Efficient coding of eyewitness narratives: A comparison of syntactic unit and word count procedures

JASON J. DICKINSON and DEBRA A. POOLE
Central Michigan University, Mt. Pleasant, Michigan

Researchers rely on two types of coding systems to evaluate eyewitness narratives. Features-of-events checklists offer coding simplicity but are impractical when target events are complex. Units-of-information (UOI) systems score all information reported, regardless of event complexity, but are difficult to implement. To test whether simpler systems would alter conclusions about memory performance, transcripts from children (3–8 years of age) who participated in an eyewitness study, originally coded using a UOI system called syntactic units (SU), were recoded using two word count procedures. Correlations between SU, modified word count, and raw word count values were high, and the proportion of information that was inaccurate was comparable across systems. Considering their high interrater reliability, procedural simplicity, and convergence with SU coding, word count procedures are efficient alternatives to UOI coding.

In the past decade, there has been a proliferation of research on the accuracy and completeness of children's eyewitness reports (e.g., Ceci & Bruck, 1995; Poole & Lamb, 1998). Although some studies have been motivated by basic developmental questions (e.g., age trends in autobiographical memory), the primary catalyst for renewed interest in children's testimony has been the challenge of investigating sexual abuse allegations. Because physical evidence is rarely available in such cases, researchers and helping professionals have tried to identify other characteristics that distinguish between true and false allegations. Two approaches have evolved for exploring which case features might be relevant when evaluating claims of sexual abuse: the indicator and the assessment approaches (Berliner & Conte, 1993; Poole & Lindsay, 1998).

The indicator approach seeks to identify symptoms, such as behavioral indicators, that discriminate sexually abused from nonabused children. At this time, however, there is a general consensus that no indicator or combination of indicators can substitute for the value of a child's report (Lamb, 1994). Consequently, the assessment approach is concerned with identifying factors associated with accurate versus inaccurate event reports. Research from this perspective typically manipulates such variables as age of witnesses, delay between events and questioning, and methods of questioning in order to determine which circumstances yield the most reliable reports. Contrived events (e.g., a magic show; Memon, Holley, Wark, Bull, & Kohnken, 1996) or natural occurrences (e.g., a pe-

diatric examination; Saywitz, Goodman, Nicholas, & Moan, 1991) furnish eyewitness investigations with recorded events against which to evaluate participants' autobiographical recall.

Coding the narratives generated from eyewitness studies requires investigators to consider two practical issues: (1) the unit of analysis for measurement and (2) the method for coding these units. Two coding strategies have become common practice: checklists based on features of events (FOE), in which narratives are typically described in terms of the percentage of correct and incorrect features recalled from a predetermined list, and coding schemes based on the number of correct and incorrect units of information (UOI) contained within each participant's narrative.

Characteristics of the Features-of-Events and Units-of-Information Coding Systems

To illustrate the FOE and UOI coding schemes, consider a common paradigm in which children are interviewed following a physical examination. Gordon et al. (1993) examined whether interviews conducted with dolls facilitate children's recall of a physical examination. Because the target event was an authentic pediatric check-up, Gordon et al. had physicians, nurses, and parents complete checklists that referenced the features of each child's examination (e.g., check ears, listen to chest). Some children were questioned verbally with the use of a doll, whereas other children were questioned verbally with no props. For each child, autobiographical recall was assessed as the proportion of features recalled from the checklist, and the data were evaluated as a function of age and interview condition.

Using a UOI coding system, Poole and White (1993) described the reports of participants who experienced an event 2 years earlier. Poole and White (1993) counted syn-

Correspondence concerning this article should be addressed to J. J. Dickinson, Department of Psychology, Florida International University, University Park Campus, DM256, Miami, FL 33199 (e-mail: jdicki01@fiu.edu).

tactic (i.e., grammatical) units (SUs), which included words that describe an actor (*she*), an action (*left*), or a direct object (*the apple*). Thus, "she left the apple" contained three SUs. In addition, individual SUs included quotes, descriptions (her hair was *black*), qualifiers (she was *very* beautiful), temporal information (*now*), and prepositional phrases (*on the bus*). Research assistants first parsed narrative responses into SUs, then marked each SU as accurate or inaccurate and calculated the proportion of information that was accurate in each child's responses.

In sum, FOE coding schemes are commonly rooted in a predefined set of events, often staged in the laboratory, and investigators typically report the proportion of correct and incorrect items recalled from a predefined list of event features (e.g., Greenstock & Pipe, 1996). With UOI coding procedures, however, investigators count features as they are mentioned by the witness, without reference to a list (e.g., Tobey & Goodman, 1992). Consequently, UOI systems code narratives as if there were an infinite feature list that included witness-specific details, such as individual reactions and conversations.

Pros and Cons of Features-of-Events and Units-of-Information Coding

Although a variety of dependent measures can be derived from FOE and UOI coding procedures, each offers discrete advantages and disadvantages. Both strategies are capable of providing the proportion of total information reported that is accurate or inaccurate. FOE coding methods, however, can also assess the completeness of a report by providing a measure of the proportion of events recalled from the pool of target events that were identified, *a priori*, by the investigators. As Baker-Ward, Ornstein, Gordon, Follmer, and Clubb (1995) reasoned, UOI coding procedures are generally not applicable to assessments of completeness, because target events are not specified. According to Baker-Ward et al. (1995), FOE coding also makes it possible for investigators to measure which information is lost or weakened over time, therefore providing information about consistency across repeated interviews. Although consistency is especially easy to track with FOE coding, such information can be coded using UOI systems, but the process is more cumbersome and prone to coding errors.

A major advantage of UOI coding procedures lies in their versatility in assessing circumstances ranging from events produced in the laboratory (e.g., a science demonstration; Poole & Lindsay, 1995) to natural occurrences (e.g., a hurricane; Bahrick, Parker, Fivush, & Levitt, 1998). With no need for a predefined checklist, investigators can code complex events, events that vary from witness to witness, or unrecorded events. When events are staged or videotaped, accuracy is determined by validating testimony against descriptions and photographs of the target events (e.g., Poole & Lindsay, 1995). In field research, however, details of the events are not always known,

which limits the ability to analyze testimony accuracy. Nonetheless, UOI coding can be used in field research when sheer amount of information is the major variable of interest (Fisher, Geiselman, & Amador, 1989; George, 1991) or when determinations of overall accuracy or inaccuracy are made on the basis of corroborating evidence (Yuille & Cutshall, 1986). Because no *a priori* list of relevant events is required for UOI coding, such systems have the advantage of counting all the information provided by witnesses, including descriptions of witnesses' emotional reactions and spontaneous elaborations of event details that were not deemed central by the investigators. Therefore, UOI coding schemes may provide more meaningful measures of how elaborated children's testimonies are as a function of age and condition.

Unfortunately, there are two reasons why UOI systems are difficult to implement. First, there are a large number of coding rules. The SU system, for example, attempts to code information the same regardless of variations in how witnesses phrase that information. In addition to other rules, coders are instructed to insert missing subjects into compound sentences (so that "He took the game. He put it on the table." is coded the same as "He took the game and put it on the table"), convert indirect quotes into direct quotes and count both as a unit, and delete information that is implied by the question (because the informativeness of a witness's answer is not affected by whether or not they included this redundant information). Many assistants lack the grammatical knowledge to learn this system. Second, even skilled coders must make several passes through a transcript, applying a subset of rules each time, because they have difficulty keeping all of the rules in mind simultaneously. Consequently, it is costly to train assistants and, in our laboratory, requires about 2 h to transcribe and code a typical eyewitness transcript.

The Present Study

The research question and the nature of the target event often dictate which coding method will be selected to assess autobiographical recall. An important question remains, however: To what extent do alternative coding procedures influence conclusions about memory performance? To answer this question, Baker-Ward et al. (1995) recoded data previously coded as FOE (i.e., Baker-Ward, Gordon, Ornstein, Larus, & Clubb, 1993), using the UOI criterion described by Saywitz et al. (1991). Correspondence between the two systems was evaluated by comparing rank orderings of children's levels of free recall. The resulting Spearman correlation coefficient of .91 indicated good agreement between the two methods of coding. Baker-Ward et al. (1995, citing a personal communication with Saywitz) also reported that results in another study were identical whether data were scored by UOI or by checking off how many possible propositions from the target event were mentioned by the participant. On the basis of these results, Baker-Ward et al. (1995)

concluded that general patterns of free recall performance across ages and events are unlikely to be artifacts of particular coding systems.

As the call for greater ecological validity has prompted investigators to use more elaborated events, UOI procedures have gained popularity, despite their complexity. The purpose of the present project was to develop new coding procedures that would result in faster data preparation when events are too complex for FOE coding. First, a system consisting of six rules was developed that could be implemented by scoring narratives on a computer keyboard, using common word-processing capabilities. This system, which we call modified word count (MWC), scrubs out much of the information that is not coded in the SU system prior to counting the remaining words. Second, we recoded narratives from a recent study of children's eyewitness memory, recording the original SU values and counting both MWC and raw word count (RWC) values. Finally, we conducted a series of analyses to compare results from the three coding systems. The primary question was whether greater coding efficiency could be obtained without altering interpretations of memory performance across ages and conditions.

METHOD

The Transcripts

A stratified random sample of 48 transcripts (8 each from ages 3–8) from children who had participated in an eyewitness study (Poole & Lindsay, in press), originally coded with a UOI system, were recoded with the MWC and RWC procedures. The transcripts were drawn from interviews of 114 children who had participated in three experimental sessions. In Session 1, each child interacted with "Mr. Science" (a male research assistant) for 16 min. This session began with Mr. Science setting a timer, followed by participation in four science demonstrations (e.g., blowing up balloons by mixing baking soda and vinegar in a bottle). Subsequently, a female interviewer asked each child nonsuggestive questions, thus collecting baseline data on children's initial reports of the target events.

In Session 2, which transpired in each participant's home approximately 3 months after Session 1, the impact of exposure to misinformation on children's reports was measured. Prior to this interview, parents read a story to their children on 3 consecutive days that described two science demonstrations their children had experienced, two fictitious (i.e., nonexperienced) science demonstrations, and one of two descriptions of nonexperienced touching (e.g., a description of how Mr. Science put something yucky into the child's mouth). Interviews for Session 2 consisted of four parts: rapport building, open-ended questions (e.g., "Just tell me what you saw or heard or did the time you played in the science room with Mr. Science"), leading questions (e.g., "Did Mr. Science make a balloon blow up?"), and source-monitoring questions that asked the children to report which events actually happened and which events were described in the book. Session 3 took place 1 month later and consisted of the same procedures as those in Session 2, thereby providing data on the stability of accurate and inaccurate reports following a delay with no further exposure to misinformation. The narrative responses that were rescored for this paper were answers elicited by the 5 open-ended questions in each of the three sessions (i.e., 15 open-ended responses per child) and narrative answers to the 10 leading questions in Sessions 2 and 3 (i.e., 20 leading question responses per child).

Coding

Syntactic unit coding. The transcripts were originally coded as SUs. As was described earlier, SUs are predefined syntactic (i.e., grammatical) units that vary in length, such as the subject of a sentence, a verb or verb phrase, a descriptor, a prepositional phrase, or a direct or indirect quote (Poole & White, 1991). SU coding is implemented in four steps. Coders first cross off any information that will not be coded (e.g., requests for clarification, off-topic conversation, information implied by the question). Next, narratives are divided into SUs. Coders then read through and mark inaccurate SUs, and finally, accurate and inaccurate SUs are counted to calculate the proportion of information that is accurate.

Development of the modified word count method. The MWC procedure consists of six rules for deleting and modifying text. After scrubbing the narratives by applying these rules, the remaining words are counted, and these word counts are the unit of analysis for assessing autobiographical recall. (See the Appendix for the MWC coding manual. A list of suggestions for using word-processing conventions to implement the MWC procedure is available on request from the first author.)

Initially, four rules for coding were constructed and refined by coding a subset of the original 114 transcripts from Poole and Lindsay (in press). Briefly, (1) deletion of off-topic information excludes text that contributes no relevant data (e.g., off-topic comments); (2) deletion of conjunctions eliminates words that contribute no information by themselves and thus are not counted in the SU system (e.g., *but*, *and*, *because*); (3) deletion of false starts is an extension of off-topic deletion, but false starts are distinct in that they commonly appear at the beginning of a sentence and often are characterized by repetitive phrases and words (e.g., *ah*, *but*, *well*, *um*); and (4) repeated words and phrases are excluded, because redundant information should be counted only once.

Once these rules were refined, a research assistant familiar with the SU coding procedure and the authors attempted to apply the rules to sample transcripts. This prompted the addition of two rules regarding (5) yes/no and leading questions and (6) contracted words. Rule 5 specifies that responses such as "yes," "no," and "I don't know" are not scored as narrative information. (Yes vs. no responses are analyzed separately.) Contracted words are retyped as two words for consistency in scoring them as two words. In sum, the six rules were developed and modified through a repetitive process of coding, analysis of consistency between coders, and revision of the rules.

To implement these six rules, the MWC procedure uses editorial features commonly found in word-processing programs—namely, the *hidden text* and *automated-word-count* features. Narratives are first transcribed verbatim into a word-processing program (e.g., Microsoft Word). To apply Rules 1–5, research assistants then *delete* or *hide* text by using an editorial feature known as the *hidden text* function. This function is carried out by highlighting the desired text and clicking a shortcut key selected by the user to activate the hidden text feature. Hidden text is denoted by a dotted underline and can be concealed from the program's word count. Coders then apply Rule 6 to standardize the typing of specific words. Subsequently, the desired narrative is highlighted, and the word count function is activated to yield a word count for individual answers or blocks of answers.

After development of the coding system, two coders applied the system to a stratified sample (ages 3–8) of 10 transcripts. An interrater agreement of 93% was achieved for responses to open-ended questions ($n = 150$ questions). Similarly, an interrater agreement of 95% was achieved for responses to leading questions by coding a subset of 5 of these transcripts ($n = 100$ questions).

Raw word count values. To evaluate whether scrubbing with the MWC procedure is necessary, we also recorded RWC values, which were merely the number of words in each answer that described target events. Our first attempt to score RWC values detected inconsis-

tencies in how coders handled such comments as "I don't remember much about that." We therefore defined RWC values as all the words in an answer except off-topic comments (e.g., "Purple is one of my favorite colors") and statements that merely indicated the child had no knowledge of an event (e.g., "That's all I can say about that"). Thus, RWC values were all words that described target events, with no attempt to uncontract words or scrub out false starts, repeated phrases, or conjunctions. The interrater agreements were 95% for answers to open-ended questions and 97% for answers to leading questions.

The differences among coding systems. To illustrate the differences among coding systems, consider a sample answer from our data set. In response to the probe, "Tell me all the things you heard in the science room," one 8-year-old child responded, "And then we, he asked me how, no, let me think, he asks me the next thing we played, it was how to blow up balloons."

There are nine underlined SUs in this example: "then (temporal terms count as one SU) he (subject of sentence is one SU, although 'we' would count as two) asks (verb or verb phrase) me the next thing we played (a quote, coded as though the child said, 'He asked me, 'what is this next thing?')", it was how to blow up balloons." MWC coding eliminates conjunctions and uninformative phrases, including false starts and repeated words. Thus, 16 MWC values are counted: "then he asks me the next thing we played, it was how to blow up balloons." RWC values do not exclude conjunctions, false starts, or repeated words (i.e., 22 words are counted: "And then we, he asked me how, he asks me the next thing we played, it was how to blow up balloons"). Therefore, the RWC system merely scrubs out statements that do not describe target events. The MWC system also does this, but it then standardizes the typing of contractions and eliminates repeated words, phrases, and conjunctions. The SU system includes MWC rules but goes further by eliminating information that is implied by the question, inserting implied subjects into compound sentences, and parsing information on the basis of a list of grammatical units. All of these systems require coders to use their own judgment to code unusual or ambiguous items, but reliability is high for all the systems because such information is a small percentage of spoken words. Compared with the time it takes to complete RWC coding on a transcribed interview (about 1 min per transcript page), the MWC system increases coding time approximately 300%, whereas the SU system increases coding time approximately 500%.

When would RWC values yield different results than MWC or SU values? If verbal meandering is more common among older than among younger witnesses, failing to delete redundant information could affect the magnitude of age trends. Also, the proportion of information in a narrative that is accurate will be inflated if unnecessary words occur more often in descriptions of true rather than of suggested events. If superfluous words are a fixed proportion of words across ages and type of information reported, however, it is unnecessary to remove them from transcripts, because conclusions about differences across ages, interviewing conditions, or other independent variables will not be influenced by their presence.

Overview of the Data Analyses

Data set preparation. The sample of 48 transcripts, originally coded as SUs, were recoded with the MWC and RWC procedures. The process of coding information as accurate or inaccurate does not vary across systems, and interrater reliabilities were high for assigning information to accuracy categories in the original data set. We therefore assigned words to accuracy categories according to the original coding of the transcripts: Words in SUs that were originally coded as accurate were marked accurate, and the remaining words were assigned to one of four error categories. Subsequently, the number of words in each category were counted for each question, and these values were substituted for SUs to produce three parallel data sets: the original data set from Poole and Lindsay (in press), which was based on SUs, and the data sets in which SU values were replaced with MWC and RWC values. These data sets

were then compared, to determine whether the selection of a coding system would result in different interpretations about memory performance across age, delay, or type of question.

Statistical analyses. For the first three sections of the results, MWC and RWC values were individually compared against SU values. First, Pearson correlation coefficients were computed to assess agreement between coding systems. Second, the transformations between SU and word count values were computed and compared across three age groups (3–4 years, 5–6 years, and 7–8 years). These analyses determined the average number of words per SU and indicated whether this number was constant across age. Next, 10 analyses that were of interest in the original investigation (Poole & Lindsay, in press) were recomputed for each of the three data sets, to explore whether conclusions about memory performance would vary as a function of coding system. Finally, the proportion of information that was inaccurate was compared across coding systems.

RESULTS

Because responses to open-ended questions are generally substantially longer than responses to specific questions (see, e.g., Lamb et al., 1996), comparisons between SU and word count values were computed separately for these two question types. For each coding system, four dependent measures were calculated for both open-ended (i.e., free-recall) and leading (i.e., specific) questions: total information, proportion of *detail* errors, proportion of *suggested* errors (i.e., errors from the misleading story), and proportion of *other* errors (e.g., off-topic information that would not be recognized as such by someone who did not know the target event). For leading questions only, the proportion of *wrong question* information was also computed (i.e., accurate descriptions of an event other than the one in question when it was not clear that the child was reporting about another event).

Correlations Between Coding System Values

For the first set of analyses, agreement between SU and word count values for responses to individual questions was assessed with Pearson correlation coefficients. (Preliminary analyses showed that correlations computed with participant as the unit of analysis were slightly higher than correlations with individual answers as the unit of analysis. For brevity, we report only the later analyses.) Because the presence of zero values for questions with no narrative response could exaggerate the degree of agreement between coding systems, correlations were computed both with and without zero values; we report only those without zero values, which tended to be slightly lower.

For responses to open-ended questions, agreement between SU and word count systems exceeded .90 for all variables (see Tables 1 and 2), even when computed within individual age groups. Agreement was also excellent for responses to leading questions, exceeding .89 for all variables (see Tables 3 and 4). Agreement between SU and RWC values was surprisingly high, never more than .05 lower than correlations between SU and MWC values.

Table 1
Correlations Between Syntactic Unit and Modified Word Count Values for Responses to Open-Ended Questions Across Five Dependent Variables (Zero Values Omitted From the Calculations)

Dependent Variable	Total Sample		3-4 Years		5-6 Years		7-8 Years	
	Correlation	<i>N</i>	Correlation	<i>n</i>	Correlation	<i>n</i>	Correlation	<i>n</i>
Total information	.97*	450	.99*	101	.96*	177	.98*	172
Proportion of detail errors	.90*	90	—	—	.90*	33	.95*	49
Proportion of coached information	.93*	35	—	—	—	—	—	—
Proportion of other errors	—	—	—	—	—	—	—	—

Note—Correlations were not calculated for cells with less than 30 observations. * $p < .01$.

Table 2
Correlations Between Syntactic Unit and Modified Word Count Values for Responses to Open-Ended Questions Across Five Dependent Variables (Zero Values Omitted From the Calculations)

Dependent Variable	Total Sample		3-4 Years		5-6 Years		7-8 Years	
	Correlation	<i>N</i>	Correlation	<i>n</i>	Correlation	<i>n</i>	Correlation	<i>n</i>
Total information	.97*	450	.97*	101	.97*	177	.97*	172
Proportion of detail errors	.90*	90	—	—	.90*	33	.90*	49
Proportion of coached information	.94*	35	—	—	—	—	—	—
Proportion of other errors	—	—	—	—	—	—	—	—

Note—Correlations were not calculated for cells with less than 30 observations. * $p < .01$.

Transforming Syntactic Units to Word Count Values

In order to compare data across studies, it is useful to know the average number of word count values per SU. Separate regressions of MWC and RWC values onto SUs showed that word count values could be expressed as multiplicative values of SUs (i.e., the presence of zero values forced the intercepts near zero). For this transcript set, there were approximately 1.75 modified words per SU and 2.00 raw words per SU.

To determine whether these transformations were constant across the ages tested, each dependent variable for open-ended and leading questions was expressed as the ratio of the two coding system values (i.e., MWC/SU values). Because only nonzero data could be analyzed, the data were combined across sessions to produce adequate sample sizes. Linear age trends for MWC values were as follows: The means for total information reported in response to open-ended questions for ages 3-4, 5-6, and 7-8, respectively, were 1.6, 1.8, and 1.9

Table 3
Correlations Between Syntactic Unit and Modified Word Count Values for Responses to Leading Questions Across Six Dependent Variables (Zero Values Omitted From the Calculations)

Dependent Variable	Total Sample		3-4 Years		5-6 Years		7-8 Years	
	Correlation	<i>N</i>	Correlation	<i>n</i>	Correlation	<i>n</i>	Correlation	<i>n</i>
Total information	.90*	587	.90*	167	.95*	253	.90*	167
Proportion of detail errors	.94*	57	—	—	—	—	—	—
Proportion of coached information	.94*	117	.95*	34	.93*	57	—	—
Proportion of other errors	.90*	62	—	—	—	—	—	—
Proportion of wrong questions	—	—	—	—	—	—	—	—

Note—Correlations were not calculated for cells with less than 30 observations. * $p < .01$.

Table 4
Correlations Between Syntactic Unit and Raw Word Count Values for Responses to Leading Questions Across Six Dependent Variables (Zero Values Omitted From the Calculations)

Dependent Variable	Total Sample		3-4 Years		5-6 Years		7-8 Years	
	Correlation	<i>N</i>	Correlation	<i>n</i>	Correlation	<i>n</i>	Correlation	<i>n</i>
Total information	.90*	587	.90*	167	.90*	253	.90*	167
Proportion of detail errors	.93*	57	—	—	—	—	—	—
Proportion of coached information	.93*	117	.93*	34	.93*	57	—	—
Proportion of other errors	.89*	62	—	—	—	—	—	—
Proportion of wrong questions	—	—	—	—	—	—	—	—

Note—Correlations were not calculated for cells with less than 30 observations. * $p < .01$.

[$F(1,43) = 10.57, p < .01$]; the means for total information reported in response to leading questions were 1.5, 1.7, 1.7 [$F(1,44) = 4.16, p < .05$]. Thus, there were slightly more words per SU among older than among younger children, indicating that word counts would increase the magnitude of age differences, relative to using SUs. Age trends were less pronounced for RWC values. Means for the three age groups for open-ended questions were 1.9, 2.1, and 2.2 [$F(1,43) = 6.69, p < .05$]. Means for the three age groups for leading questions were 1.7, 1.9, and 2.0 [$F(1,44) = 3.09, p = .09$].

The Practical Significance of Syntactic Unit Versus Word Count Coding

To further evaluate the implications of using word count values, we compared a set of 10 analyses of variance (ANOVAs) conducted on SU values—two linear age trend analyses and eight linear age trend $\times$ session analyses—with the corresponding 10 ANOVAs computed on word count values. These analyses yielded comparisons for 26 main effects and interactions that were of interest in the original investigation.

For SU versus MWC coding, decisions to reject the null hypothesis ($\alpha = .05$) were concordant for all analyses except one. For the proportion of suggested information reported in response to leading questions, the main effect of session just reached significance for SU values ($p < .05$), but not for MWC values ($p = .06$). All of the main effects of age and interactions involving age yielded comparable results across these two coding systems. For RWC values, decisions to reject were concordant for all 26 main effects and interactions. Effect sizes for the age trends also showed remarkable agreement among coding systems. For example, effect sizes for RWC and SU values differed by only .000–.019 for nine analyses and by .03 and for the remaining analysis. For total information in response to open-ended questions in Session 1, for example, the effect sizes were .34 for SUs and .31 for RWC; for leading questions in Sessions 2 and 3, these values were .01 and .02.

The Proportion of Inaccurate Information Across Coding Systems

None of the previous analyses directly compared the proportions of accurate information across coding systems. Intuitively, it seems that false starts and redundant information would be more common in reports of accurate information than in reports of suggested information, because witnesses might simply repeat suggested information verbatim. If this were the case, failing to scrub out superfluous information in word count procedures would inflate the proportion of accurate information, relative to SU procedures.

The proportions of detail, other, and suggested information in responses to open-ended questions were entered separately into age trend (3–4, 5–6, and 7–8) $\times$ session (Sessions 2 vs. 3) $\times$ coding system ANOVAs, first comparing SUs against MWC values, then compar-

ing SUs against RWC values. As was expected, there were significant effects of coding system on the proportion of detail errors for both word count values ($ps < .001$). This is a consequence of the fact that detail errors tend to be single words but there are more accurate words than accurate SUs. Nonetheless, the absolute difference between systems was trivial: .016 of SUs were detailed errors, as compared with rates of .011 and .009 for MWC and RWC values, respectively.

The influence of coding system on other errors, which tend to be entire phrases, was less pronounced. The main effect of SUs versus MWC values did not reach significance ($p = .052$), whereas the main effect of SUs versus RWC barely did ($p = .050$). Again, the absolute difference between systems was small, with intrusion errors consisting of .007, .004, and .004 of SU, MWC, and RWC values, respectively.

Coding system did not alter estimates of the proportion of total information that described suggested events. Error rates were .054 and .051 for SU and MWC values, respectively ($p = .25$); for SU and RWC values, error rates were .054 and .049 ($p = .08$). There were no significant interactions involving coding for any of the aforementioned analyses. Therefore, although word count systems credit older children with mentioning proportionately more information than they credit younger children relative to SU coding, this had no major implications for measuring age trends in the accuracy of witness narratives.

DISCUSSION

The purpose of this project was to test new ways of coding eyewitness narratives that would result in more efficient data preparation than is provided by current UOI procedures, but without sacrificing the ability to code all of the information mentioned by eyewitnesses. Pearson correlation coefficients that assessed agreement between UOI and two word count methods showed strong congruence between coding procedures. Word count procedures credited older children with slightly more words per SU than they credited younger children, a result that might be partially explained by the fact that SU coding counts phrases that occur more frequently in elaborated narratives as single units (e.g., quotes and prepositional phrases). However, ANOVAs conducted on a number of dependent measures revealed that patterns of statistical significance between SU and MWC results were comparable for 25 out of 26 main effects and interactions. All 26 results were comparable between SU and RWC values.

These results document excellent agreement between coding methodologies. Although a larger sample size might have detected more discrepancies between SU and word count procedures, our sample sizes were typical for eyewitness research involving children (e.g., Goodman, Hirschman, Hepps, & Rudy, 1991; Memon, Wark, Bull, & Koehnken, 1997). Unfortunately, the data set we reanalyzed did not include narratives from an adult sample, and therefore further research is necessary to ensure

that comparisons between adults and children are not influenced by the choice of coding system.

Our most surprising finding was that correlations between SU and RWC values were almost as high as correlations between SU and MWC values, and all the systems yielded similar estimates of the proportion of suggested information that was inaccurate. It is counterintuitive that raw word counts would perform this way, because these counts include false starts and redundant information, which are commonly believed to be more characteristic of true versus false events. For example, criterion based content analysis is a system that attempts to distinguish between true and false events on the basis of these and other narrative features (e.g., Raskin & Esplin, 1991). Recently, though, many investigators have argued that children's reports of suggested events contain many of the features that are commonly associated with accurate testimony, making differences between accurate and inaccurate testimony too small to permit accurate diagnosis (e.g., Bruck, Hembrooke, & Ceci, 1997; Lamb et al., 1997). Our data therefore join a growing body of literature that demonstrates how difficult it is to distinguish between real and implanted memories, at least in young children's testimonies.

Investigators who use a word count procedure will enjoy several advantages over existing UOI coding procedures. Even the more complicated MWC protocol is straightforward, involving deleting (i.e., hiding) and re-typing text on the basis of only six rules that can be summarized on a one-page reference chart (see Table A1 in the Appendix). As a result, research assistants are spared having to learn complex rules based on grammatical categories. Coding is further simplified by taking advantage of computer word processing features. For example, human error can be reduced by using search and replace functions to code frequently appearing words, and the automated word count feature quickly tallies units of information and dramatically reduces coding time. Furthermore, by having assistants work directly on the computer, rather than by marking on hard copies of transcripts, fully coded transcripts can be printed as needed and are immediately available for archiving.

Word count procedures have limitations, though, and obviously the research question at hand should dictate which coding procedure is selected. For example, when target events can be outlined in terms of specific features, FOE coding may be more desirable, because the consistency of reports can be assessed more easily. In contrast, investigations of naturally occurring events may require the versatility of word count systems, which are capable of measuring information even when there is no record of the target event. Although RWCs performed well in our analyses, we suspect that some investigators will nonetheless prefer to use MWC values, because the cleaner transcripts from this system might be less likely to provoke criticism about the choice of coding method.

To conclude, word count procedures may not be optimal coding systems for all investigations of children's recall of events. Nonetheless, their high interrater reliability,

procedural simplicity, and the robust convergence with the SU procedure make them an efficient alternative to current UOI coding.

REFERENCES

- BAHRICK, L. E., PARKER, J. F., FIVUSH, R., & LEVITT, M. (1998). The effects of stress on young children's memory for a natural disaster. *Journal of Experimental Psychology: Applied*, **4**, 308-331.
- BAKER-WARD, L., GORDON, B. N., ORNSTEIN, P. A., LARUS, D. M., & CLUBB, P. A. (1993). Young children's long-term retention of a pediatric examination. *Child Development*, **64**, 1519-1533.
- BAKER-WARD, L., ORNSTEIN, P. A., GORDON, B. N., FOLLMER, A., & CLUBB, P. A. (1995). How shall a thing be coded? Implications for the use of alternative procedures for scoring children's verbal reports. In M. S. Zaragoza, J. R. Graham, G. C. N. Hall, R. Hirschman, & Y. Ben-Porath (Eds.), *Memory and testimony in the child witness* (pp. 61-85). Thousand Oaks, CA: Sage.
- BERLINER, L., & CONTE, J. R. (1993). Sexual abuse evaluations: Conceptual and empirical obstacles. *Child Abuse & Neglect*, **17**, 111-125.
- BRUCK, M., HEMBROOKE, H., & CECI, S. (1997). Children's reports of pleasant and unpleasant events. In J. D. Read & D. S. Lindsay (Eds.), *Recollections of trauma: Scientific evidence and clinical practice* (pp. 199-231). New York: Plenum.
- CECI, S. J., & BRUCK, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony*. Washington, DC: American Psychological Association.
- FISHER, R. P., GEISELMAN, R. E., & AMADOR, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witnesses of crime. *Journal of Applied Psychology*, **74**, 722-727.
- GEORGE, R. (1991). *A field and experimental evaluation of three methods of interviewing witnesses/victims of crime*. Unpublished manuscript, Polytechnic of East London, London.
- GOODMAN, G. S., HIRSCHMAN, J. E., HEPPS, D., & RUDY, L. (1991). Children's memory for stressful events. *Merrill-Palmer Quarterly*, **37**, 109-158.
- GORDON, B. N., ORNSTEIN, P. A., NIDA, R. E., FOLLMER, A., CRENSHAW, M. C., & ALBERT, G. (1993). Does the use of dolls facilitate children's memory of visits to the doctor? *Applied Cognitive Psychology*, **7**, 459-474.
- GREENSTOCK, J., & PIPE, M.-E. (1996). Interviewing children about past events: The influence of peer support and misleading questions. *Child Abuse & Neglect*, **20**, 69-80.
- LAMB, M. E. (1994). The investigation of child sexual abuse: An interdisciplinary consensus statement. *Child Abuse & Neglect*, **18**, 1021-1028.
- LAMB, M. E., HERSHKOWITZ, I., STERNBERG, K. J., ESPLIN, P. W., HOVAV, M., MANOR, T., & YUDILEVITCH, L. (1996). Effects of investigative utterance types on Israeli children's responses. *International Journal on Behavioral Development*, **19**, 627-637.
- LAMB, M. E., STERNBERG, K. J., ESPLIN, P. W., HERSHKOWITZ, I., ORBACH, Y., & HOVAV, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse & Neglect*, **21**, 255-264.
- MEMON, A., HOLLEY, A., WARK, L., BULL, R., & KOHNKEN, G. (1996). Reducing suggestibility in child witness interviews. *Applied Cognitive Psychology*, **10**, 503-518.
- MEMON, A., WARK, L., BULL, R., & KOHNKEN, G. (1997). Isolating the effects of the cognitive interview techniques. *British Journal of Psychology*, **88**, 179-197.
- POOLE, D. A., & LAMB, M. E. (1998). *Investigative interviews of children: A guide for helping professionals*. Washington, DC: American Psychological Association.
- POOLE, D. A., & LINDSAY, D. S. (1995). Interviewing preschoolers: Effects of nonsuggestive techniques, parental coaching, and leading questions on reports of nonexperienced events. *Journal of Experimental Child Psychology*, **60**, 129-154.
- POOLE, D. A., & LINDSAY, D. S. (1998). Assessing the accuracy of young children's reports: Lessons from the investigation of child sexual abuse. *Applied & Preventative Psychology*, **7**, 1-26.
- POOLE, D. A., & LINDSAY, D. S. (in press). Misinformation from par-

- ents' and children's eyewitness testimony. *Journal of Experimental Psychology: Applied*.
- POOLE, D. A., & WHITE, L. T. (1991). Effects of question repetition on the eyewitness testimony of children and adults. *Developmental Psychology*, *27*, 975-986.
- POOLE, D. A., & WHITE, L. T. (1993). Two years later: Effects of question repetition and retention interval on the eyewitness testimony of children and adults. *Developmental Psychology*, *29*, 844-853.
- RASKIN, D. C., & ESPLIN, P. W. (1991). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 153-164). Washington, DC: American Psychological Association.
- SAYWITZ, K. J., GOODMAN, G. S., NICHOLAS, E., & MOAN, S. F. (1991). Children's memories of a physical examination involving genital touch: Implications for reports of child sexual abuse. *Journal of Consulting & Clinical Psychology*, *59*, 682-691.
- TOBEY, A. E., & GOODMAN, G. S. (1992). Children's eyewitness memory: Effects of participation and forensic context. *Child Abuse & Neglect*, *16*, 779-796.
- YUILLE, J. C., & CUTSHALL, J. L. (1986). A case study of eyewitness memory of a crime. *Journal of Applied Psychology*, *71*, 291-301.

APPENDIX

The Modified Word Count Coding Manual and Quick Reference Guide

The following are six rules for preparing eyewitness narratives for coding. Five rules specify how to delete text that contributes no relevant information from the transcripts; one rule deals with typing conventions that will produce consistency across transcripts.

Eyewitness reports are transcribed verbatim and stored using a word-processing program. These programs offer a convenient feature that allows the transcriber to block and hide text that will not be scored, although the hidden text is retained in the file and can be printed when needed. In Word, for example, hidden text is marked on the screen by a dotted underline; in this coding manual, hidden text is underlined. Whenever transcribers mark hidden text on printed transcripts, however, deletion is represented by striking out words with a straight line. Phrases typed in parentheses in transcripts are explanatory

comments left by the interviewers or transcribers, and such comments are always deleted.

1. The first rule of deletion involves *off-topic information*. Text that contributes no relevant data is classified as off topic and should be excluded from coding. This includes questions, observations, comments, attempts at conversation, and otherwise random statements made by the speaker. In addition, statements such as "I don't know," "I'm not sure," "I think," and "That's all I know" should be deleted. Similarly, phrases that add no additional information, such as "or something" and "or whatever" are deleted. However, phrases such as "like," "kinda like," and "kind of" are legitimate descriptors of events and are coded.

And then we made paper airplanes and I got to color one. I colored mine yellow 'cause when I went to a beach I saw a yellow airplane. And anyways, yellow is my favorite color.

The first thing I think we did in this book about Mr. S . . . , in Mr. Science's room we, do they have to be in order? We played with the pulley.

He put a balloon on top and you squeezed it, and it popped. That's all. That's all I know. I think.

2. Deletion of *conjunctions* (i.e., "and," "but," "because") serves three functions: It transforms run-on sentences into smooth, legible passages, it increases the consistency of word counts between younger and older witnesses (because younger witnesses use fewer compound sentences than older witnesses do), and it eliminates words that contribute no information by themselves (e.g., "It was red and blue"). At first inspection, deletion of conjunctions may seem trivial, but the number of excessive words they contribute may be substantial.

We had a bottle with water and an eyedropper in it and we pressed it and it went down. We had tubes and we talked in them and we could hear each other.

First Mr. Science spun two and then I spun two. And then he showed me these funny glasses and I had to wear them and I spun tops with them and everything was that way, and if I looked this way, everything was that way.

The conjunction "or" is used in several different ways. Sometimes it indicates two possibilities or a range, as in "It was blue

Table A1
Quick Reference Guide: Rules for Preparing Eyewitness Narratives for Coding

Rule	Description	Example
1. Off-topic information	Text that contributes no relevant data. This includes comments such as "or something" and "or whatever."	There was this black coat, <u>and that's all. Any questions?</u>
2. Conjunctions	Delete the words "and," "but," and "because." Delete "or" and collapse across possibilities.	Then he said it was time to go <u>and</u> I had to go into his room.
3. False starts	The speaker starts, stops, and begins again. Data are coded where the speaker begins a complete sentence.	One of those things you open up it is like a, <u>it is like a, it is like</u> metal.
4. Repeated words and phrases	The speaker repeats words or phrases. These data are coded only once.	We saw a paperclip, a can. <u>A paperclip, a can, a magnet.</u>
5. Yes-no and leading questions	Do not count "yes," "no," or "I don't know" answers, even in the form, "I don't know if . . ." For questions such as "What did he do?" the word "nothing" is coded.	<u>Yes.</u> He pushed my tummy. <u>I do not think</u> he pushed my tummy. [Code as "he did not push my tummy."]
6. Contractions	Retype as two words.	You pull it up to see which <u>one's</u> [one is] the lightest.

APPENDIX (Continued)

or green" or "set it for 6 or 10 minutes." In such cases, delete the "or" and collapse across the two possibilities, so that the possibilities count as one word or one unit of information ("blue/green," "6-10 minutes"). In other cases, "or" may be a false start, as when the witness corrects himself or herself (e.g., "I heard my mom, or, some people talking"). In such cases, the "or" is also deleted, and the information is scored as a false start.

3. *False starts* usually are recognized by their repetitive nature. A sentence often starts with a few descriptive words, only to be challenged by the beginning of a new sentence. The speaker may start a sentence, stop, and begin a new sentence, often repeating the process several times. It is apparent that the speaker is stopping to reorganize words in order to better communicate the information.

False starts may also take the form of words such as "well," "but," "and," "um," or "ah." These words often occur at the beginning of a sentence and may serve a stalling function, giving the speaker more time to consider his or her answer. False starts should be deleted so that the beginning of a sentence or phrase is scored only once, regardless of how many times it may stop and begin again. For example,

Well what I did, what I did with that man right now, well what we did was we he would bring out this little thing, it was a can and it was heavy I could not even carry it.

One-word false starts should be deleted as well. For example,

Okay, I heard people talking outside.

However, false starts are commonly embedded within a sentence and do not necessarily appear at the beginning.

First Mr. Science got out a, a big pop bottle filled with water and he put, there was an eyedropper in it with little, with water in it and I had to try to get it to go up and down.

4. The next rule of deletion is excluding *repeated words and phrases*. Throughout the text, the speaker may repeat words and phrases several times; however, this information is coded only once.

Mr. Science got a balloon and the balloon we poured something into it.

We saw a paperclip, a can, and a magnet. Paperclip, can, and a magnet.

The speaker in the latter example twice reveals that Mr. Science had a balloon. Therefore, one account is deleted and the other is scored.

5. There are special considerations for scoring *answers to yes-no questions and leading questions*. Do not score the "yes," "no," or "I don't know" answer as narrative information. In response to questions such as, "Did he push your tummy?" the responses "I don't remember if he pushed my tummy" and "I don't know if he pushed my tummy" contain no codeable words. However, the response, "I don't think he pushed my tummy" is coded as though the witness said, "I think he didn't push my tummy" by crossing off "I think." If a leading question such as "Tell me what he did to your tummy" is followed by the response, "Nothing," the "nothing" is counted as a word of narrative information (because the individual is stating that nothing like that happened).

6. Finally, *contractions* are composed of two words and thus represent two units of information. Therefore, contractions are retyped and scored as two words. These words are circled on printed transcripts to mark contractions that need to be retyped in their expanded form.

(Manuscript received November 9, 1999;
revision accepted for publication July 30, 2000.)