ASSESSING THE ACCURACY OF ENGLISH-AS-A-SECOND-LANGUAGE EYEWITNESS TESTIMONIES AND CONTEMPORANEOUS OFFICER NOTES USING TWO METHODS

Meredith Allison and Cecily Basquin
Elon University

Jennifer Gerwing
Health Services Research Centre, Akershus University Hospital,
University of Victoria

Although English-as-a-Second Language (ESL) eyewitnesses interact regularly with police officers in the US and Canada, little research has examined their testimonies. This study sought to assess the testimony accuracy of 17 ESL witnesses, and the contemporaneous notes the officers made during free and cued recall questioning. We assessed accuracy using two methods: A checklist approach (CL) that has been used in past studies (e.g., List, 1986) and an inductive microanalysis of face-to-face dialogue (MFD) approach that was developed for this study. We found that witnesses gave more accurate information in free recall and made more errors in cued recall when both the CL and MFD methods of analysis were used. The same pattern of results held for the officer note data. When we directly compared the MFD and CL data, however, we found that the MFD method captured more information (both accurate and inaccurate witness details), suggesting that it provides richer accuracy data for eyewitness testimony and officer notes. Future research on ESL witness testimony using the MFD approach is discussed.

Keywords: eyewitness; ESL; officer notes; free recall; cued recall; memory

When people witness crimes in which the perpetrators are strangers, their memories are put to the test. The witnesses will be questioned repeatedly, typically by law enforcement and then, if the cases make it to trial, by lawyers. Witnesses also may be asked to view in-person or photographic lineups. While the accuracy of eyewitness identifications has been studied extensively for more than 30 years (e.g., Brewer & Wells, 2011; Buckhout,

Author note: Meredith Allison and Cecily Basquin, Department of Psychology, Elon University; Jennifer Gerwing, Health Services Research Centre, Akershus University Hospital, and Department of Psychology, University of Victoria.

In addition to acknowledging Elon University for its support of this project, we would like to thank the students and staff of the English Language Centre, University of Victoria (Canada). In particular, we thank Robbie Newton, who provided vital, concrete assistance, and feedback. Elizabeth Brimacombe and Janet Bavelas contributed constructive advice, especially in the design phase of the project. Finally, we thank John Price and Dianne Hartley for their help with creating the crime video.

Correspondence concerning this article should be addressed to Meredith Allison, Elon University, CB 2337, Elon, NC, 27244. E-mail: mallison5@elon.edu

1974; Wells et al., 1998), the accuracy of eyewitness testimonies has received far less of a focus. In psychology, Munsterberg (1908) was a pioneer in examining testimony accuracy, followed by a resurgence of interest in the 1970s (e.g., Ellis, 1975; Loftus & Palmer, 1974; Wells, Lindsay, & Ferguson, 1979). Since then, research on testimony accuracy often has involved checklists applied to the testimony itself (e.g., List, 1986). A complementary method for assessing accuracy is examining the notes that officers take when questioning witnesses. Such research has shown that officer notes often provide an incomplete accounting of the witnesses' words (e.g., Cauchi & Powell, 2009). A topic that largely has been ignored is the testimony of witnesses who speak English-as-a-Second-Language (ESL). Police officers in the United States and Canada have to interact with community members who are not proficient English speakers, so these witnesses are worthy of careful study (National Institute of Justice, 1999).

The accuracy of witness testimony is always important, but the accuracy of moderate proficiency English speakers deserves particular attention due to the increased probability of misunderstandings and language errors. Studying ESL witnesses intersects the subdisciplines of memory, language, and culture. Even though there are large numbers of ESL individuals in Canada and the United States (Karliner, Jacobs, Chen, & Mutha, 2007; Statistics Canada, 2011), relatively few psychologists in these subdisciplines have studied their interactions with law enforcement. This paper will focus on assessing the accuracy of ESL witness testimonies and contemporaneous officer notes. In the next sections, we review the literature on ESL witnesses and studies of testimony accuracy, focusing on methodological issues surrounding the measurement of accuracy.

Eyewitness Testimony Accuracy

The typical methodological approach for studying testimony accuracy involves naïve participants (often undergraduates) who view an event that includes an unexpected crime (live, photos, film, or video). They then are questioned about that event (written tests, interviews that are audio- or video-recorded). The participants' answers are scored for accuracy by comparing their responses to the actual event. Note that this paper does not include a discussion of Criterion-Based Content Analysis (Vrij, 2005) or Statement Validity Analysis (Brown, 2010), in which researchers assess the accuracy of eyewitness statements after observing an actual (i.e., out of the lab) crime. With these approaches, the truth or actual accuracy is unknown, and the researcher can only gauge the accuracy of witness statements by inference. The focus in this paper instead is on the laboratory situation, where the truth can be accessed directly - researchers know exactly what happened in the crime because they created the mock crime in the form of a videotaped or live event. Controlling the stimulus has the additional advantage of having multiple witnesses of the same event from the same point of view, creating the opportunity to discover which approaches to obtaining testimony lead to improved accuracy.

One of the most robust findings in the testimony accuracy literature has resulted from comparing witness accuracy in response to two different types of questioning: Free recall and cued recall. *Free recall questioning* typically involves open-ended questions that do not lead witnesses to particular answers. This is often accomplished by using a question

like "tell me everything you remember." Witnesses answer these kinds of questions with little direction from the interviewer or officers. In contrast, *cued recall questioning* is more specific, thereby allowing officers to probe particular areas they want to know more about (e.g., "what was he wearing?").

Several studies have noted that the accuracy of eyewitness testimony under free recall is greater than that under cued recall (e.g., Fisher, Brewer, & Mitchell, 2009; Lipton, 1977; List, 1986). Free recall questioning results in testimonies with fewer errors, but they can be incomplete, so there is a benefit to asking cued recall questions. Cued recall questioning can lead participants to remember new and accurate information because this memory task relies on cues that can help participants to remember (e.g., Thomson & Tulving, 1970). But the trade-off here is that errors can be introduced (List, 1986; Padilla-Walker & Poole, 2002). Both forms of questioning are seen as beneficial in investigative interviewing, but researchers suggest that witnesses should be questioned first with open-ended questions, followed by cued recall questions (Fisher & Schreiber, 2007). Researchers also have found that some witnesses perform well regardless of the type of questions (e.g., young adults in List, 1986), but other witnesses, such as older adults and young children, can be more prone to error under cued recall (e.g., Gordon & Follmer, 1994; List, 1986).

To date, there are no data to determine whether testimony accuracy in ESL witnesses is more accurate under free recall versus cued recall questioning. However, there are some data that focus on ESL witnesses' accuracy in other respects, such as the relationship between particular language features and memory (Fausey & Boroditsky, 2011), or the effect of misleading post-event information on accuracy (Shaw, Garcia, & Robles, 1997). Shaw et al. (1997) found that bilingual Spanish and English witnesses who were exposed to post-event information were equally likely to show a drop in accuracy without any drop in confidence, regardless of the language of the post-event information or language of testing. Outside of the eyewitness literature, however, research shows that memory accuracy may be affected by second-language speakers because of the increased cognitive load, reducing the amount that can be recalled (Service, Simola, Metsänheimo, & Maury, 2002).

Methods for testing testimony accuracy

Although several studies have reported examinations of the impact of the type of questioning on testimony accuracy, the methods used to assess accuracy vary. Some methods are removed from face-to-face interaction, such as asking witnesses to write their responses for subsequent accuracy scoring (e.g., Brewer, Potter, Fisher, Bond, & Luszcz, 1999; Gilbert & Fisher, 2006; Valentine & Maras, 2011). For example, participants in List (1986) completed questionnaires, which later were scored for accuracy using a reliable checklist (CL) of accurate information. Similarly, participants in Odinot, Wolters, and Giezen (2013) were asked to answer written questions in small units of information. These responses then were evaluated as correct or incorrect by scorers who had established interrater agreement.

Other researchers made audio- or video-recordings of witness testimony interviews, later either transcribing and analyzing them for accuracy or analyzing directly from

the recordings (e.g., Allison, Brimacombe, Hunter, & Kadlec, 2006; Brimacombe, Jung, Garrioch, & Allison, 2003; Brimacombe, Quinton, Nance, & Garrioch, 1997; Brock, Fisher, & Cutler, 1999; Gordon & Follmer, 1994; Krahenbuhl, Blades, & Eiser, 2009). Scoring can be done by watching or listening to the recordings while noting correct and incorrect statements on a detailed CL. For example, Dunning and Stern (1992) audio-recorded free recall testimony from witnesses who had viewed a police training video in which an officer was shot. They questioned witnesses on multiple occasions. Two raters used a CL to score the number of correct facts, errors, and confabulations. Similarly, Fisher and Cutler (1996) scored tape-recorded, face-to-face interviews with a CL in order to examine the relationship between consistency (over multiple interviews) and accuracy.

How the researchers operationalized accuracy varied across these studies. Some defined accuracy as the number of check marks on a check sheet (with errors being the number of incorrect answers on the CL; omissions were sometimes also calculated, Gordon & Follmer, 1994). Others defined accuracy as the proportion of accurate statements given the total number of statements made (e.g., Allison et al., 2006). Regardless of whether they used a written information or recorded interview approach, some researchers were clear that they had used detailed rules and calculated inter-rater reliability (e.g., Dunning & Stern, 1992; Krahenbuhl et al., 2009); some worked to consensus with multiple scorers (e.g., Fisher & Cutler, 1996); while others omitted this information from their published papers, so it is not clear whether this was done (e.g., Gordon & Folmer, 1994).

Another method for assessing witness accuracy is to analyze the notes that officers take during the investigative interview. The National Institute of Justice (1999) recommended that witness interviews be recorded (notes, audio, or video). However, digital recording devices are not always available in the field and may fail to capture the witnesses' words adequately (Hyman Gregory, Schreiber Compo, Vertefeuille, & Zambruski, 2011). Even if recording devices are available, they may not be turned on until after the initial interview, once officers decide to open a formal investigation (Cauchi & Powell, 2009). Thus, contemporaneous note-taking may be the only method of documenting witness testimony in some interviews. One survey of U.S. officers showed that more than 95% of them reported taking notes while interviewing witnesses (Hyman Gregory et al., 2011). The officers reported that these notes were helpful for facilitating their later memory of what had been said and for making formal police reports. When officers did take notes, they were typically not verbatim, as officers reorganized and reworded what witnesses actually said (Hyman Gregory et al, 2011). Although the notes can include a lot of accurate detail (Hyman Gregory et al., 2011), officers often make errors of omission, leaving out both peripheral and central crime details (Cauchi & Powell, 2009; Lamb, Orbach, Sternberg, Herschkowitz, & Horowitz, 2000). For example, Lamb et al. (2000) found that investigators left out 25% of the information that alleged child victims reported in sexual abuse cases in Israel. However, errors of commission are rare (Lamb et al., 2000).

Another problematic aspect of contemporaneous note-taking is that officers tend to omit documenting their actual questions (Cauchi & Powell, 2009; Hyman Gregory et al., 2011; Lamb et al., 2000). The type of questioning is not trivial: Later triers of fact will want

to know whether the witness spontaneously offered up the fact in free recall or whether it only emerged in cued recall questioning where post-event information may have been introduced (Cauchi & Powell, 2009). Some officers incorrectly may record information as coming from free rather than cued recall when, in fact, those details were elicited via focused questioning (Lamb et al., 2000). In Lamb et al. (2000), officers neglected to portray accurately how information was elicited from the alleged child victims. These studies show that officer notes capture accurate but incomplete information and may contain attribution errors. Few studies have systematically examined officer notes, and none have examined officer notes when the witness is a non-native English speaker. Clearly more research on officer note-taking is needed to more fully describe their accuracy.

Current Study

This review of the literature shows considerable variability in how testimony accuracy has been defined and measured and, to date, no one has systematically examined the testimony accuracy of ESL witnesses. In this paper, we propose a new method for assessing testimony accuracy: Using microanalysis of face-to-face dialogue (MFD), which is an open-ended, inductive method for the "detailed and replicable examination of any aspect of observable communication" (Bavelas, Gerwing, Healing, & Tomori, 2016, pp. 129-130). Here, we used MFD to examine participants' speech in context, closely examining each utterance's meaning. The method connected the accuracy analyses to the participants' speech and used annotation software to track decisions. We also approached the data inductively, by first creating accuracy operational definitions based on pilot data, then calculating interrater reliability for those definitions, and finally analyzing the complete data set. Inductive analysis can lead to new insights (Bavelas, Kenwood, Phillips, 2002) that top-down approaches (like CLs) may miss.

The purpose of the current study was, therefore, to examine the accuracy of ESL witness testimony and contemporaneous officer notes using two methods: A checklist (CL) and microanalysis of face-to-face dialogue (MFD). We also compared accuracy as a result of free and cued recall questioning and directly compared officer notes with the witnesses' testimonies. The use of MFD was exploratory because we created new operational definitions inductively from the data. We moved away from an analysis of paper transcripts of eyewitness testimony, instead using annotation software to assess accuracy, which may lead to a more precise and comprehensive analysis. We hypothesized that the witnesses would be more accurate in free than in cued recall (Lipton, 1977). We also hypothesized that the officers' notes would be incomplete when compared with direct eyewitness testimonies, failing to completely capture what the witnesses had said (Lamb et al., 2000).

METHOD

Participants

Seventeen dyads participated. Each dyad consisted of one non-native English speaker, who was assigned to the role of eyewitness, and one native-English speaker, who was assigned to the role of officer. The ESL eyewitnesses were recruited from an English Language center at a western Canadian university and received a gift card for their partici-

pation. The police officers were recruited using the same university's psychology online participant pool and received course credit in return for their participation.

The ESL eyewitnesses were recruited from 400-500 level English courses, placing them at approximately the B2 level in the Common European Framework of References for Languages. They were on average 24.94 years old (*SD* = 5.84), and their self-reported first languages were: Spanish (*n* = 5), Japanese (*n* = 3), Portuguese (*n* = 3), Korean (*n* = 3), Cantonese/Mandarin (*n* = 2), and Arabic (*n* = 1). Eleven of the witnesses were female, and six were male. All police officers were native English speakers, and their average age was 21.52 years (*SD* = 3.87). Seven were male, and ten were female.

Eyewitness Procedure

The following procedure was approved by the research ethics boards of the authors' universities. The participants (one ESL eyewitness and one police officer) arrived at the laboratory and signed consent forms. The ESL eyewitnesses were moved to another room where they watched a mock crime video in which a woman's purse was stolen. These eyewitnesses had not been told that the video they would be viewing included a crime. During this time, in the main room, the police officers prepared for an unrelated task. The witnesses returned, and both participants spent the next 15 minutes doing an unrelated task that involved first-aid instructions. This task served as a rapport builder and as a way for time to pass and for memory to degrade. Then, the experimenter returned and explained that the ESL students earlier had witnessed a mock crime and that the other student would now play the role of a police officer. The officer would interview the ESL eyewitness to find out about the crime. At this point, the eyewitnesses and officers were seated so that they were recorded by lab cameras simultaneously on a split screen. Once the experimenter left the room, they began the interview.

The police officers used a structured interview protocol that directed them to ask an open-ended, free-recall question (*tell me everything you can remember about the movie you saw earlier*), followed by specific cued recall questions to probe for eyewitness descriptions of the perpetrator and his vehicle. The police officers were instructed to make notes on the interview protocol in the spaces provided. These interviews lasted an average of 11.76 minutes (*SD* = 4.87). At the conclusion of the study, the participants were debriefed. They watched the video of their participation and completed a video release form, specifying how their videos could be used in the future. All participants allowed for their videos to be analyzed.

Eyewitness Testimony Accuracy Analysis

In order to measure the accuracy of each eyewitness account, two methods for analyzing accuracy were used: A CL of accurate statements and MFD, namely a set of operational definitions, detailed analytical procedures, and reliability that we developed. We used ELAN software (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) for annotating participants' speech and all analyses.

Checklist. The CL of accurate statements was created by watching the original crime video and selecting all important event and crime details. The CL included 78 accu-

rate details about the setting, victim, criminal, actions, and the criminal's vehicle. To analyze for accuracy using the CL, each interview was watched from the beginning, paying close attention to the statements the witness made. If the statement corresponded with an accurate statement on the CL, a checkmark was placed next to the corresponding item. Different color checkmarks were used to distinguish between accurate and inaccurate statements in response to the free recall and cued recall questions (contact the first author for a copy of the CL rules). Inter-rater reliability was calculated to ensure the CL and rules were reliable. Two of the authors independently analyzed two groups (11.7% of the data) and compared answers by calculating percent agreement (# agreements / # agreements + # disagreements). Each individual item was compared across raters, and only exact matches were counted as agreements (i.e.,reliability was not calculated by correlating the total number of accurate items, but instead by checking each individual item for accuracy). This resulted in a more precise and conservative assessment of reliability. An overall reliability score of 89.7% agreement was achieved. Any disagreements were resolved through discussion, and the final data include the agreed-upon decisions. Once the rules were found to be reliable, all of the remaining dyads were analyzed for accuracy using the CL method by one of the authors.

Microanalysis. Using ELAN software, analysts transcribed the participants' speech, linking it (and all subsequent analysis) directly to the video. For all measures described here, we first developed them on pilot data and one real dyad before moving on to analyzing the remaining 16 dyads.

The analysis of accuracy began by identifying all idea units. We defined *idea units* as single thoughts related to the crime (Dritschel, 1991). For example, "I saw a woman and she was talking on the phone," was coded as two idea units, because the statement contained two items of information. Although most idea units came from the witnesses, sometimes the idea unit was initiated by the police officer and subsequently accepted or rejected by the witness. Accepted idea units were included as idea units in this analysis. We excluded parts of the conversation that were not direct statements of information, such as hedges, disclaimers, qualifiers (um, I think, I don't know) and backchannels (uh huh, yeah). Detailed rules for marking idea units were created, and all idea units were identified by consensus (contact the first author for the idea unit decision rules).

Once the idea units were identified, we analyzed each for accuracy. An idea unit could be categorized as *accurate*, *inaccurate*, *repeated accurate*, or *repeated inaccurate*. The categories were mutually exclusive; that is, each idea unit could be placed into only one category. An accurate idea unit was a single item or event that was congruent with what occurred in the crime video (e.g., "I saw a woman" when the victim in the video was female). An inaccurate idea unit was defined as a single item that was incongruent with what was on the video (e.g., "he wore *jeans*" when the perpetrator wore green chinos). If a participant's idea unit was partly correct and partly incorrect (e.g., "his sweater was *red*" when the perpetrator wore a beige sweater), the idea unit was categorized as inaccurate.

An idea unit was categorized as a repeated accurate idea unit when the witnesses restated an accurate fact that they had mentioned earlier in questioning (either in free or

cued recall). For example, one participant said in her free recall testimony that the victim "put her bag in the path." Later, during cued recall, this same witness repeated that "her bag is in the path." Thus, the information was accurate, but it was not new. Similarly, a repeated inaccurate idea unit was when a participant repeated an inaccurate statement. One witness said that the perpetrator's vehicle was "like Cherokee, that kind of car." She later said "like this kind of Cherokee." Again, the information was not new and was still inaccurate.

Two of the authors analyzed all of the free and cued recall portions of the interviews independently for two groups (11.7% of the data) using MFD. Percent agreement was calculated as with the CL approach. Each idea unit was compared across raters, and exact matches were counted as agreements. Note that most disagreements involved distinctions between new and repeated, rather than between accurate and inaccurate. An acceptable level of reliability was reached (85.3%). All disagreements were resolved, and the final agreed-upon decisions were included in the analyses.

Officer Notes Accuracy Analysis

As with the direct eyewitness testimonies, we analyzed the officer notes for accuracy using the CL approach. For a bottom-up comparison, we could not use MFD, because the officer notes were not face-to-face dialogues; instead, we used an inductive discourse analytic (DA) approach. This method used the same rules described above, with the discourse analysis completed on paper. Two of the authors independently analyzed the officers' notes on two randomly selected groups (11.7% of the data). We calculated inter-rater reliability for the CL approach using percent agreement, and the level of agreement was 89.5%. We also analyzed two groups using the DA approach and reached an acceptable level of agreement (85.5%). All disagreements were resolved, and then one author analyzed the remaining notes.

RESULTS

Eyewitness Testimony Accuracy

Checklist Method. Descriptive statistics were run on the number of accurate and inaccurate CL items in free and cued recall and are presented in Table 1. A repeated measures Analysis of Variance (ANOVA) of question type (free vs. cued recall) on accurate CL items was conducted and was significant, Wilk's Lambda = .21, $F(1, 16) = 60.23$, $p < .001$, partial eta-squared = .79. As hypothesized, there were more accurate CL items in free than cued recall (see Table 1). Another repeated measures ANOVA of question type (free vs. cued) on inaccurate CL items was conducted and also was significant, Wilk's Lambda = .33, $F(1, 16) = 32.26$, $p < .001$, partial eta-squared = .67. Consistent with past studies, there were significantly more errors in cued than free recall.

Table 1. *Means and Standard Deviations for Accurate and Inaccurate Free and Cued Recall Checklist Items in Direct Eyewitness Testimony*

	Free Recall *M* (*SD*)	Cued Recall *M* (*SD*)	ANOVA *p* value
Accurate	17.41 (5.29)	5.94 (1.68)	$p < .001$
Inaccurate	0.94 (0.66)	3.88 (2.12)	$p < .001$

Microanalysis of face-to-face dialogue. Descriptive statistics on the accuracy measures in free and cued recall are presented in Table 2. A series of repeated measures ANOVAs of question type (free vs. cued recall) on the idea unit measures were conducted. Question type significantly affected the number of accurate (non-repeated) idea units, Wilk's Lambda = .54, $F(1, 16) = 13.39$, $p < .01$, partial eta-squared = .46. As hypothesized, there were more accurate idea units in free than cued recall (see Table 2). Question type did not affect the number of repeated accurate idea units ($p > .05$). Turning to errors, there was no significant difference in the number of inaccurate idea units in free versus cued recall ($p > .05$), but there was a significant impact of question type on inaccurate *repeated* idea units, Wilk's Lambda = .76, $F(1, 16) = 5.06$, $p < .05$, partial eta-squared = .24. There were more inaccurate repeated idea units in cued recall than in free recall.

Table 2. *Means and Standard Deviations for Accurate and Inaccurate Free and Cued Recall MFD Idea Units in Direct Eyewitness Testimony*

Idea Unit	Free Recall *M* (*SD*)	Cued Recall *M* (*SD*)	ANOVA Significance (*p*)
Accurate	17.18 (7.38)	9.76 (4.29)	$p < .01$
Inaccurate	8.71 (8.20)	5.59 (5.23)	n.s.
Accurate Repeated	4.35 (3.87)	5.06 (2.95)	n.s.
Inaccurate Repeated	.41 (.71)	1.24 (1.52)	$p < .05$

Checklist versus Microanalytic Methods. We hypothesized that MFD would yield a richer picture of the data than the CL approach. We compared the two methods directly by conducting several paired samples *t*-tests, as shown in Table 3. We found that the MFD approach captured more accurate and inaccurate information overall and within cued recall. The MFD approach led to capturing higher inaccuracy scores in free recall, but there were no differences between MFD and CL in accurate free recall scores. Thus on 5/6 *t*-tests, the MFD captured more information than the CL approach, showing strong support for our hypothesis.

Table 3. *Means, Standard Deviations, and Tests of Significance for Accurate and Inaccurate Free and Cued Recall MFD Idea Units and CL Items in Direct Eyewitness Testimony*

Idea Unit/Item	Microanalysis *M* (*SD*)	Checklist *M* (*SD*)	ANOVA Significance (*p*)
Total Accurate	26.94 (7.85)	23.35 (4.95)	$t(16) = -2.83$, $p < .05$
Total Inaccurate	8.64 (5.64)	4.82 (2.30)	$t(16) = -4.33$, $p < .01$
Free Recall Accurate	16.53 (7.03)	17.41 (5.29)	n.s.
Free Recall Inaccurate	4.18 (3.91)	0.94 (0.66)	$t(16) = -3.41$, $p < .01$
Cued Recall Accurate	9.76 (4.29)	5.94 (1.68)	$t(16) = -3.98$, $p < .01$
Cued Recall Inaccurate	5.00 (2.87)	3.88 (2.12)	$t(16) = -2.15$, $p < .05$

Officer Note Accuracy

We ran the same accuracy analyses on the officer notes as we did above on the direct eyewitness testimonies. We found a very similar pattern of results for free versus cued recall for both the CL and inductive DA methods. These results are reported in Tables 4 and 5. These results confirm that there were more accurate items in free than cued recall and more errors in cued than free recall. However, some differences between the direct testimony data and officer note data emerged when we compared the DA and CL approaches: There were fewer significant findings in the officer note data. For the total number of accurate and inaccurate DA and CL items using paired-samples *t*-tests, the results were not significant ($p > .05$). Similarly, there were no significant differences when comparing accurate free and cued recall DA and CL items ($p > .05$). However, there were two significant findings. The DA method yielded higher scores for inaccurate free recall idea units ($M = 2.88$, $SD = 2.74$), when compared with the inaccurate free recall CL items ($M = 1.71$, $SD = 2.20$), $t(16) = -4.52$, $p < .001$. In addition, there were higher scores for inaccurate cued recall DA idea units ($M = 3.59$, $SD = 2.35$), when compared with inaccurate cued recall CL items ($M = 2.12$, $SD = 1.36$), $t(16) = -2.82$, $p < .05$.

Table 4. *Means and Standard Deviations for Accurate and Inaccurate Free and Cued Recall Checklist Items for the Officer Notes*

	Free Recall *M* (*SD*)	Cued Recall *M* (*SD*)	ANOVA *p* value
Accurate	11.29 (3.57)	4.29 (2.34)	$p < .001$**
Inaccurate	1.71 (2.20)	2.12 (1.36)	n.s.

Note. **Wilk's lamba, $F(1, 16) = 29.88$, $p < .001$, partial eta-squared = .65

Table 5. *Means and Standard Deviations for Accurate and Inaccurate Free and Cued Recall DA Idea Units for the Officer Notes*

Idea Unit	Free Recall *M* (*SD*)	Cued Recall *M* (*SD*)	ANOVA Significance (*p*)
Accurate	11.47 (4.77)	4.94 (2.80)	$p < .001$**
Inaccurate	2.88 (2.74)	3.59 (2.35)	n.s.
Accurate Repeated	0.82 (0.88)	1.06 (1.20)	n.s.
Inaccurate Repeated	0.00 (0.00)	0.65 (1.06)	$p < .05$*

Notes. **Wilk's lamba, $F(1, 16) = 20.63$, $p < .001$, partial eta-squared = .56. *Wilk's lamba, $F(1, 16) = 6.37$, $p < .05$, partial eta-squared = .29.

Eyewitness Testimony vs. Officer Notes

We compared the direct eyewitness testimony data with the officer notes data for both CL and inductive methods (also comparing cued and free recall) using a series of paired samples *t*-tests. The scores on 13 accuracy variables were significantly higher for direct eyewitness testimony than the officer notes (Table 6). These data showed that the officer notes did not capture as much information as the witnesses actually provided in their testimonies on a number of measures in both the CL and inductive approaches and across cued and free recall.

Table 6. *Means and Standard Deviations for Direct Eyewitness Testimony and Officer Notes Comparisons*

	Direct Eyewitness Testimony *M* (*SD*)	Officer Notes *M* (*SD*)	*t*(df), *p*
CL			
Total accurate CL items	23.35 (4.95)	15.53 (0.72)	8.65(16), $p < .001$
FR accurate CL items	17.41 (5.29)	11.29 (3.57)	5.87(16), $p < .001$
CR accurate CL items	5.94 (1.68)	4.29 (2.34)	2.96(16), $p < .05$
CR inaccurate CL items	3.88 (2.12)	2.12 (1.36)	3.16(16), $p < .01$
MFD (or DA)			
Accurate IUs	26.94 (7.85)	16.41 (5.10)	6.72(16), $p < .001$
Inaccurate IUs	8.94 (5.64)	6.47 (3.64)	2.38(16), $p < .05$
Accurate repeated IUs	16.24 (13.35)	1.88 (1.22)	4.44(16), $p < .001$
Inaccurate repeated IUs	1.65 (1.73)	0.65 (1.06)	4.12(16), $p < .01$
FR accurate IUs	16.53 (7.03)	11.47 (4.77)	3.19(16), $p < .01$
FR accurate repeated IUs	10.59 (10.23)	0.82 (0.88)	3.97(16), $p < .01$
FR inaccurate repeated IUs	0.41 (.62)	0.00 (0.00)	2.75(16), $p < .05$
CR accurate IUs	9.76 (4.29)	4.94 (2.79)	3.93(16), $p < .01$
CR accurate repeated IUs	5.59 (5.23)	1.06 (1.20)	4.03(16), $p < .01$
CR inaccurate repeated IUs	1.24 (1.52)	0.65 (1.06)	2.42(16), $p < .05$

Note. MFD = microanalysis, DA = Discourse Analysis, CL = checklist, FR = free recall, CR = cued recall, IU = idea unit

DISCUSSION

This paper examined the testimony accuracy of ESL witnesses, a relatively new focus for psychology and law researchers. We confirmed the results of past studies, in that free recall accuracy was higher than cued recall accuracy (Fisher et al., 2009). This finding held when the testimony was provided by moderately proficient speakers of English. This suggests that it is to the officers' advantage to ask open-ended questions first, giving ESL witnesses the opportunity to express what they can without interruption. When officers did ask cued recall questions, the number of errors increased, so officers should carefully choose which questions to ask and should be selective in their word choices so as to reduce misunderstandings and errors.

We also developed a new method of assessing testimony accuracy in ESL witnesses that can be used with other types of participants. We created an approach using microanalysis of face-to-face dialogue (MFD) that involved a more in-depth analysis of the participants' words. This approach yielded more information and appeared to be more detailed than the checklist (CL) approach. Although the CL was a precise measurement tool, two raters could achieve agreement by checking the same item on the CL, even if the two checks were based on two different observations (different witness utterances). Thus, this level of agreement could hide actual errors and be misleading. Such errors did not occur

with the MFD because each decision was directly connected to an utterance (an idea unit). That is, the level of inter-rater reliability was a precise reflection of analysts' decisions about each idea unit. We found MFD to be more flexible than the CL, which had no means for scoring items witnesses mentioned but were not on the list. In contrast, MFD helped us to create operational definitions that were built from the ground-up, providing a close connection between the analysis and the data. We need to replicate these findings on other samples to ensure that the superiority of the MFD approach is not restricted only to the ESL sample. MFD is also designed for inductive research, providing a means for researchers to discover new variables worthy of study that they had not anticipated (Bavelas et al., 2002). Using open-ended software like ELAN also facilitated adding in new variables after the fact because the software allows for multiple tiers and sub-tiers.

Overall, MFD in combination with the CL approach could give a more comprehensive description of accuracy. The CL approach can be used to provide the number of overall items remembered correctly, incorrectly, and those that were omitted. Researchers can further break down the CL into free and cued recall (or central and peripheral details) and describe how much information of each type is remembered. The MFD approach would then involve careful transcription and would provide a solid measurement of word count (broad measure of verbosity) and idea units (specific measure of crime-relevant ideas). These idea units could be analyzed, and researchers could count the number of units that are repeated (both correctly repeated and incorrectly repeated). With the MFD approach, researchers could choose to move beyond simply the accuracy of idea units and analyze qualifiers that mark confidence in those idea units, elaborations, and hedges. Also they could analyze sequences of interaction between the interviewer and witness that were not related to the crime. Combining the CL and MFD approaches would give researchers broad strokes and fine-tuned analyses of testimony accuracy.

This paper adds to the small body of literature that shows that contemporaneous officer notes capture some accurate information but are incomplete representations of what the witnesses actually conveyed in their testimonies (Cauchi & Powell, 2009). This finding has been replicated in basic research that revealed that native-speaking listeners might not pick up all the representative information conveyed by non-native speakers (Lev-Ari & Keysar, 2012). We did not find strong evidence that the inductive discourse analysis (DA) approach was more effective at capturing the accuracy of the officer notes, perhaps because the officers' notes were relatively brief when compared with the testimonies of the witnesses. Thus, it may be that the CL approach is adequate for assessing officers' written text and that an in-depth DA method may not be necessary.

Findings regarding officer notes support the best practices recommendations of digitally recording all interactions with witnesses (National Institute of Justice, 1999). One possible reason officers missed documenting information in their notes could be that they were in a cognitively and socially demanding situation: Each officer in this study needed to complete the assigned task (ask specific questions and take notes), while undertaking the additional challenge of interacting with an ESL witness. Officers had to ensure that they understood what the witnesses were saying (e.g., dealing with pronunciation issues, unin-

tended errors in vocabulary, ambiguous speech), while ensuring that their own questions were understood by witnesses. Future studies that directly compare the accuracy of officer notes taken when the witness is a native versus a non-native English speaker would add to our understanding of officer-ESL witness interactions.

Another possible avenue of investigation would be to examine the quality and accuracy of the notes that a second officer takes while the first officer interviews. A second officer note-taker could reduce the cognitive load of the interviewer. However, this approach would have to recognize that the second officer would be merely an *overhearer* of the testimony, rather than an *addressee* (which the first officer would be). Research has shown that overhearers' understanding is significantly less accurate than addressees' understanding, simply because of their inability to interact with the speaker (Schober & Clark, 1989). A further avenue of investigation could be to test whether officers should be trained on *what* their notes should contain. Should officers try to write down everything the witnesses say because it could turn out to be forensically important? Officers should also be sure to write down their own questions and clarifications (e.g., Lamb et al., 2000), particularly with ESL witnesses. For example, ESL witnesses may be unfamiliar with some vocabulary in officers' questions, and officers may have to find ways to define these terms before witnesses are able to answer. In this study's material, ESL witnesses often needed help understanding the term "license plate." If officers fail to note this understanding problem in their notes (recording simply that the witnesses did not see the license plate), then they could be recording inaccurate information. They also could misrepresent the witnesses' lan-guage skills by suggesting that this term was used and understood. Further, officers should be trained on *how* to take notes, such as using standard abbreviations (Cauchi & Powell, 2009). Better officer training is needed to get more uniformity and better quality officer notes. This training should involve practice and feedback so that officers can improve their skills (Fisher, 1995).

Limitations and Future Directions

The sample size in this study was small, limiting our ability to generalize to all ESL witnesses of moderate English proficiency. Two practical issues explain the small sample size. First, ESL participants were difficult to recruit: The language barrier itself was an impediment, with many ESL students lacking the confidence to go to an unfamiliar part of the campus to participate in a study during which they would be recorded speaking their second language. A second reason is that MFD requires many hours of labor-intensive and careful work. We should note that our sample size is in line with other MFD studies (e.g., Bavelas, Gerwing, Sutton, & Prevost, 2008; see Appendix 9A in Bavelas et al., 2016) and other studies examining officer notes (e.g., Hyman Gregory et al., 2011; Schreiber Compo, Hyman Gregory, & Fisher, 2012). However, future studies could aim to increase the sample size, while still including participants with a variety of first languages. Another limitation is that the sample was composed of witnesses who spoke English with moderate proficiency. How the results would hold up with ESL witnesses of lower or higher proficiency is an open question.

In the future, researchers should replicate these findings on other groups of ESL witnesses to see whether accuracy is still superior in free recall or whether the free-cued recall distinctions start to fade as proficiency decreases. These findings should also be replicated on groups other than ESL witnesses to see whether the MFD approach yields more precise and complete info than the CL approach on native English speakers. Another avenue for future research would be to show the videotaped testimonies to triers-of-fact. Will mock jurors perceive ESL witnesses to be accurate in both free and cued recall? And will they be seen as credible? Past studies suggest that ESL witnesses may be seen as less credible (Lev-Ari & Keysar, 2010) or as more likely to be deceptive (Evans & Michael, 2014), particularly when their accents are strong (Brennan & Brennan, 1981). Finally, more research into officer note accuracy and training should be conducted as it is likely that officers will continue to use their notes when creating police reports (Hyman Gregory et al., 2011). Note-taking may be more cognitively taxing when the witnesses are of lower English proficiency, so future research on witnesses with a range of language skills should be conducted. Further, trained officers (rather than students) would likely interview and take notes in different ways, so the validity of future studies with ESL witnesses would be enhanced by including trained criminal investigators.

Conclusion

The number of officer interactions with ESL witnesses will likely increase as more and more non-native English speakers live in the US and Canada. Officers should be careful when interviewing them, asking open-ended, free recall questions first, followed by essential cued recall questions. Officers should also be careful in their note-taking, being sure to write down as much information as possible from both officers and witnesses. In addition, researchers should further investigate the best ways to assess accuracy in the lab, as there is a lot of variability in the methods used. We suggest that a combination of the MFD and CL approaches would yield the most complete picture of witness accuracy.

REFERENCES

Allison, M., Brimacombe, E., Hunter, M. A., & Kadlec, H. (2006). Young and older adult eyewitnesses' use of narrative features in testimony. *Discourse Processes, 41,* 289-314. doi: 10.1207/s15326950dp4103_3

Bavelas, J. B., Gerwing, J., Healing, S., & Tomori, C. (2016). Microanalysis of face-to-face dialogue: An inductive approach. In C.A. Van Lear & D. J. Canary (Eds.), *Researching communication interaction behavior: A sourcebook of methods and measures* (pp. 129-157). Thousand Oaks, CA: Sage.

Bavelas, J. B., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58,* 495-520.

Bavelas, J. B., Kenwood, C., & Phillips, B. (2002). Discourse analysis. In M. Knapp & J. Daly (Eds.), *Handbook of interpersonal communication* (pp.102-129). Newbury Park, CA: Sage.

Brennan, E. M., & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language and Speech, 24,* 207-221.

Brewer, N., Potter, R., Fisher, R. P., Bond, N., & Luszcz, A. M. (1999). Beliefs and data on the relationship between consistency and accuracy of eyewitness testimony. *Applied Cognitive Psychology, 13,* 297-313. doi: 10.1002/(SICI)109907

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science, 20,* 24-27. doi: 10.1177/0963721410389169

Brimacombe, C. A. E., Quinton, N., Nance, N., & Garrioch, L. (1997). Is age irrelevant? Perceptions of young and old eyewitnesses. *Law and Human Behavior, 21,* 619-634.

Brock, P., Fisher, R. P., & Cutler, B. L. (1999). Examining the cognitive interview in a double-test paradigm. *Psychology, Crime & Law, 5*, 29-45. doi: 10.1080/10683169908414992

Brown, J. M. (2010). Statement validity analysis. In J. M. Brown & E. A. Campbell (Eds.), *The Cambridge handbook of forensic psychology* (pp. 319-326). New York, NY: Cambridge University Press.

Buckhout, R. (1974). Eyewitness testimony. *Scientific American, 231,* 23-31.

Cauchi, R., & Powell, M. B. (2009). An examination of police officers' notes of interviews with alleged child abuse victims. *International Journal of Police Science & Management, 11,* 505-515. doi: 10.1350/ijps.2009.11.4.147

Dritschel, B. H. (1991). Autobiographical memory in natural discourse. *Applied Cognitive Psychology, 5,* 319-330. doi: 0888-4080/91/040319-124$06.00

Dunning, D., & Stern, L. B. (2006). Examining the generality of eyewitness hypermnesia: A close look at time delay and question type. *Applied Cognitive Psychology, 6,* 643-657. doi: 10.1002/acp.2350060707

Ellis, H. D. (1975). Recognizing faces. *British Journal of Psychology, 66,* 409-426.

Evans, J. E., & Michael, S. W. (2014). Detecting deception in non-native English speakers. *Applied Cognitive Psychology, 28,* 226-237. doi: 10.1002/acp.2990

Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eyewitness memory. *Psychonomic Bulletin Review, 18,* 150-157. doi: 10.3758/s13423-010-0021-5

Fisher, R. P. (1995). Interviewing victims and witnesses of crime. *Psychology, Public Policy, and Law, 1,* 732-764. doi:10.1037/1076-8971.1.4.732

Fisher, R. P., Brewer, N., & Mitchell, G. (2009). The relation between consistency and accuracy of eyewitness testimony: Legal versus cognitive explanations. In T. Williamson, R. Bull, & T. Valentine (Eds.), *Handbook of psychology of investigative interviewing: Current developments and future directions* (pp. 121-136). Oxford, UK: Wiley-Blackwell. doi: 10.1002/9780470747599.ch8

Fisher, R. P., & Cutler, B. L. (1996). The relation between consistency and accuracy of eyewitness testimony. In G. Davies, S. Lloyd-Bostock, M. McMurran,, & C. Wilson (Eds.), *Psychology and law: Advances in research* (pp. 21-28). Berlin, Germany: De Gruyter.

Fisher, R. P., & Schreiber, N. (2007). Interview protocols to improve eyewitness memory. In M. P. Toglia, J. D. Read, D. F. Ross, & R. C. L. Lindsay (Eds.), *The handbook of eyewitness psychology, Vol 1: Memory for events* (pp. 53-80). Mahwah, NJ: Erlbaum.

Gilbert, J. A. E., & Fisher, R. P. (2006). The effects of varied retrieval cues on reminiscence in eyewitness memory. *Applied Cognitive Psychology, 20,* 723-739. doi: 10.1002/acp.1232

Gordon, B. N., & Follmer, A. (1994). Developmental issues in judging the credibility of children's testimony. *Journal of Clinical Child Psychology, 23,* 283-294. doi:10.1207/s15374424jccp2303_6

Hyman Gregory, A., Schreiber Compo, N., Vertefeuille, L., & Zambruski, G. (2011). A comparison of US police interviewers' notes with their subsequent reports. *Journal of Investigative Psychology and Offender Profiling, 8,* 203-215. doi: 10.1002/jip.139

Karliner, L. S., Jacobs, E. A., Chen, A. H., & Mutha, S. (2007). Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. *Health Services Research, 42,* 727-754. doi: 10.1111/j.1475-6773.2006.00629.x

Krahenbuhl, S., Blades, M., & Eiser, C. (2009). The effect of repeated questioning on children's accuracy and consistency in eyewitness testimony. *Legal and Criminological Psychology, 14,* 263-268. doi: 10.1348/135532508X398549

Lamb, M. E., Orbach, Y., Sternberg, K. J., Herschkowitz, I., & Horowitz, D. (2000). Accuracy of investigators' verbatim notes of their forensic interviews with alleged child abuse victims. *Law and Human Behavior, 24,* 699-708.

Lev-Ari, S., & Keysar, B. (2012). Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Processes*, *49*(7), 523-538.

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

Lipton, F. P. (1977). On the psychology of eyewitness testimony. *Journal of Applied Psychology, 62,* 90-95.
List, J. A. (1986). Age and schematic differences in the reliability of eyewitness testimony. *Developmental Psychology, 22,* 50-57. doi: 0012I649/86/S00.75
Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585-589.
Munsterberg, H. (1908). *On the witness stand: Essays on psychology and crime.* New York, NY: Doubleday.
National Institute of Justice. (1999). Eyewitness evidence: A guide for law enforcement. U.S. Department of Justice, Office of Justice Programs Research Report. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/178240.pdf
Odinot, G., Wolters, G., & Giezen, A. V. (2013). Accuracy, confidence, and consistency in repeated recall of events. *Psychology, Crime, and Law, 19,* 629-642. doi:10.1080/1068316X.2012.660152
Padilla-Walker, L. M., & Poole, D. A. (2002). Memory for previous recall: A comparison of free and cued recall. *Applied Cognitive Psychology, 16,* 515-524. doi: 10.1002/acp.809
Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211-232.
Schreiber Compo, N., Hyman Gregory, A., & Fisher, R. (2012). Interviewing barriers in police investigators: A field study of a current US sample. *Psychology, Crime and Law, 18,* 359-375. doi: 10.1080/1068316X.2010.494604
Service, E., Simola, M., Metsänheimo, O., & Maury, S. (2002). Bilingual working memory span is affected by language skill. *European Journal of Cognitive Psychology*, *14*(3), 383-408. doi:10.1080/09541440143000140
Shaw, J. S., Garcia, L. A., & Robles, B. E. (1997). Cross-language post event misinformation effects in Spanish-English bilingual witnesses. *Journal of Applied Psychology, 82,* 889-899. doi: 10.1037/0021-9010.82.6.889
Statistics Canada. (2011). *Linguistic characteristics of Canadians. Retrieved April 17, 2017 from http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.cfm*
Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86,* 255-282.
Valentine, T., & Maras, K. (2011). The effect of cross-examination on the accuracy of adult eyewitness testimony. *Applied Cognitive Psychology, 25,* 554-561. doi: 10.1002/acp.1768
Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11,* 3-41. doi: 10.1037/1076-8971.11.1.3
Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *6*, 603-647. doi:10.1023/A:1025750605807
Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology, 64,* 440-448.
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). *ELAN: A professional framework for multimodality research*. In Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation (pp. 1556-1559). 22-26 May, Genoa, Italy.

Date Received: 06/2016
Date Accepted: 01/2017

Suggested citation: Allison, M., Basquin, C., & Gerwing, J. (2017). Assessing the accuracy of English-as-a-second-language eyewitness testimonies and contemporaneous officer notes using two methods. [Electronic Version]. *Applied Psychology in Criminal Justice, 13*(1), 1-17.