken eyewitness testimony across five Likert scale metrics, namely accuracy, credibility, deceptiveness, physical attractiveness, and prestige, as well as a binary determination of the defendant's guilt. Participants were mock jurors who evaluated taped eyewitness statements from purportedly real court cases.

There were three "eyewitnesses", one each of German, Mexican, and Lebanese background. The mock witnesses were highly proficient bilinguals who recorded two versions of the videos, keeping the script constant. One video was "accent-free" with American English affectation, and one featured an accent that more traditionally corresponded to their national origin. The researcher found a significant main effect of perceived ethnic background, with the Lebanese witness rated as less credible than the other witnesses. In addition, accent predicted ratings of each of the five credibility metrics. In a follow-up study, Frumkin and Stone (2020) asked participants to rate "eyewitnesses" from audio clips rather than the videos that Frumkin (2007) used. The paradigm was further modified to test interactions between three demographic variables: accent status, race, and age. These mock witnesses were either white or Black, older (aged 50 to 55) or younger (aged 20 to 25), and had accents typically associated with higher or lower social status in Britain. Participants rated the eyewitnesses on eight dimensions, including accuracy and credibility, which were collapsed into a single "favorability" score following factor analysis. Higher status accent was associated with higher favorability than lower status, and white witnesses were rated more favorably than Black witnesses.

In sum, evidence linking dialect and language status in credibility ratings indicates that the way a person uses language in the courtroom may be just as important as what they say. Juror decisions and perceptions may be affected by unconscious biases that lead to disparities in courtroom outcomes, negatively impacting members of certain groups. The present study investigates an overlooked dimension of eyewitness statements that may play a role in juror perceptions of (and biases regarding) credibility: lexical characteristics. In the courtroom, lexical aspects of spoken witness statements occur alongside other linguistic dimensions such as fluency, morphosyntactic accuracy, and pragmatic cues that cannot be fully captured by a written transcript. These dimensions can also be difficult to separate when investigating the role each plays in how jurors perceive a witness, leading to potential confounds in research. Written statements represent a related, yet distinct, register of legal language (see also Chen, 2021, on subregisters such as court transcripts). While these statements are not directly analogous to spoken testimony, they represent a useful way to isolate lexical characteristics from additional communicative variables, offering valuable insights into how language production is evaluated in legal settings (e.g., Brönnimann et al., 2013; Builes et al., 2024).

2.2. Lexical Diversity

Lexical diversity encompasses several aspects of quantifying the lexical variety in a text. Jarvis (2013a) draws connections between patterns of ecological diversity and naturally occurring lexical diversity, noting that diversity arises not solely from variety or differences, but instead from the *combination* of varied elements and the relationships between

those elements. In an overview, Jarvis (2013b) identifies seven dimensions of lexical diversity that have been used in prior research: *volume*, *abundance*, *variability* (later renamed "variety-repetition"; Akbary & Jarvis, 2023), *evenness*, *disparity*, *dispersion*, and *specialness*. *Volume* refers to the number of words, or tokens, present within a given text, while *abundance* refers to the number of unique tokens. *Variety-repetition* describes the relationship between *volume* and *abundance* and is often adapted from type-token ratio (TTR). *Evenness*, a concept borrowed from studies of ecological diversity, quantifies the balance of type distribution throughout a text. *Disparity* calculates the semantic similarity between words. *Dispersion* analyzes the distance between types as they occur throughout a text. *Specialness* refers to specific words that make a text "feel" more diverse, an aspect of diversity that remains elusive in terms of quantification (Jarvis, 2017). Each of these metrics might alone fulfill a simplistic definition of "lexical diversity", but combining them allows for a multifaceted and nuanced idea of diversity within a given text. When untrained human raters are asked to evaluate the lexical diversity in a text, their judgments show high interrater reliability and pattern with quantitative measurements of lexical diversity dimensions at overall correlations of between .492 and .847 (Jarvis, 2017; Kyle et al., 2021). Put plainly, human raters can perceive and identify diversity in a text to an extent that is highly correlated with objective measures.

There are some limits to quantifying lexical diversity. One of these limits is the length of a text: shorter samples (<100 words) should not be used for the more complex lexical diversity measures (i.e., *variety-repetition*, *evenness*, *disparity*, and *dispersion*), as they are dependent on text length (Kyle et al., 2021; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021). For example, some measures of *variety-repetition*, such as Measure of Textual Lexical Diversity, use a technique whereby words are evaluated consecutively, leading to issues with the last several words in the text (McCarthy & Jarvis, 2010; Vidal & Jarvis, 2020). An alternate *variety-repetition* measure, Moving Average Type-Token Ratio (MATTR; Covington & McFall, 2010), rapidly computes the type-token ratio for a window of variable length (which is adjusted for the length of each text) to avoid sampling biases; MATTR is considered a contemporary "gold standard" in assessing *variety-repetition* in the field of lexical diversity.

Lexical diversity is known to vary across populations. Durán et al. (2004) report trends in lexical diversity for both first and second language development. They found that L1 speech was characterized by sequentially increasing lexical diversity for typically developing children aged 18- to 60-months. In contrast, lexical diversity in speech produced by children with specific language impairment was shown to develop non-linearly and demonstrated wide variation compared to the other child group. For adult L2 learners with low proficiency, there were register differences in lexical diversity (e.g., academic text demonstrated the highest scores), but their level of overall lexical diversity was comparable to the typically developing 5-year-olds' L1 speech. These results indicate that lexical diversity might be used to infer information about a person's language development or status. In adults, higher L2 proficiency generally correlates with higher

lexical diversity in both spoken and written language production. Treffers-Daller (2013) investigated the usefulness of two measures of *variety-repetition* for assessing language knowledge; they ultimately explained 62% of the variance in participants' L2 proficiency scores. Read and Nation (2006) found that higher production of types and tokens during L2 speech assessment was linked to higher language proficiency scores. Yoon (2017) identified patterns in lexical diversity that characterized both L2 proficiency level and register. Though the corpus analysis showed some overlap in *variety-repetition* across adjacent proficiency levels, there were clear distinctions between higher and lower proficiency.

Data from attrition populations, composed of individuals whose language knowledge has declined, are somewhat consistent with these results. Schmid and Jarvis (2014) tested participants who were L2 dominant following L1 attrition, comparing them to a group of monolinguals. They report that lexical diversity scores were better predictors of group membership than measures of current proficiency. While *evenness* and *disparity* were most characteristic of the attrition group, there was also wide variation between group members with regards to lexical diversity. Gharibi and Boers (2019) echo these results with their study of child heritage language users aged 6 to 18 whose language knowledge had attrited. They found that *variety-repetition* was lower compared to monolingual controls, though they produced a larger range compared to the monolingual group. These data also showed an effect of age, with older children producing language with higher lexical diversity, indicating that the age effects observed by Durán et al. (2004) persist beyond age five. Taken together, these results indicate that lexical diversity is often associated with specific groups of language users.

As humans are sensitive to differences in lexical diversity, these dimensions might serve as cues for a writer or speaker's identities even without any additional context. Even absent explicit information or assumptions about a person's background, such as (perceived) accent or dialect, lexical diversity may shape how a juror evaluates the credibility of an individual. The current study seeks to determine the extent to which six dimensions of lexical diversity may be related to juror ratings of written eyewitness statements.

2.3. Forensic Lexical Variation

In addition to studies of language development and proficiency, aspects of lexical diversity have been investigated in relation to veracity in a forensic context. Morgan et al. (2013) employed modified cognitive interviewing (MCI) in an effort to improve detection of *veracity* and *deception*. The MCI technique involves prompting interviewees to recall a given event across visual, auditory, emotional, and temporal dimensions. For instance, rather than the traditional "Tell me what happened", an interviewer using MCI might ask questions such as "What did that look like?" or "If I were there, what might I have seen?" to elicit specific visual details,

with similarly structured questions pertaining to the other dimensions. In the study, researchers collected MCI data involving either true or false recall, then asked highly skilled and experienced human raters to determine whether the witness was being deceptive. In addition, the interviewers were directed to judge the veracity of the eyewitness statement immediately following the interaction, and the *volume*, *abundance*, and *variety-repetition* (TTR) of the statements were computed. The results reveal the fallibility of human raters: despite these participants' background in the security sector, their classification accuracy was 52% for truthful statements and 41.7% for untruthful statements – performing worse than chance. The lexical diversity analysis, however, was fruitful. While none of the three dimensions predicted veracity in responses to visual or emotional prompts, number of types and tokens *did* predict veracity in statements involving auditory and temporal recall, in addition to predicting full recall (collapsed across the four dimensions). All three dimensions predicted veracity in the initial non-MCI unprompted recall. When the researchers found differences in lexical diversity, TTR was lower and number of types and tokens were higher for truthful statements. The model with just types and tokens as predictors of veracity had 84.4% classification accuracy, with 27% sensitivity and 98% specificity; in other words, the model was highly accurate at determining truthfulness but had more false "positives" when identifying someone as deceptive.

Morgan et al. (2015) also used lexical diversity to assess variation in MCI statements from three groups of people: those who had completed a task and described it, those who had not completed a task but said they had, and those who had completed a task but said they had not. The tasks were either cognitive or manual in nature. Three human raters trained in MCI evaluated the statements for veracity, with an average accuracy of 76% for truthful accounts and 46.7% for untruthful accounts. When there were significant differences between true and false statements, numbers of types and tokens were higher in the truthful statements compared to the untruthful statements, much like Morgan et al.'s (2013) findings. In the 2015 study, TTR was higher for all truthful statement types except a prompt where participants were asked to reflect on any possible errors in their recall. These results demonstrate that measurements of lexical diversity can be a helpful tool in understanding how language varies in the forensic context.

Few publications investigate length or word count of typical eyewitness accounts. The Morgan et al. studies report wide variability in mean statement lengths, with responses to individual MCI prompts as between 49.5 and 729.7 words (2013) and full interview statement lengths (when responding to interlocutor prompts) as between 590.1 (SD = 132) words for truthful statements and 1288.5 (SD = 132) words for deceptive statements. However, as these statements were collected in a lab setting with prompts intended to elicit extended contextual information, they may not accurately reflect naturalistic testimony length. In an analysis of statements from real court cases, Builes et al. (2024) note veracity-related differences in transcribed spoken testimony, with fewer words per sentence in dishonest testimony. However, the authors did not analyze overall length of statement because they were not the same across honest and dishonest testimony, with no mention of directional trends. Brönnimann et al. (2013) also analyzed transcripts from real court testimony; however, their analysis controlled

for statement length, with no exact values or ranges reported. Thus, while number of tokens may be a helpful indicator of veracity, few studies have quantified this lexical diversity dimension in real-life statements.

Beyond lexical diversity, social psychologists have another, related construct that is directly related to how witness testimony is perceived. Reyes et al. (1980) found that variation in accounts describing the same event led to drastically different outcomes in a mock juror paradigm. The authors presented participants with "vivid" detailed eyewitness accounts and "pallid" accounts with few details. Their mock jurors were able to remember more about the vivid accounts in both immediate and delayed recall, and they also found the vivid statement to be more persuasive: the juror was more likely to side with the party whose witness presented a vivid account, regardless of whether they were the defendant or plaintiff. Bell and Loftus (1985) proposed that these results reflect how detail-rich vivid descriptions can evoke stronger emotional responses than pallid descriptions, which may in turn lead to positive perceptions about the witness's knowledge or memory abilities. In a subsequent study, the authors at least partially confirmed this hypothesis by collecting and examining qualitative responses from mock jurors that detailed *why* they found statements to be more or less persuasive (Bell & Loftus, 1988). Decades later, Colwell et al. (2007) collected eyewitness accounts that were either truthful, obfuscated (such that the perpetrator was difficult to identify), or fabricated. The researchers operationalized statement *vividness* as both total number of details and proportion of extraneous details to total details. They found that *vividness* strongly predicted perceived and objective veracity and guilt; that is, that higher *vividness* was associated with both veracity and greater *perceived* veracity. There is, however, a limit to the effect of *vividness* in perceptions of eyewitness accounts: Peace et al. (2015) found that *vividness* must co-occur with plausibility, as their participants rated statements with "bizarre" details as less credible despite high levels of *vividness*. While these studies do not situate *vividness* within the context of lexical diversity, there may be some relationship between quantifications of *vividness* (parts of speech) and lexical diversity (lexemes).

Altogether, previous work offers compelling evidence that some aspects of lexical diversity are linked to veracity in a forensic context and that juror perceptions of witness credibility can be influenced by the way a witness sounds or looks. However, it remains unknown whether lexical diversity also plays a role in perceptions of witness credibility, which dimensions of lexical diversity matter most, and whether any effects of lexical diversity can be separated from broader biases related to language status. Additionally, findings on lexical characteristics of witness testimony from social psychology have not been directly integrated with lexical variation research. The current study brings together these separate lines of inquiry to investigate how variation in lexical diversity relates to perceptions of credibility and associated metrics. It expands the application of forensic lexical diversity analysis to include other, more sophisticated measures of lexical diversity and contributes a more comprehensive understanding of the role of lexical diversity in how written eyewitness statements are perceived.

DOI: 10.14762/jll.2026.86

2.4. The Current Study

This research extends the work of Morgan et al. (2013, 2015), as well as the social psychology literature on *vividness* and perceived credibility (e.g., Bell & Loftus, 1988; Colwell et al., 2007; Reyes et al., 1980) and research on the relationship between juror perceptions of eyewitness identities and credibility (e.g., Frumkin, 2007; Frumkin & Stone, 2020), to uncover whether the lexical diversity of a written eyewitness statement predicts juror perceptions about that witness. It is the first to expand the operationalization of lexical diversity in the forensic context to include three additional, theoretically distinct dimensions of lexical diversity (*evenness*, *disparity*, and *dispersion*; Jarvis, 2013b), and it further distinguishes between perceptions of credibility and other factors (such as eloquence) that may influence juror judgment. Specifically, this study asks the following research questions:

(1) Does lexical diversity in a written eyewitness statement predict perceived accuracy, credibility, deceptiveness, eloquence, and prestige?
(2) Does *vividness* (quantified as number of words adding non-essential context) predict perceived credibility, accuracy, deceptiveness, eloquence, and prestige?
(3) When controlling for lexical diversity, do mock jurors rate statements from L1 and L2 English writers differently?

For (1), following previous research findings, it was predicted that perceptions of credibility would be positively associated with *variety-repetition* and number of tokens and types. In an exploratory analysis, *disparity*, *dispersion*, and *evenness* were evaluated as possible predictors of perceptions of credibility and related metrics. For (2), it was predicted that higher *vividness* would be associated with higher perceived accuracy and credibility. For (3), it was predicted that L2 statements would be rated lower in prestige and eloquence.

3. Method

3.1. Stimuli

The stimuli consisted of eight short statements. Four of the statements were written by L1 English speakers, and four were written by L2 English speakers. Each statement was written by a different undergraduate student and described a video of a separate event, such as a car accident or a theft. Statement A was written by a female L1 English speaker aged 19 who is fluent in heritage language Jamaican Patois. Statement B was written by a male monolingual English speaker aged 19. Statement C was written by a male L1 English speaker aged 18 who has limited knowledge of L2 Spanish. Statement D was written by a female monolingual English speaker aged 36 who has limited knowledge of L2 Spanish, Italian, and Korean. Statement E was written by a female L1 Chinese speaker aged 21 with self-described conversational fluency in L2 English. Statement F was written by a female L1 Spanish speaker aged 20 who is fluent in L2 English and has limited knowledge of L2 Portuguese and Turkish. Statement G

was written by a male L1 Cantonese speaker aged 21 who is fluent in L2 English. Statement H was written by a male L1 Chinese speaker aged 22 who has limited knowledge of L2 English.

The writers were asked to recall what they had seen in as much detail as possible, writing at least 100 words (range[1]: 100 to 179), the minimum length necessary for analysis of a text's lexical diversity (McCarthy & Jarvis, 2010; Zenker & Kyle, 2021). The statements had originally been written in the third-person (e.g., "There was a pedestrian crossing the street in front of *the* car"), and they were converted to first-person accounts (e.g., "There was a pedestrian crossing the street in front of *my* car") for the purposes of the study per the findings of Colwell et al. (2007) regarding obfuscation in eyewitness statements. The text samples were then scored using six measures of lexical diversity proposed by Jarvis (2013): *abundance* (types), *volume* (tokens), *variety-repetition* (MATTR),[2] *evenness*, *disparity*, and *dispersion*. Table 1 reports lexical diversity metrics for each of the eight statements.[3]

Table 1: *Lexical Diversity and Vividness Values for Stimuli Statements*

Statement	Types	Tokens	HD-D	MATTR	MTLD-W	Evenness	Disparity	Dispersion	Vividness
A	46	100	26.1328	28.63	23.37	0.9236	1.002	18	0.3
B	43	102	26.1328	31.11	26.24	0.9364	1.004	9.8	0.235
C	76	100	36.5055	42.18	106	0.9834	1.008	5	0.313
D	69	148	28.4157	30.18	21.79	0.9184	1.012	15.54	0.257
E	73	100	36.2209	42.96	102	0.9857	1.025	4	0.293
F	67	179	28.309	31.28	31.22	0.9391	1.023	16.2	0.32
G	60	117	28.5455	31.72	29.17	0.9308	1.005	9.4	0.265
H	52	115	27.1543	31.02	25.61	0.9321	1.016	21.74	0.252

Note: Statements A-D were written by an L1 English speaker, and Statements E-H were written by an L2 English speaker. All participants were presented with a total of 4 written statements and told that they were real eyewitness accounts from a court case. Two of these four were written by L1 English speakers, and two were written by L2 English speakers. The four statements were counterbalanced and randomized within those lists to avoid order effects. Each statement and its related questions appeared on a single, separate screen. All text appeared in black 16-pt Times New Roman font on a white background.

[1] As limited data is available regarding typical length of eyewitness accounts, especially in real testimony, a comparison between these statements' lengths and those of typical eyewitness statements would be useful but is not possible at this time.

[2] In addition to MATTR, hypergeometric distribution of diversity (HD-D; McCarthy & Jarvis 2007) and measure of textual lexical diversity-moving window technique (MTLD-W; Vidal & Jarvis 2020) scores were calculated to evaluate the *variety-repetition* dimension of lexical diversity (see Table 1). These indices were chosen due to their reliability and stability regardless of text length (Kyle et al., 2021). The correlations between these 3 measures (MATTR, HD-D, MTLD-W) were very high (r = .977 to .990, all ps < .001), and models with HD-D or MTLD-W instead of MATTR did not change the overall results.

[3] The statements are also available in an OSF repository. Available at osf.io/udcyk/?view_only=853421513a8f49649be5463c4ca0ae8f (accessed 11. April 2026).

3.2. Participants

Participants were recruited via Prolific.com, an online subject pool that has been shown to provide high-quality data from human subjects (Peer et al., 2022). To mitigate low-quality or automated responses, participants had to have previously completed at least 100 Prolific surveys with an approval rate of 98% or more to be eligible for this study. Responses submitted by participants were also screened manually to ensure that demographic information in the survey matched the demographics in their Prolific participant profile, that attention checks were successful, and that the completion time was not significantly shorter than other participants' submissions.

Prior to viewing any surveys, Prolific requires all potential participants to complete a lengthy questionnaire, with items ranging from education level to political affiliation to specific medical conditions. Researchers on the platform can then use filters to show a study to a custom group of Prolific users who meet some set of criteria. For the current study, the following parameters were set: participants' nationality was United States; participants responded "yes" or "no" (but not "I don't know" or "decline to respond") to whether they had served on a jury, had been the victim of a crime, or had been imprisoned for committing a crime; and participants were comfortable taking part in a study where deception was used. The recruitment message stated that participants would answer demographic questions, then read texts and answer questions about the texts' authors.

Following an *a priori* power analysis to calculate necessary sample size, 64 participants were recruited for the study. All participants successfully passed two attention checks, and therefore none were excluded. Participants were aged 19 to 79 ($M = 37.03$, $SD = 13.66$). There were 23 male, 39 female, and 2 nonbinary participants. Two participants were born outside of the United States (in the Philippines and Trinidad), but all 64 were United States citizens. Six participants had previously served on US juries, and an additional 31 had been summoned for jury duty but not selected. Twenty-two participants indicated that they or someone they knew had been a defendant in a criminal trial, and 18 reported knowing someone who had been sentenced to prison.

3.3. Procedure

Data were collected via Qualtrics survey during asynchronous online sessions. Participants had no time limit for how long they could stay on each page, but they were not allowed to return to a screen once they had progressed to the next item. Prior to the start of the study, all participants read an informed consent document approved by the appropriate Institutional Review Board and gave their consent to participate. After the consenting procedure, the study's questions and structure largely followed that of Frumkin (2007). Participants first completed a demographic questionnaire covering

age, gender, education, country of origin, language, and ethnicity. Participants then responded to the items detailed above regarding jury duty, experience in criminal cases, and criminal convictions. After completing the demographic questionnaire, participants were told that the survey required them to act as a mock juror and read four eyewitness statements from "real crimes", then answer questions about the writers of the statements. At this point, they were given the choice to either reaffirm their willingness to participate or decline and end the study. Next, each participant was "sworn in" and given "Judge's instructions", as they would in a real courtroom.

Finally, the participants were presented with the statement-rating task. After reading each statement, participants rated its writer from 0 (not at all) to 10 (very) for how accurate, credible, deceptive, eloquent, and prestigious they thought the "witness" was. The dimensions of accuracy, credibility, deceptiveness, and prestige were included following Frumkin (2007), and they were presented using the same wording. Unlike the current text-based study, Frumkin was collecting data based on observing and listening to, rather than reading, eyewitness statements. Physical attractiveness was therefore omitted as a dimension, instead replaced by eloquence. Table 2 reports the by-statement credibility ratings.

Table 2: *By-Statement Ratings*

Statement	Accuracy	Credibility	Deceptiveness	Eloquence	Prestige
A	6.188(1.89)	6.344(1.75)	1.469(1.68)	4.719(2.54)	4.156(1.71)
B	7.906(1.42)	7.656(1.66)	0.750(1.27)	5.812(1.99)	5.156(2.00)
C	4.344(2.04)	4.844(2.34)	2.594(2.46)	5.344(2.57)	3.938(2.29)
D	7.594(1.52)	7.312(1.73)	1.000(0.98)	6.500(1.92)	5.094(2.02)
E	3.812(1.80)	4.125(2.04)	1.906(2.13)	1.094(0.86)	2.000(1.78)
F	7.031(1.91)	6.562(2.09)	1.750(1.95)	4.500(2.06)	4.562(1.70)
G	6.031(1.67)	6.219(1.81)	1.375(1.18)	3.906(2.05)	3.656(1.99)
H	5.250(1.67)	5.688(1.89)	1.562(1.74)	2.156(1.65)	2.469(1.78)

Note: M(SD). All scores range from 0 (lowest) to 10 (highest).

At the end of the study, participants read a debriefing form explaining that the study had employed deception in that the statements were not, in fact, from real court cases. They were then asked whether they wanted to allow their data to be included for analysis or to be deleted; all participants agreed to their inclusion.

3.4. Analysis

All quantitative data were analyzed using R (R Core Team, 2025). Prior to analysis, a correlation matrix of the five perception metrics was computed, using |.3| as the correlation threshold (Tabachnick & Fidell, 2007). As there were several correlations above the threshold, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy was calculated to ensure that Exploratory Factor Analysis would be appropriate for identifying underlying factors for the five metrics; the KMO threshold (Kaiser, 1974) was .5 and used principal axis factoring. The KMO statistic was .76, between "middling" and "meritorious" on Kaiser's scale, which indicated that the data would be suitable for factor analysis. The scree plot in Figure 1 shows a sharp decline between the eigenvalues for factors 1 and 2, with the eigenvalue for factor 2 at just above 1.0. The plot levels off after factor 2, indicating that 2 underlying factors are most representative of these data. Factor 1, henceforth *INTEGRITY*, was composed of perceived accuracy, credibility, and (inverted) deceptiveness, which had respective loadings of .71, .74, and .66. Perceived eloquence and prestige comprised Factor 2, henceforth *STATUS*, with respective loadings of .79 and .86. Together, these factors account for .65 of the cumulative variance. The two factors were used as dependent variables for the subsequent analyses. The mean INTEGRITY and STATUS ratings for each statement are presented in Table 3. The dataset is available in the linked OSF repository.

Figure 1: *Scree Plot for Factors Underlying the Five Perceptual Dimensions.*

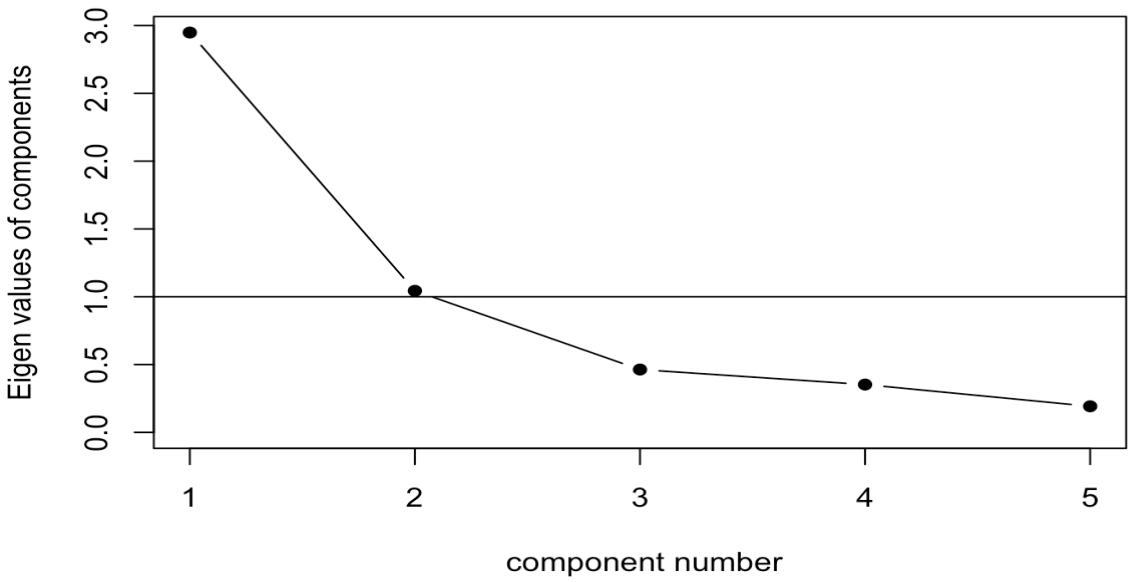


Table 3: *By-Passage Ratings for* INTEGRITY *and* STATUS

Statement	INTEGRITY	STATUS
A	7.021(1.44)	4.438(1.90)
B	8.271(1.29)	7.656(1.70)
C	5.531(2.04)	5.484(2.00)
D	7.969(1.27)	5.797(1.74)
E	5.344(1.46)	1.547(1.08)
F	7.281(1.65)	4.531(1.64)
G	6.958(1.34)	3.781(1.82)
H	6.458(1.28)	2.313(1.62)

Note: M(SD). All scores range from 0 (lowest) to 10 (highest).

Linear regression models were employed to answer (1). Each lexical diversity metric was a predictor, and INTEGRITY and STATUS were outcome variables; there were thus 12 models total, one for each of the lexical diversity metrics per dependent variable. This method was selected due to high multicollinearity between the measurements of lexical diversity, e.g., a higher number of types necessarily requires a higher number of tokens. A Holm-Bonferroni correction was applied to mitigate the possibility of Type I errors resulting from this method.

To answer (2), each of the statements was tagged for part of speech using TreeTagger (Schmid, 1997), and the following parts of speech were subsequently included in the calculation of *vividness*: prepositions, subordinate conjunctions, adjectives, predeterminers, possessive pronouns, and adverbs. The sum of the vivid parts of speech was divided by the total number of words in the text to calculate a *vividness* proportion. Linear regression models were then employed to determine whether *vividness* proportion significantly predicted INTEGRITY or STATUS.

(3) investigated whether participants rated L1 and L2 statements differently; while lexical diversity provides broad insight about lexical patterns in language production, it cannot capture all variation within texts. The statements themselves had a range of values for the lexical diversity dimensions, and lexical diversity was controlled across L1 and L2 statements to the extent possible. Per two-sample *t*-tests, there were no significant differences between the groups for any of the 6 lexical diversity dimensions (all *p*s > .05). This decision therefore allowed for delineating the effect of lexical diversity versus the effect of L1/L2 STATUS. To answer (3), a two-sample *t*-test was performed to determine whether participants rated statements written by L1 and L2 English speakers differently for INTEGRITY and STATUS.

An exploratory analysis was also performed to examine the extent to which *vividness* would be correlated with each of the lexical diversity dimensions. The operationalization of *vividness* in this study was based on parts of speech, while the lexical diversity operationalizations were each based on individual lexemes. *Vividness* was therefore not anticipated to be directly analogous to any of the lexical diversity measures, but it was thought that the latter's dimensions of *volume*, *abundance*, *variety-repetition*, and *disparity* might show some overlap with the vividness score.

4. Results

4.1. Lexical Diversity

(1) asked whether lexical diversity would be associated with participants' perceptions of the statements. In the Bonferroni-corrected linear regression analyses ($\alpha = .0083$), five of the six lexical diversity dimensions predicted integrity: *volume*, *abundance*, *variety-repetition*, *evenness*, and *disparity* (see Table 4). *Dispersion* was marginally significant prior to the correction for multiple analyses (original $p = .028$) but failed to reach significance at the adjusted alpha level. In the linear regression analyses for STATUS, three of the lexical diversity predictors reached significance after the Bonferroni correction ($\alpha = .0083$): *variety-repetition*, *evenness*, and *disparity* (see Table 5). *Abundance* (original $p = .118$; corrected $p = .707$), *volume* (original $p = .002$; corrected $p = .011$), and *dispersion* (original $p = .056$; corrected $p = .339$) were not significant predictors of STATUS.

Table 4: *Regression Analysis for Lexical Diversity and* INTEGRITY

Variable	Estimate	SE	95% Confidence Interval		Adjusted *p*
			LL	UL	
Volume	.016	.005	.006	.025	.011
Abundance	-.018	.012	-.041	.005	.707
Variety-repetition	-.118	.025	-.168	-.069	< .001
Evenness	25.79	5.363	-36.347	-15.224	< .001
Disparity	1.642	.423	0.808	2.475	< .001
Dispersion	.016	.008	-.000	.032	.339

Table 5: *Regression Analysis for Lexical Diversity and* STATUS

Variable	Estimate	SE	95% Confidence Interval		Adjusted *p*
			LL	UL	
Volume	.015	.004	-.064	-.029	< .001
Abundance	-.046	.009	.001	.023	< .001
Variety-repetition	-.150	.018	-.191	-.117	< .001
Evenness	-32.90	4.043	-40.857	-24.932	< .001
Disparity	2.12	.327	1.477	2.765	< .001
Dispersion	.015	.007	.002	.028	.170

These results indicate that higher perceived INTEGRITY is associated with greater *volume* (i.e., more words overall), lower *abundance* (i.e., fewer unique words), lower *variety-repetition* (i.e., fewer unique words compared to total words), lower *evenness* (i.e., uneven distribution of unique words throughout the text), and higher *disparity* (i.e., more synonyms). For integrity, *variety-repetition* and *evenness* were the strongest predictors, each explaining about 20% of the variation in INTEGRITY ratings. The results also reveal that higher perceived STATUS is associated with lower *variety-repetition*, higher *evenness* (i.e., more even distribution of unique words throughout the text), and higher *disparity*. For STATUS, there were no predictors that explained a large portion of the variance, with the highest amount of variance (7.7%) explained by *variety-repetition*.

4.2. Vividness

(2) asked whether the proportion of vivid words in a statement predicted its perceived INTEGRITY and STATUS. A linear regression model indicated that *vividness* indeed was significantly associated with INTEGRITY ratings, $\beta = -17.97$, $SE = 3.633$, $t(254) = -4.95$, $p < .001$; $R^2 = .084$. However, there was no significant relationship between STATUS and *vividness*, $\beta = -5.65$, $SE = 4.681$, $t(254) = -1.21$, $p = .228$; $R^2 = .002$. These results indicate that statements with more vivid detail were perceived as being less deceptive and more credible and accurate, explaining about 8% of variance in these ratings. Conversely, the *vividness* of a description had no significant bearing on whether the writer was perceived as more eloquent or prestigious.

4.3. Language STATUS

(3) investigated whether participants perceived L1 and L2 writers to have differing levels of INTEGRITY and STATUS beyond the lexical diversity in the statements. Results from a two-tailed *t*-test indicated that INTEGRITY was significantly different between the groups, $t(248.5) = -3.17$, $p = .002$. The L1 statements were rated as higher ($M = 7.20$) than the L2 statements ($M = 6.51$). The effect size for this difference was small, Hedges' $g = -.39$, 95% CI [-0.64, -.15].

A two-tailed *t*-test between the L1 and L2 groups for STATUS also showed a significant difference, $t(253.82) = -8.49$, $p < .001$. Again, the L1 statements were rated significantly higher ($M = 5.09$) than the L2 statements ($M = 3.04$). This difference had a large effect size, Hedges' g = -1.06, 95% CI [-1.32, -.80], representing more than one standard deviation from the mean. Figures 2 and 3 show the differences in means for the L1 and L2 groups via boxplots.

Figure 2: *INTEGRITY Ratings by L1/L2 STATUS*

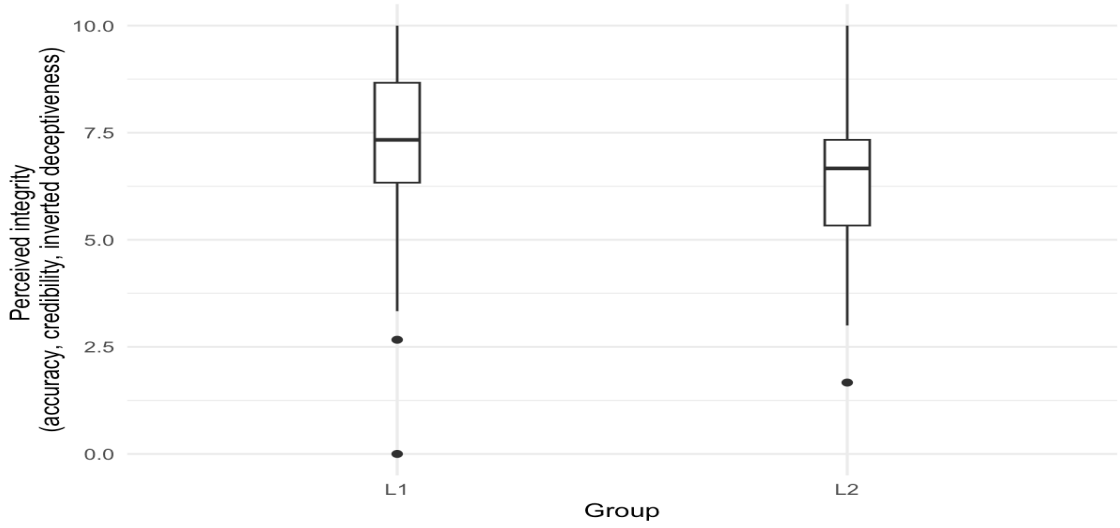
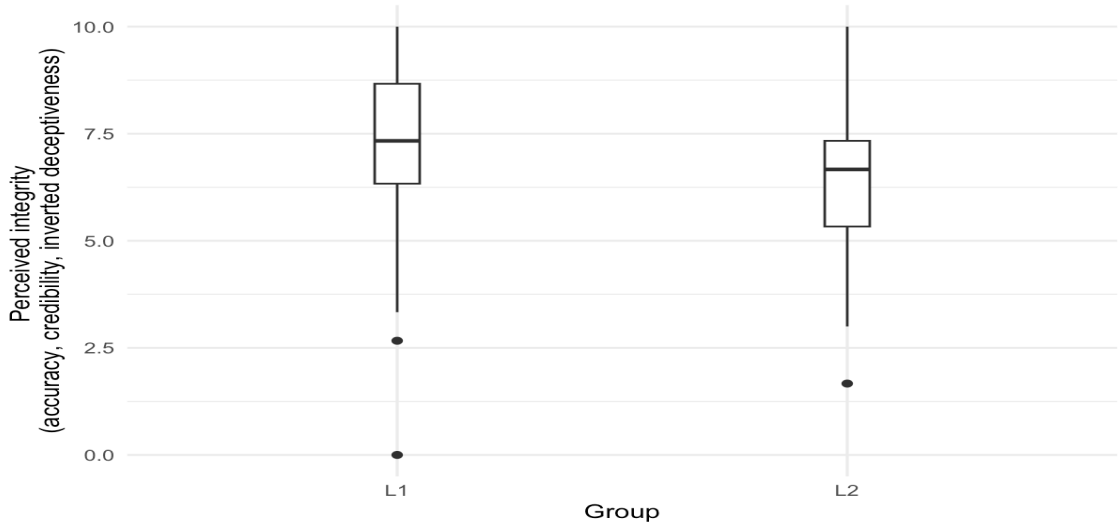


Figure 3: *STATUS Ratings by L1/L2 STATUS*



4.4. Relationship Between Vividness and Lexical Diversity

The final analysis was an exploratory comparison of *vividness* and the six lexical diversity dimensions to determine the degree of overlap. To accomplish this, correlations were computed between *vividness* and each of the diversity measures. While *vividness* was significantly correlated with all six dimensions, it was most highly correlated with Types ($r(254) = .54$, $p < .001$), *evenness* ($r(254) = .47$, $p < .001$), *variety-repetition* ($r(254) = .41$, $p < .001$), *dispersion* ($r(254) = .36$, $p < .001$), *tokens* ($r(254) = .22$, $p < .001$), and *disparity* ($r(254) = -.13$, $p = .030$). These results demonstrate that, while *vividness* has some overlap with lexical diversity, it represents a separate aspect of lexical variety.

5. Discussion

This study shows that some lexical features of an eyewitness statement are associated with juror perceptions of credibility and related attributes even when information such as appearance and accent is unavailable. Additionally, statements made by L2 English writers were generally rated less favorably than those made by L1 English writers, even when lexical diversity was controlled. However, deceptiveness (the perceptual dimension most analogous to veracity) was associated more strongly with perceived credibility and accuracy rather than perceived prestige and eloquence, suggesting that, to some extent, jurors distinguish between stylistic elements and perceived INTEGRITY in their evaluations.

5.1. Lexical Diversity

(1) asked which, if any, lexical diversity dimensions predicted perceptions of INTEGRITY and STATUS. After controlling for multiple analyses, the linear regressions indicated that five of the six lexical diversity dimensions (*volume*, *abundance*, *variety-repetition*, *evenness*, and *disparity*) significantly predicted INTEGRITY (accuracy, credibility, and deceptiveness). INTEGRITY was positively associated with 1) longer statements that had 2) fewer unique words, 3) a less even distribution of the same word across the statement, and 4) more semantically related words (i.e., synonyms). The linear regressions for STATUS (eloquence and prestige) found fewer predictors among the lexical diversity dimensions, namely *variety-repetition*, *evenness*, and *disparity*. STATUS was positively associated with 1) a lower ratio of unique to total words, 2) a more even distribution of the same word across the statement, and 3) a higher number of semantically related words.

Overall, the lexical diversity results reveal that mock jurors do indeed consider lexical diversity on some level when evaluating witness statements. Certain aspects of lexical diversity, such as statement length, are seen as indicating credibility, while others, such as the number of unique words, appear to (perhaps counterintuitively) seem suspicious. At the same time, related words were received positively and associated with both higher credibility and STATUS. The *evenness* of the distribution of repeated words across the statement had opposite effects on perceptions of INTEGRITY and STATUS, positively associated with STATUS and negatively associated with INTEGRITY, though this accounted for much more variation in INTEGRITY ratings (20.8%) versus STATUS ratings (8%). Similarly, *variety-repetition* played a large role in INTEGRITY (20.6%) compared to STATUS (7.7%) ratings, though both were positive correlations.

The findings related to credibility are somewhat consistent with prior literature, albeit with some important caveats. Most crucially, these results may point to a key difference between the lexical characteristics associated with perceived INTEGRITY and ob-

jective veracity: lower *abundance* may lead jurors to *perceive* a true statement as inaccurate. The participants in the Morgan et al. (2013, 2015) studies performed near chance (52%) when identifying untrue statements but higher when identifying true statements (72%), and the current study's lexical diversity findings may therefore help explain generally poor human judgments about whether a statement is true (e.g., Bond et al., 1985). Future work may investigate how other linguistic characteristics of eyewitness statements, such as vocabulary choice, syntax, and grammar, affect perceived INTEGRITY and STATUS.

Finally, these results have broader implications regarding linguistic variation and bias. The lexical diversity of language production has been shown to often vary systematically due to factors such as language proficiency or STATUS (which will be discussed further in 5.3). The results from this study show that jurors may unintentionally penalize certain types of lexical features, even when those features have no bearing on the truth of the statement. These findings are therefore relevant not only to (forensic) linguists, but also to legal professionals and members of law enforcement tasked with eliciting and evaluating witness statements. Through an understanding of how lexical diversity can bias perceived credibility ratings, including how these perceptions differ from objective measures of veracity, perhaps these effects could be mitigated, ultimately leading to more just forensic outcomes.

Beyond lexical qualities and witness characteristics such as pronunciation and appearance, juror evaluation of courtroom discourse also relies on cues including attorney-witness interactions, presuppositions, co-constructed event descriptions, sequential organization of responses, question and answer structure, and response completeness and appropriateness when constructing perceptions of credibility, STATUS, and accuracy (e.g., Ehrlich, 2011). The stimuli used in this study were static, decontextualized statements that do not reflect the full breadth of information available during synchronous testimony. Instead, these results identify lexical diversity as one of many co-occurring variables that contribute to how credibility is built and negotiated through naturalistic courtroom interactions. Future research may consider how lexical diversity interacts with pragmatic attributes of courtroom dialogue to gain a more holistic understanding of how jurors use this information to construct perceptions of integrity.

5.2. Vividness

(2) asked whether *vividness* significantly predicted perceptions of INTEGRITY and STATUS. A linear regression model indicated that, while *vividness* significantly predicted perceived INTEGRITY (accuracy, credibility, deceptiveness), it did not predict STATUS (prestige and eloquence). While *vividness* did significantly predict perceived integrity, the small effect size indicates that the amount of descriptive detail is only one of many characteristics used to draw these conclusions.

These results are consistent with the findings of Colwell et al. (2007), who had found higher *vividness* to be associated with objective veracity in witness statements. In that study, *vividness* was operationalized as the number of details and proportion of details within the text. The current study's findings reflect that jurors consider more detailed statements to be more accurate and credible and less deceptive. Together, *vividness* appears to both indicate objective veracity as well as lead to increased perceived integrity. However, the current study extends this work to include perceived STATUS, with results showing no relationship between *vividness* and STATUS. Taken together, the *vividness* data suggest that jurors are relying on multiple lexical attributes (e.g., not simply number of [unique] words, but *which* words appear) as they evaluate witness statements, and they do not evenly consider these attributes in the process.

Additionally, the current findings underscore that *vividness* is not interchangeable with lexical diversity. An exploratory analysis investigated how *vividness* is related to the six dimensions of lexical diversity used in this study. To date, no other known publications have attempted to link the concept of *vividness* as introduced by social psychologists to the quantifiable aspects of lexical diversity long used by linguists. Correlations revealed that *vividness* was significantly correlated with each of the six lexical diversity measures, ranging from moderate correlations (types, .54; *evenness*, .47; *variety-repetition*, .41; *dispersion*, .36) to low correlations (tokens, .22; *disparity*, -.13). These results, while generally expected, are somewhat surprising in light of the perceived STATUS data. Recall that *variety-repetition*, *evenness*, and *disparity* significantly predicted STATUS, and *vividness* did not, despite moderate correlations with *variety-repetition* and *evenness*. These results suggest that the predictive power of *variety-repetition* and *evenness* instead lie in aspects of lexical diversity that are independent from *vividness*. Overall, this links *vividness* to measurable lexical variation; future work in this domain from the perspective of social psychology may thus benefit from quantitative, in addition to impressionistic or qualitative, evaluations of *vividness*. Future research may also compare this operationalization, applied to unstructured elicited statements, to the way Colwell et al. (2007) operationalized *vividness* for their more controlled stimuli statements.

5.3. Language STATUS

(3) asked whether, beyond lexical diversity, participants would rate L1 and L2 statements differently. Each participant saw two L1 and L2 statements, and the statements were not labeled for language STATUS. The results showed significant differences for both INTEGRITY and STATUS: L1 statements were rated higher for both despite the "blinded" nature of the evaluation. Notably, the effect size for the difference in STATUS was large, indicating that these ratings may reflect perceptions about the *writer* rather than the *statement*; however, the statements were still negatively affected.

In this study, jurors penalized L2 writers even though language STATUS was not disclosed and lexical diversity between the L1 and L2 statements was matched. These results demonstrate that biases in courtroom interactions can be present in written testimony and are not restricted to the auditory or visual cues such as accent, appearance, and pronunciation investigated by Frumkin (2007) and Frumkin and Stone (2020). Instead, these perceptions reflect broader biases connected to the way language is used, demonstrating the vulnerable position of L2 speakers in the courtroom (Powell, 2008) even when they do not require language support. The results also echo Rickford and King (2016) in that jurors may dismiss testimony based on language STATUS or identity rather than its content. These biases can negatively impact individuals' courtroom outcomes, leading to systematic inequities.

It is clear that the courtroom is not an unbiased setting. How can biases be mitigated? Juror training may be a good place to start. The concept of explicitly addressing implicit biases is not novel, and indeed recent work has attempted, albeit unsuccessfully, to combat unconscious juror bias via pre-trial video (Kirshenbaum & Miller, 2021; Ruva et al., 2024). Perhaps targeted information about ways that implicit bias may lead to unfair and disproportionate negative courtroom outcomes through characteristics such as language may be more successful. More research must be carried out in this area to increase the efficacy of related interventions and efforts. For now, the current work contributes to the literature on unconscious linguistic bias related to L2 STATUS in the courtroom, particularly as it relates to the evaluation of witness statements.

5.4. Limitations

The primary limitation of the current study was that the mock trial setting may not directly reflect a real courtroom. The mock jurors had time to reread statements that might usually be read aloud, and they were asked to evaluate the statements without the additional context afforded by trial activities such as attorney questions and other witness testimony. This method meant that participants were not influenced by the perceptions of fellow jurors or factors such as eyewitness accent or appearance, and the statement evaluations reflected only the textual content; however, subsequent studies may introduce additional variables to investigate whether they lead to different results. Additionally, the writers of the witness statements were all undergraduate students, and future work might explore whether education level affects perceived INTEGRITY and STATUS. Finally, all the descriptions in these written statements were true, an intentional decision to control for objective veracity. However, the inclusion of untrue statements may reveal additional differences or nuances in lexical diversity, especially the dimensions of *evenness*, *disparity*, and *dispersion* that have not previously been used to evaluate witness statements. While these methodological constraints may limit the direct transferability of the results, they offer controlled evidence for the role that lexical characteristics play in perceptions of credibility and STATUS separate from potential confounds related to a spoken modality.

6. Conclusion

The current study makes four major contributions to current understanding of the linguistic factors that shape perceived eyewitness credibility. First, it demonstrates that lexical diversity can shape juror perceptions of witnesses even when visual and auditory cues about witness identity are unavailable. Second, it distinguishes between perceived INTEGRITY and perceived STATUS as separate evaluative constructs influenced by different dimensions of lexical diversity. Third, it reveals persistent biases against L2 writers even when controlling for a statement's lexical diversity. Finally, it identifies a specific lexical diversity dimension that differs between perceived credibility and objective veracity: while higher *abundance* (i.e., number of types) was associated with less credibility in this study, prior research indicates it is a marker of higher veracity. Altogether, these findings help clarify which evaluations are about the witness and which about their statement, providing insight into the factors that may influence juror evaluations of courtroom testimony.

There is much room for future studies, especially in the realm of investigating how other linguistic dimensions such as syntactic variation, complexity, or lexical richness might influence juror perceptions. These dimensions, alongside lexical diversity analyses, could also be applied to other aspects of the judicial process such as dialogue from interrogations or accuracy of court reporter transcriptions. Furthermore, future research might investigate juror perceptions of regional or dialectal variation to understand how these elements can influence perceived credibility and related aspects. There are clearly many factors at play in how jurors perceive witnesses and their statements, and identifying which specific factors lead to increased bias can help forensic linguists and other interested parties work towards linguistic equity in the courtroom.

References

Akbary, Mary & Jarvis, Scott (2023). Lexical diversity as a predictor of genre in TV shows. *Digital Scholarship in the Humanities*, 38(3), 921–936. DOI: 10.1093/llc/fqad004.

Bell, Brad E. & Loftus, Elizabeth F. (1985). Vivid persuasion in the courtroom. *Journal of Personality Assessment*, 49(6), 659–664. DOI: 10.1207/s15327752jpa4906_16.

Bell, Brad E. & Loftus, Elizabeth F. (1988). Degree of detail of eyewitness testimony and mock juror judgments. *Journal of Applied Social Psychology*, 18(14), 1171–1192. DOI: 10.1111/j.1559-1816.1988.tb01200.x.

Bond, Charles F.; Kahler, Karen Nelson & Paolicelli, Lucia M. (1985). The miscommunication of deception: An adaptive perspective. *Journal of Experimental Social Psychology*, 21(4), 331–345. DOI: 10.1016/0022-1031(85)90034-4.

Brönnimann, Rebecca; Herlihy, Jane; Müller, Julia & Ehlert, Ulrike (2013). Do testimonies of traumatic events differ depending on the interviewer? *The European Journal of Psychology Applied to Legal Context*, 5(1), 97–121.

Buckhout, Robert (1974). Eyewitness testimony. *Scientific American*, 231(6), 23–31.

Builes, Juan Camilo Carvajal; Barreto, Idaly & Gutiérrez De Piñeres, Carolina (2024). Deception detection based on the linguistic style of honest and dishonest stories. *The Journal of Forensic Practice, 26(1)*, 46–59. DOI: 10.1108/JFP-07-2023-0035.

Chen, Meishan (2021). Is courtroom discourse an 'oral' or 'literate' register? The importance of sub-register. *Discourse Studies*, 23(3), 1–25. DOI: 10.1177/1461445620982097.

Colwell, Kevin; Hiscock-Anisman, Cheryl; Memon, Amina; Rachel, Alexis & Colwell, Lori (2007). Vividness and spontaneity of statement detail characteristics as predictors of witness credibility. *American Journal of Forensic Psychology*, 25(1), 5–30.

Covington, Michael A. & McFall, Joe D. (2010). Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100. DOI: 10.1080/09296171003643098.

Drobnjak, Marko (2024). *Vloga Percepcije Govora Pri Oceni Verodostojnosti Pričanja. [The Role of Speech Perception in Assessing the Credibility of Testimony]* [Thesis, Univerza v Ljubljani]. Available at repozitorij.uni-lj.si/IzpisGradiva.php?id=159142 (accessed 12 April 2026).

Durán, Pilar; Malvern, David; Richards, Brian & Chipere, Ngoni (2004). Developmental trends in lexical diversity. *Applied Linguistics*, 25(2), 220–242. DOI: 10.1093/applin/25.2.220.

Ehrlich, Susan (2011). Courtroom Discourse. In Wodak, Johnstone & Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 361–372). SAGE Publications Ltd. DOI: 10.4135/9781446200957.

Frumkin, Lara A. (2007). Influences of accent and ethnic background on perceptions of eyewitness testimony. *Psychology, Crime & Law*, 13(3), 317–331. DOI: 10.1080/10683160600822246.

Frumkin, Lara A. & Stone, Anna (2020). Not all eyewitnesses are equal: Accent status, race and age interact to influence evaluations of testimony. *Journal of Ethnicity in Criminal Justice*, 18(2), 123–145. DOI: 10.1080/15377938.2020.1727806.

Frumkin, Lara A. & Thompson, Amanda (2020). The impact of different British accents on perceptions of eyewitness statements. *Journal of Language and Discrimination*, 4(1).

Gharibi, Khadijeh & Boers, Frank (2019). Influential factors in lexical richness of young heritage speakers' family language: Iranians in New Zealand. *International Journal of Bilingualism*, 23(2), 381–399. DOI: 10.1177/1367006917728395.

Jarvis, Scott (2013a). Capturing the Diversity in Lexical Diversity. *Language Learning*, 63(s1), 87–106. DOI: 10.1111/j.1467-9922.2012.00739.x.

Jarvis, Scott (2013b). Defining and measuring lexical diversity. In Jarvis & Daller (Eds.), *Vocabulary Knowledge: Human Ratings and Automated Measures* (pp. 13–41). John Benjamins Publishing Company.

Jarvis, Scott (2017). Grounding lexical diversity in human judgments. *Language Testing*, 34(4), 537–553. DOI: 10.1177/0265532217710632.

Jones, Taylor; Kalbfeld, Jessica Rose; Hancock, Ryan & Clark, Robin (2019). Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2), e216–e252.

Kaiser, Henry F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. DOI: 10.1007/BF02291575.

Kassin, Saul M.; Ellsworth, Phoebe C. & Smith, Vicki L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist*, 44(8), 1089–1098. DOI: 10.1037/0003-066X.44.8.1089.

Kirshenbaum, Jacqueline M. & Miller, Monica K. (2021). Judges' experiences with mitigating jurors' implicit biases. *Psychiatry, Psychology and Law*, 28(5), 683–693. DOI: 10.1080/13218719.2020.1837029.

Kyle, Kristopher; Crossley, Scott A. & Jarvis, Scott (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2), 154–170. DOI: 10.1080/15434303.2020.1844205.

Leung, Esther S. (2008). Interpreting for the minority, interpreting for the power. In Gibbons & Turell (Eds.), *Dimensions of Forensic Linguistics* (pp. 197–211). John Benjamins.

Lev-Ari, Shiri & Keysar, Boaz (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093–1096. DOI: 10.1016/j.jesp.2010.05.025.

Loftus, Elizabeth F. (1996). *Eyewitness Testimony*. Harvard: University Press.

Loftus, Elizabeth F. (2019). Eyewitness testimony. *Applied Cognitive Psychology*, 33(4), 498–503. DOI: 10.1002/acp.3542.

McCarthy, Philip M. & Jarvis, Scott (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. DOI: 10.3758/BRM.42.2.381.

Morgan, Charles A. III; Rabinowitz, Yarin; Hilts, Deborah; Weller, Craig E. & Coric, Vladmir (2013). Efficacy of modified cognitive interviewing, compared to human judgments in detecting deception related to bio-threat activities. *Journal of Strategic Security*, 6(3), 100–119.

Morgan, Charles A. III; Rabinowitz, Yarin; Palin, Beau & Kennedy, Kirk. (2015). Who should you trust? Discriminating genuine from deceptive eyewitness accounts. *The Open Criminology Journal*, 8, 49–59.

O'Neill Shermer, Lauren; Rose, Karen C. & Hoffman, Ashley (2011). Perceptions and credibility: Understanding the nuances of eyewitness testimony. *Journal of Contemporary Criminal Justice*, 27(2), 183–203. DOI: 10.1177/1043986211405886.

Pavlenko, Aneta; Hepford, Elizabeth & Jarvis, Scott (2019). An illusion of understanding: How native and non-native speakers of English understand (and misunderstand) their Miranda rights. *International Journal of Speech Language and The Law*, 26(2), 181–207. DOI: 10.1558/ijsll.39163.

Peace, Kristine A.; Brower, Krista L. & Rocchio, Alexandra (2015). Is truth stranger than fiction? Bizarre details and credibility assessment. *Journal of Police and Criminal Psychology*, 30(1), 38–49. DOI: 10.1007/s11896-014-9140-7.

Peer, Eyal; Rothschild, David; Gordon, Andrew; Evernden, Zak & Damer, Ekaterina (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643–1662. DOI: 10.3758/s13428-021-01694-3.

Powell, Richard (2008). Bilingual courtrooms: In the interests of justice? In Gibbons & Turell (Eds.). *Dimensions of Forensic Linguistics* (pp. 131–159). John Benjamins. Available at jbe-platform.com/content/books/9789027291158-aals.5.10pow (accessed 12 April 2026).

R Core Team (2025). *R: A Language and Environment for Statistical Computing* [Computer Software]. R Foundation for Statistical Computing. Available at R-project.org/ (accessed 12 April 2026).

Read, John & Nation, Paul (2006). An investigation of the lexical dimension of the IELTS Speaking Test. In McGovern & Walsh (Eds.), *IELTS Research Reports, 6*. IELTS Australia and British Council.

Reyes, Robert M.; Thompson, William C. & Bower, Gordon H. (1980). Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology*, 39(1), 2–12. DOI: 10.1037/0022-3514.39.1.2.

Rickford, John R. & King, Shareese (2016). Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4), 948–988.

Ruva, Christine L.; Sykes, Elizabeth C.; Smith, Kendall D.; Deaton, Lillian R.; Erdem, Sumeyye & Jones, Angela M. (2024). Battling bias: Can two implicit bias remedies reduce juror racial bias? *Psychology, Crime & Law*, 30(7), 730–757. DOI: 10.1080/1068316X.2022.2115494.

Schmid, Helmut (1997). Probabilistic part-of-speech tagging using decision trees. In Jones & Somers (Eds.), *New Methods in Language Processing*. Routledge.

Schmid, Monika S. & Jarvis, Scott (2014). Lexical access and lexical diversity in first language attrition. *Bilingualism: Language and Cognition*, 17(4), 729–748. DOI: 10.1017/S1366728913000771.

Tabachnick, Barbara G. & Fidell, Linda S. (2007). *Using Multivariate Statistics, 5th Ed*. Allyn & Bacon/Pearson Education.

Treffers-Daller, Jeanine (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In Scott Jarvis & Michael Daller (Eds.), *Studies in Bilingualism*, 47, 79–104. John Benjamins Publishing Company. DOI: 10.1075/sibil.47.05ch3.

Vidal, Karina & Jarvis, Scott (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, 24(5), 568–587. DOI: 10.1177/1362168818817945.

Wells, Gary L. & Turtle, John W. (1987). Eyewitness testimony research: Current knowledge and emergent controversies. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, 19(4), 363–388. DOI: 10.1037/h0080000.

Yoon, Hyung-Jo (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66, 130–141. DOI: 10.1016/j.system.2017.03.007.

Zenker, Fred & Kyle, Kristopher (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, 100505. DOI: 10.1016/j.asw.2020.100505.