Journal of Experimental Psychology: Applied
2012, Vol. 18, No. 4, 390–403

1076-898X/12/$12.00 DOI: 10.1037/a0029801

Extracting the Truth From Conflicting Eyewitness Reports: A Formal Modeling Approach

Berenike Waubert de Puiseau
Max Planck Institute for Research on Collective Goods, Bonn, Germany, and University of Mannheim

André Aßfalg and Edgar Erdfelder
University of Mannheim

Daniel M. Bernstein
Kwantlen Polytechnic University

Eyewitnesses often report details of the witnessed crime incorrectly. However, there is usually more than 1 eyewitness observing a crime scene. If this is the case, one approach to reconstruct the details of a crime more accurately is aggregating across individual reports. Although aggregation likely improves accuracy, the degree of improvement largely depends on the method of aggregation. The most straightforward method is the majority rule. This method ignores individual differences between eyewitnesses and selects the answer shared by most eyewitnesses as being correct. We employ an alternative method based on cultural consensus theory (CCT) that accounts for differences in the eyewitnesses' knowledge. To test the validity of this approach, we showed 30 students 1 of 2 versions of a video depicting a heated quarrel between 2 people. The videos differed in the amount of information pertaining to the critical event. Participants then answered questions about the critical event. Analyses based on CCT rendered highly accurate eyewitness competence estimates that mirrored the amount of information available in the video. Moreover, CCT estimates resulted in a more precise reconstruction of the video content than the majority rule did. This was true for group sizes ranging from 4 to 15 eyewitnesses, with the difference being more pronounced for larger groups. Thus, through simultaneous consideration of multiple witness statements, CCT provides a new approach to the assessment of eyewitness accuracy that outperforms standard methods of information aggregation.

Keywords: cultural consensus theory, general Condorcet model, eyewitness memory, eyewitness testimony, formal modeling

On November 22, 1963, John Fitzgerald Kennedy, then President of the United States, was assassinated during a motorcade through the city center of Dallas, Texas. Two days later, the only suspect, Lee Harvey Oswald, was assassinated by Jack Ruby, a local criminal. With Oswald's death, chances to clarify the true chain of events diminished greatly. Thus, the President's Commission on the Assassination of President Kennedy that investigated the murder had to rely almost exclusively on eyewitness reports to reconstruct the crime. After several months of investigation, the commission published the Warren Report, which stated that three shots were fired, all from a sixth floor window at the southeast corner of the Texas School Book Depository (President's Commission on the Assassination of President Kennedy, 1964).

Due to lack of agreement among the eyewitnesses, the Warren Report based some of its conclusions on conflicting eyewitness accounts. Indeed, large discrepancies emerged among the 552 witness with respect to descriptions of the number of shots fired and the location from which the shots were fired: "The consensus among the witnesses at the scene was that three shots were fired. However, some heard only two shots, whereas others testified that they heard four and perhaps as many as five or six shots" (President's Commission on the Assassination of President Kenney, 1964, p. 110).

Although the previous example describes a very rare situation in which hundreds of eyewitnesses were present during the assassination, it also epitomizes one of the problems that the criminal justice system faces when eyewitnesses testify, namely, the fallibility of memory: "If a hundred people were to see the same automobile accident, no two reports would be identical" (E. F. Loftus, 1996, p. 153).

One major reason for such discrepancies is the unreliability of eyewitness accounts. Inspired by Münsterberg's (1908) pioneering

This article was published Online First October 22, 2012.

Berenike Waubert de Puiseau, Max Planck Institute for Research on Collective Goods, Bonn, Germany, and Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany; André Aßfalg and Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim; Daniel M. Bernstein, Department of Psychology, Kwantlen Polytechnic University, Surrey, British Columbia, Canada.

The authors are grateful to Wolfgang Degen from the Wiesbadener Kurier for providing the video materials and to George Karabatsos for providing the MCMC analysis software. We thank William H. Batchelder and two anonymous reviewers for their valuable comments on a previous draft of this article.

Correspondence concerning this article should be addressed to Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim, 68131 Mannheim, Germany. E-mail: erdfelder@psychologie.uni-mannheim.de

work, Loftus' seminal experiments (i.e., E. F. Loftus, 1975; E. F. Loftus & Palmer, 1974) spawned hundreds of studies demonstrating the inability of eyewitnesses to reproduce observations reliably (for an overview, see Frenda, Nichols, & E. F. Loftus, 2011; Pansky, Koriat, & Goldsmith, 2005). However, in contrast to the literature that indicates poor memory performance, the general public believes eyewitness memory to be accurate, thus leading people to accept eyewitness testimony (Schmechel, O'Toole, Easterly, & E. F. Loftus, 2006; Simons & Chabris, 2011; Wells, Memon, & Penrod, 2006; Wise & Safer, 2004). In a recent study, Simons and Chabris (2011) investigated the public's knowledge of how memory works. In telephone interviews, 1,500 respondents from a representative sample of the general American public indicated their agreement with six statements on memory functioning. Their responses deviated strongly from those of experts on memory research and only 1.5% of participants in the representative sample correctly disagreed with all statements. In fact, 63.0% of the representative sample believed that "human memory works like a video camera," 47.6% thought that a memory formed of an event does not change, and 37.1% of the participants agreed with the statement that "the testimony of one confident eyewitness should be enough evidence to convict a defendant of a crime" (Simons and Chabris, 2011, p. 5). Psychological knowledge (i.e., the number of psychology books read and the number of psychology classes taken) was positively correlated with the proportion of correct answers, indicating that education about memory improves understanding of its functioning. Similarly, Schmechel et al. (2006) surveyed a sample of 1,007 juror-eligible U.S. citizens to explore their knowledge about eyewitness evidence. After asking 20 questions about factors that influence eyewitness reliability, the authors found that a large proportion of potential jurors hold misconceptions about eyewitness evidence, including the functioning of memory and factors influencing eyewitness reliability.

This combination of eyewitness unreliability and perceived high accuracy is dangerous and may lead, in its most extreme form, to false convictions (Dripps, 1999). To curtail the damage and to objectify legal decision making, the Innocence Project, an "organization dedicated to exonerating wrongfully convicted people,"[1] was founded, making use of the scientific improvements on forensic evidence. Across the 292 postconviction DNA exonerations in the United States resulting from the project as of June 20, 2012, eyewitness misidentification testimony was a factor in 72% of the cases, making it the leading cause of these wrongful convictions.[2]

Past attempts to improve witness memory are promising but fraught with problems. The cognitive interview (Geiselman, Fisher, MacKinnon, & Holland, 1985), for example, relies on context reinstatement: The more cues that match between encoding and retrieval contexts, the more information one remembers (Dando, Wilcock, & Milne, 2009; Milne & Bull, 2002; Morris, Bransford, & Franks, 1977; Tulving & Thomson, 1973). A meta-analysis of 42 studies revealed a clear superiority of the cognitive interview over standard procedures (Köhnken, Milne, Memon, & Bull, 1999). Today, the cognitive interview is a well-established interviewing procedure (Fisher, Milne, & Bull, 2011). Yet even though eyewitnesses report more correct details, it remains unclear which of the potentially conflicting pieces of information obtained through the cognitive interview are true and which are false (see also Bernstein & E. F. Loftus, 2009). The assessment of statement reliability is based solely on the information provided by the eyewitnesses.

In principle, this problem could be remedied by making use of reported confidence to distinguish between presumably competent and incompetent eyewitnesses, assuming that the former are more likely to provide accurate and comprehensive testimonies and therefore hold higher levels of confidence. The confidence-accuracy link has been investigated extensively in the psychology and law arena (e.g., Brewer, & Wells, 2006; Roebers, 2002; Sauer, Brewer, Zweck, & Weber, 2010; Sporer, Penrod, Read, & Cutler, 1995). Despite improvements in the investigation of links between eyewitness confidence and accuracy (Brewer & Wells, 2006; Weber & Brewer, 2004), some issues remain unresolved (Brewer & Weber, 2008; Brewer & Wells, 2011; R. C. L. Lindsay, Wells, & Rumpel, 1981). Some studies suggest that confidence ratings are more informative when provided immediately after an eyewitness's testimony or identification, instead of after some retention interval. However, the very same studies also suggest that these ratings may also be fallible (Brewer & Weber, 2008). Overall, the quality of confidence ratings as an index of accuracy depends on contextual cues, some of which have been identified (Brewer & Wells, 2006; Brewer & Weber, 2008; D. S. Lindsay, Read, & Sharma, 1998). Furthermore, most studies explore the confidence-accuracy relationship for identification only, whereas research investigating this link for event recall or recognition is scarce (e.g., Hollins & Perfect, 1997; Roebers, 2002; Smith, Kassin, & Ellsworth, 1989). Thus, independent cues to the quality of eyewitness accounts would be "extremely valuable" (Brewer & Weber, 2008, p. 827).

The presence of several eyewitnesses may ameliorate these troubles. In a sample of 773 Australian students, three-quarters had previously been eyewitnesses to a crime (Paterson & Kemp, 2006). The number of cowitnesses varied between 0 and 100, with a median of 3. Despite this knowledge, researchers have focused primarily on single testimonies without considering response patterns of multiple eyewitnesses.

To overcome these problems and limitations, we introduce and evaluate a model-based procedure that assesses the accuracy of several eyewitnesses simultaneously. According to this model, *eyewitness accuracy* (i.e., the actual proportion of correct responses) depends on several parameters, the most important being *eyewitness competence*, a latent variable measuring the actual knowledge about a crime, as influenced by situational and cognitive factors (e.g., perception, attention, memory). The approach is based on the assumption that high eyewitness competence results in congruence between eyewitnesses' testimonies, whereas guessing due to absence of knowledge results in stochastically independent responses (i.e., congruence at chance level). Conversely, the more congruent eyewitness reports are across items, the more likely it is that these responses are based on actual knowledge (i.e., high eyewitness competence). Obviously, this assumption is suspect when factors other than knowledge produce congruence between erroneous responses, such as schema influences, cowitness

[1] More information on the Innocence Project is available at http://www.innocence-project.org.

[2] See http://www.innocenceproject.org/Content/Facts_on_PostConviction_DNA_Exonerations.php.

talk, effects of leading questions, or exposure to misinformation. Hence, to apply the method properly, one must design interview questions that minimize schema influences and other biasing effects on responding. Moreover, the interviewer must ensure that eyewitnesses testify independently, that is, no witness should know about other witnesses' responses to the same event. Ideally, no witness should have been exposed to reports of the critical event (e.g., in newspapers, radio, or TV) before the interview.

Based on the underlying model, competence parameters are estimated for each eyewitness from their response patterns across items. Roughly speaking, these competence estimates are then used to aggregate the eyewitnesses' reports, thereby giving more weight to reports of more competent eyewitnesses than to reports of less competent eyewitnesses. Thus, based on a set of conflicting eyewitness responses to dichotomous questions addressing the same target event (e.g., a crime), the ultimate goal is the identification of the response pattern that describes the target event truthfully. This is achieved not only by examining how frequently a detail was reported but also by weighting each piece of information with respect to the competence of the person who reported it.

We organize this article in four sections. In the first section, we describe a simple model to evaluate eyewitness competence when one knows the true answers to a set of crime-related questions a priori—a rare situation in practice. This model is known as the two-high threshold model (2-HTM) of recognition (e.g., Snodgrass & Corwin, 1988). In the second section, we generalize the 2-HTM to the more realistic and interesting scenario in which one does not know the true answers a priori. Here, one must estimate both the eyewitnesses' competences and the answer key based on the eyewitnesses' reports. Fortunately, we can use an extant method to achieve this goal, namely, consensus analysis. Consensus analysis was developed in the context of cultural consensus theory (CCT; Romney, Weller, & Batchelder, 1986). As we will show, although originally designed for anthropological research, consensus analysis addresses problems that are structurally and conceptually isomorphic to problems in the field of eyewitness testimony. In the third section, we describe an eyewitness recognition memory experiment designed to evaluate consensus analysis in the context of eyewitness testimony. Finally, in the Discussion section, we summarize the conclusions that can be drawn from the results and their implications for using consensus analysis in legal contexts. Because consensus analysis can be applied not only to eyewitness reports but also to any kind of witness report, we use the terms "eyewitness" and "witness" interchangeably.

The Two-High Threshold Model (2-HTM)

Consider a typical recognition memory experiment. Participants learn a list of items (e.g., target words) and later receive a yes–no recognition test including both the target items and new distractor items randomly arranged. Participant i, $i = 1, \ldots, N$, should respond "yes" to each target item and "no" to each distractor item. Performance is usually measured in terms of some function of hits, $H_i = P(\text{"yes"} \mid \text{target})_i$, and false alarms, $F_i = P(\text{"yes"} \mid \text{distractor})_i$, such as $H_i - F_i$ or $d'_i = z(H_i) - z(F_i)$. This recognition memory paradigm resembles a witness recognition test in which the witness should respond "yes" to statements describing a critical target event correctly and "no" to incorrect statements about this event. If investigators know the actual truth values of these statements a priori, they can measure witness performance in terms of hits, $H_i = P(\text{"yes"} \mid \textit{true})_i$, and false alarms, $F_i = P(\text{"yes"} \mid \textit{false})_i$, just like in a recognition memory experiment.

The 2-HTM (Snodgrass & Corwin, 1988) explains hits and false alarm rates in terms of the joint influence of (a) the ith participant's discrimination competence between the two sets of stimuli, D_i, and (b) the ith participant's "yes-guessing" tendency in case of discrimination failure, g_i. Applied to witness judgments, a hit occurs when (a) an eyewitness correctly detects a true statement with probability D_i, or (b) fails to detect a true statement with probability $(1 - D_i)$ but, in addition, guesses correctly with probability g_i. Hence, the overall probability of a hit is $H_i = D_i + (1 - D_i)g_i$. Conversely, a false alarm occurs only if the witness fails to detect a false statement with probability $(1 - D_i)$ and incorrectly guesses "yes" with probability g_i. By implication, the model equation for false alarms is $F_i = (1 - D_i)g_i$.

By solving both 2-HTM model equations for the two parameters D_i and g_i, we can write these parameters as functions of H_i and F_i. Specifically, the probability D_i of witness i knowing the actual truth value of a statement—henceforth referred to as the *competence* of witness i—can be derived by calculating the difference between both equations,

$$D_i = H_i - F_i. \tag{1}$$

Analogously, the *response tendency* of witness i guessing "true" after failing to recognize the actual truth value can be derived as

$$g_i = \frac{F_i}{1 - H_i + F_i}. \tag{2}$$

By replacing H_i and F_i with observed relative frequencies of hits and false alarms (with the latter not exceeding the former), we can determine maximum likelihood estimates of the competence parameter D_i and the response tendency parameter g_i from Equations 1 and 2 for each witness i. Because the 2-HTM is well established and has been validated successfully in various contexts (e.g., Bröder & Schütz, 2009; Snodgrass & Corwin, 1988; see also Erdfelder et al., 2009; Erdfelder, Cüpper, Auer, & Undorf, 2007; Erdfelder, Küpper-Tetzel, & Mattern, 2011; Hilbig, 2012), it is an appropriate tool for assessing and evaluating witness competence and response tendency whenever one knows the answer key to a set of questions a priori. Even more important in the present context, one can generalize the 2-HTM to the more realistic scenario in which assessors lack knowledge of the answer key a priori, that is, when they do not know whether a certain response was correct or not. We address this more fundamental problem in the following section.

Consensus Analysis

CCT (Romney et al., 1986) was originally designed to explore the culture of unknown ethnic communities in anthropological research. In this section, we argue that CCT also provides an appropriate conceptual and methodological framework to assess witness competence and response tendency whenever the answer key to a set of crime-related questions is unknown a priori. Specifically, we hypothesize that a crime scene resembles a culture known only to those involved and the eyewitnesses, and, in particular, unknown to the witness interviewer. In standard anthropo-

logical research projects, anthropologists explore the shared knowledge that constitutes culture through systematic interviews of people from that community. Presumably, consensus among informants in this case is less than perfect because individual community members hold different levels of expertise about their culture (Romney, Batchelder, & Weller, 1987). Analogously, due to factors such as individual differences in memory, attention, motivation, or visual perspective, witnesses to crimes are not in perfect agreement with each other, either. Furthermore, both the anthropologist and the witness interviewer do not know the correct answers to the questions asked and they are unaware of the informants' knowledge about their culture or the crime, respectively.

To clarify the concept of the shared pool of information, we return to the President Kennedy assassination. The circumstances and sequence of events during the assassination constitute a pool of shared information. Individual eyewitness reports diverged from this shared pool to some degree. Consequently, the alleged truth had to be derived from data that were contaminated with individual errors. Visual perspective likely determined the amount and quality of information that witnesses observed. Additionally, individual differences in attention and memory strength also influenced the quality of memory.

In such cases where several people witnessed a crime, CCT may provide a solution and yield estimates of eyewitness competence and the correct responses that one may use to reconstruct the actual crime. Thus, the present experiment explores whether methods developed in the framework of CCT can improve the conclusions drawn from witness testimony.

The General Condorcet Model

For dichotomous items with true–false or yes–no answer options, CCT is formalized in terms of the general Condorcet model (GCM, Batchelder & Romney, 1986; Karabatsos & Batchelder, 2003; Romney et al., 1986). Other versions of the GCM and different formal models have been developed to accommodate different response formats (Batchelder, Kumbasar, & Boyd, 1997; Batchelder & Romney, 1988; Batchelder, Strashny, & Romney, 2010). However, we employ the GCM for dichotomous items, because it is the most extensively investigated formal specification of CCT (e.g., Shafto & Coley, 2003; Weller, Romney, & Orr, 1987).

Stemming from anthropological research, the GCM was designed to utilize patterns of congruence and incongruence in response data to detect what constitutes culture in unknown ethnic communities. The basic idea is that, because of their shared knowledge about the culture, highly competent informants will provide responses that tend to agree (and reflect the truth). Conversely, incompetent informants will essentially guess with respect to all questions and therefore generate uncorrelated responses. Thus, roughly speaking, the extent to which informants agree with each other mirrors the correlation of their responses with the truth. Hence, the probability of knowing the correct answer, the competence, can synonymously be interpreted as a function of shared knowledge (Weller, 1987). However, as the present work explores the possibility of applying the GCM to witness testimony, we will outline the basic principles of the GCM from that perspective. The GCM and the associated methods of statistical analysis have been described in detail elsewhere (cf. Aβfalg & Erdfelder, 2012; Batchelder & Romney, 1986, 1988, 1989; Karabatsos & Batchelder, 2003; Romney et al., 1986, 1987). We will thus illustrate only the basics, thereby relying on the notation previously introduced in the literature (cf. Karabatsos & Batchelder, 2003).

The GCM can be easily combined with the 2-HTM discussed in the previous section. Assume each witness independently replies to M statements about a critical event. An unknown number of these statements is actually true, whereas the others are false. Accordingly, a witness responds either with a "true" or "false" statement. Assuming N witnesses and M statements, then X_{ik} is the response of witness i to statement k, producing the response matrix $\mathbf{X} = (X_{ik})_{N\times M}$ with

$$X_{ik} = \begin{cases} 1, \text{ if witness } i \text{ answers "true" to statement } k \\ 0, \text{ if witness } i \text{ answers "false" to statement } k \end{cases}. \quad (3)$$

Furthermore, the unknown answer key $\mathbf{Z} = (Z_k)_{1\times M}$ is defined as

$$Z_k = \begin{cases} 1, \text{ if the correct response to statement } k \text{ is "true"} \\ 0, \text{ if the correct response to statement } k \text{ is "false"} \end{cases}. \quad (4)$$

The major assumptions underlying the GCM are

1. Common truth: There is a single answer key to a set of questions that applies to all witnesses. Thus, all witnesses refer to the same event.

2. Local independence:

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z}) = \prod_{i=1}^{N} \prod_{k=1}^{M} P(X_{ik} = x_{ik} \mid Z_k = z_k), \quad (5)$$

for all possible response matrices $\mathbf{x}$ and answer key vectors $\mathbf{z}$. As outlined previously, one implication of this assumption is that each witness must provide responses independently of all other witnesses. Thus, relationships between witnesses' answers are determined solely by the extent to which each witness's responses correlate with the true answer key (Romney, 1999), not, for example, by shared stereotypes, cowitness talk, or collaborative memory processes (Janssen, Kirschner, Erkens, Kirschner, & Paas, 2010).

There are two major differences between the GCM and the 2-HTM introduced in the previous section. First, the answer key $\mathbf{Z}$ is a latent variable in the context of the GCM, whereas it is a manifest (i.e., observable) variable in the context of the 2-HTM. Second, the most general version of the GCM, introduced by Karabatsos and Batchelder (2003), accounts for variance in item difficulties in addition to the participants' competences and response tendencies. Because the probability of correct discriminations may vary across items, Karabatsos and Batchelder (2003) propose a parameter D_{ik} that depends on both the participant i and the item k. In contrast, the 2-HTM, as outlined previously, assumes homogeneous item difficulties and therefore competence parameters D_i that depend on the participants only.

More precisely, according to Karabatsos and Batchelder's (2003) most general version of the GCM, the probability of witness i knowing the correct response to statement k is

$$D_{ik} = \frac{\theta_i(1-\delta_k)}{\theta_i(1-\delta_k)+(1-\theta_i)\delta_k}, \quad (6)$$

where δ_k is the item difficulty parameter of statement k and θ_i is the competence parameter of witness i with $0 < \delta_k, \theta_i < 1$. In contrast to D_{ik}, the response tendency parameter g_i and all other assumptions of the 2-HTM remain unchanged in Karabatsos and Batchelder's (2003) model.

Incidentally, as shown by Crowther, Batchelder, and Hu (1995), Equation 6 is equivalent to the well-known Rasch model (Fischer & Molenaar, 1995; Rasch, 1960) applied to the latent probability D_{ik} that witness i is correct on item k. Hence, just like the Rasch model, the GCM based on Equation 6 is not identified without an additional parameter constraint. To ensure identifiability, we therefore used the first item as a reference item and set $\delta_1 = .5$ in all our analyses (see Karabatsos & Batchelder, 2003).

Based on these GCM assumptions, the probability of a correct response of witness i to statement k can be derived as

$$p_{ik} = D_{ik}^{Z_k} + g_i(1 - D_{ik})(2Z_k - 1). \quad (7)$$

Note that for true statements (i.e., $Z_k = 1$), this model equation reduces to the 2-HTM model equation for hits, $H_{ik} = D_{ik} + (1 - D_{ik})g_i$, whereas for false items (i.e., $Z_k = 0$), it reduces to the complement of the 2-HTM model equation for false alarms, $1 - F_{ik} = 1 - (1 - D_{ik})g_i$. This clearly shows that the GCM can be seen as a generalization of the 2-HTM for the case of unknown answer keys.

Parameter Estimation and Model Selection

How does one obtain estimates of the GCM parameters θ_i (witness competence), g_i (witness guessing bias), δ_k (item difficulty), and Z_k (answer key)? Following Karabatsos and Batchelder (2003), we estimated all parameters simultaneously from the observed eyewitness response matrix $\mathbf{X}$ using a Bayesian procedure, the Markov Chain Monte Carlo (MCMC) method. Starting from Equation 7, the likelihood function of the observed $N \times M$ response matrix $\mathbf{X}$ given the full GCM parameter vector $\Omega = \langle\langle\theta_i\rangle_{i=1}^{N},\langle g_i\rangle_{i=1}^{N},\langle\delta_k\rangle_{k=1}^{M},\langle Z_k\rangle_{k=1}^{M}\rangle$ can be derived as

$$L(\mathbf{X}|\Omega) = \prod_{i=1}^{N}\prod_{k=1}^{M} p_{ik}^{Z_k X_{ik} + (1-Z_k)(1-X_{ik})} \times (1-p_{ik})^{Z_k(1-X_{ik}) + (1-Z_k)X_{ik}}. \quad (8)$$

The MCMC method provides estimates of the posterior distribution of Ω based on this likelihood function and the prior distribution of Ω. Again following Karabatsos and Batchelder (2003), we used uninformative priors in our analyses. In this case, the shape of the posterior distribution of Ω is solely determined by the likelihood function. In line with standard practice, we used the means of their estimated marginal posterior distributions as point estimates of the GCM parameters. The only exception is the discrete answer key for which we used the mode of the marginal posterior distribution as a point estimate.

Alternative estimation procedures for different variants of the GCM exist (see Aβfalg & Erdfelder, 2012; Batchelder & Romney, 1986, 1988, 1989; Romney et al., 1986, 1987). Among these, the factor analytical approach appears to be used most frequently. In contrast to the MCMC procedure, however, this method does not account for possible differences in response tendencies and item difficulties. Given these limitations of the factor analytical approach, we prefer the MCMC procedure. Instead of simply assuming that, for example, response tendencies do not influence the data, as the factor analytic approach does, the MCMC procedure provides for tests of this assumption. Use of appropriate model selection criteria prevents both model misfit and overfit of the data, thus suggesting the most parsimonious version of the GCM necessary to adequately describe the data. Model selection within the MCMC approach can be accomplished with the distance information criterion (DIC; Karabatsos & Batchelder, 2003). The DIC measure penalizes for the number of parameters and therefore helps to prevent overfitting (Spiegelhalter, Best, Carlin, & van der Linde, 2002). By computing DIC for all versions of the GCM and selecting the model version with the lowest DIC, we achieve a balance between model fit and model parsimony.

Eyewitness Recognition Experiment

We hypothesize that the GCM is a valid tool to aggregate testimonies of several eyewitnesses. In general, data aggregation can reduce idiosyncrasies such as false claims of single witnesses. Furthermore, the GCM accounts for competence differences between witnesses, thus acknowledging that witnesses may differ not only with regard to memory, attention, or motivation but also with regard to environmental factors such as visual perspective that influenced the eyewitnesses' knowledge. Enhancing the assessment of witness testimony not only contributes to the study of eyewitness memory in the lab but also may improve the assessment of witnesses in real-life situations.

Of course, the GCM will fail, just as any other method fails, if all eyewitnesses lack competence. In this case, the testimonies are essentially random and stochastically independent of each other. However, this extreme scenario is rather unlikely in practice. If at least some of the witnesses show moderate degrees of competence, there is a basis for estimating the true answer key. A more serious problem is that congruence between individuals may not be driven by competence but instead by common schemas, cowitness talk, or possibly motivated lying (Wells et al., 2006). Any systematic congruence that is not due to competence would violate the assumption of local independence and may therefore bias the GCM estimates. As outlined in the introduction to this article, the best way to address this problem is to design the interview questions and both the temporal and spatial context of the interview in such a way so as to minimize the risk of violations of independence. However, because it is difficult to eliminate such undesired influences entirely—especially outside the lab—robustness tests of the GCM methods under violations of independence would certainly be helpful. Note, however, that this desideratum is not specific to the GCM approach. Of course, any method of analyzing witness testimonies can be adversely affected by schema influences, cowitness talk, or lying.

In the present experiment, we presented a video of a critical event to our participants and subsequently asked them questions about the event. Because all event details can be determined based on the objective video content, the true answer key to all questions is known in this case. This enables an evaluation of the correctness of the answer key as estimated from the participants' responses using the GCM. Moreover, GCM estimates of participant competence and guessing bias can be compared with the corresponding 2-HTM estimates as derived from the true answer key.

We expected three outcomes if CCT is a viable theoretical approach for assessing eyewitness testimony. First, the GCM

estimates of eyewitness competences and response tendencies obtained by ignoring the answer key should closely resemble the 2-HTM estimates based on the true answer key. Thus, the GCM should uncover the true eyewitness competences and guessing tendencies, even though the correct answers are treated as unknown in the GCM. Second, the GCM's answer key estimates should outperform those of plausible alternative aggregation rules, in particular, the majority rule—the answer pattern provided by the majority of eyewitnesses irrespective of their competences. Third, the GCM's answer key estimates should be consistent. Thus, the more witnesses there are, the better the GCM answer key estimate approximates the true answer key.

We presented participants with one of two videos followed by an unexpected memory test. One of the videos showed only a few details of the critical event (*low information*), whereas the other provided the viewer with many details (*high information*). This manipulation was intended to induce competence differences in participants. Competence differences are important for two reasons. First, due to the students' similarity on cognitive characteristics in comparison with the overall population, the sample was possibly more homogeneous than actual groups of eyewitnesses. Heterogeneity was further likely to be reduced because laboratory studies commonly reduce individual variability as a consequence of standardization. Second, we predicted that the experimental manipulation of the participants' competences affects the competence estimates of the GCM.

Method

Sample

Thirty undergraduate students from the University of Mannheim participated as part of a study requirement. Of these participants, 73.3% ($n = 22$) were female. Participants were recruited during lectures and by means of posters on campus. They were randomly assigned to two video conditions. Twenty-seven of the participants majored in psychology (90.0%), two in sociology (6.7%), and one in a nonspecified subject. Ages ranged between 18 to 28 years ($M = 21.13$, $SD = 2.65$). There were no significant differences between the two video conditions on any of these demographic variables.

Design and Procedure

We presented the videos on 19-in. monitors using adjustable headphones. Partition walls separated the individual workspaces. Before the experiment started, participants provided informed consent and answered several demographic questions. As an incentive to perform as well as possible, participants were promised a shopping voucher worth €10 for the highest scores in the task. Given the different amounts of information provided in the two video versions, the participant with the highest score in each video condition received a voucher.

In the study phase of the experiment, we administered an incidental learning task by telling participants to watch a video and to look for a coffeehouse of a famous brand. Participants then watched one of two video versions. Half the participants watched the video in the high-information condition, the other half in the low-information condition. At the video's end, we asked participants to help the police by serving as a witness and by answering 110 dichotomous questions concerning the critical event shown in the video.

Participants completed a 2-min brainstorming task before answering the recognition questions. To enhance memory performance, we instructed participants to imagine themselves in the scene they had just observed. This builds upon the first mnemonic of the cognitive interview. It requires witnesses to mentally reinstate the environmental and personal context of the crime-scene observation (cf. Campos & Alonso-Quecuty, 1999; Geiselman, Fisher, MacKinnon, & Holland, 1986). Participants then completed the items at their own pace. The two response buttons appeared beside each other on the screen. Participants made their selection by clicking on one of these buttons. The duration of the experiment ranged between 20 and 40 min.

Materials

To create the two video versions, we used footage from two different sources. The first part of both videos shows the perspective of a pedestrian walking through the city center of a midsized German city and lasts about 7 min. In the second part of both videos, this person observes the critical event that lasts for about 2 min. In the third and final part, the person from whose perspective the video is shot leaves the scene of the critical event and, approximately 1 min later, reaches a café where the video ends. The research team filmed the first and the last parts of the video in the same downtown area where the critical event occurred. We took the scenes of the second part depicting the critical event from a project on moral courage sponsored by a local newspaper. The critical event consisted of two parts. It depicts a fight between a man—the perpetrator—and a woman—the victim—who seem to know each other. The man starts an argument by asking the woman to stop and talk to him. When she asks him to leave her alone, he pushes and grabs her. The woman then defends herself physically and verbally. Three bystanders stop and intervene by telling the man to calm down. For a few seconds, the camera moves away from the scene, but then returns to the man and the woman, who restart the fight. Again, a bystander intervenes. After the second part of the fight, the person from whose perspective the video is shot leaves the scene.

The two video versions differed only in terms of the visual perspective from which the critical event in the second part of the video was observed; the first and the last part were identical in both video versions. All recognition test items referred to the critical event only. The high-information video version showed the perpetrator and the victim from above and from the front. This video was recorded out of a window on the first floor of a building and provides the observer with a clear view of the scene. The two people in the video walk toward the camera; therefore, the observer can see the bystanders clearly. The observer can also perceive the colors of clothes and hair and the words spoken by the people involved. The low-information video version shows the crime scene from a perspective that resembles the viewpoint of somebody sitting in a café and shows the perpetrator and the victim from behind. Because the video was recorded by a video camera hidden in a bag, the view is fuzzy. The man and woman are partially hidden behind the other people walking on the street. Colors are hard to see, because the brightness of the video is reduced compared with the high-information video. Furthermore, only fragments of the sentences are recorded and understandable, resulting in less information obtained by watching the low-information video compared with the high-information video. Figure 1

Figure 1. Screenshots from the high-information (left) and low-information (right) video conditions, depicting a quarrel between a man and a woman from different viewing angles.

shows screen shots from one scene recorded from the two different angles.

Each of the 110 recognition test items addressed one of the following topics: crime scene (18 items; e.g., "Did both quarrels happen at the same crime scene?"), perpetrator (24 items; e.g., "Did the perpetrator wear glasses?"), victim (24 items; e.g., "Did the victim wear a scarf?"), course of action (20 items; e.g., "Did the perpetrator push the victim?"), or general information (24 items; e.g., "Did the bystander who stopped the second quarrel have a beard?"). Given the suboptimal perspective, some of the information necessary to answer the questions properly was barely perceivable in the low-information condition.

Results

We analyzed participant-by-item response data using the MCMC method of GCM consensus analysis described previously. George Karabatsos (Karabatsos & Batchelder, 2003) provided the analysis software, adapted to run in the R statistics environment (R Development Core Team, 2011). We used several evaluation criteria for the GCM. First, the competence and response tendency estimates of the 2-HTM based on the true answer key served as criteria to evaluate the competence and response tendency estimates of the GCM. Second, to evaluate the GCM's item difficulty estimates, we compared them with the actual item difficulties, that is, the proportion of correct responses per item. Third, we evaluated the GCM's answer key by comparing it with both the answer key estimate based on the majority rule and the true answer key. However, before applying these evaluation criteria, we first identified the version of the GCM that best accounted for our data.

Model Selection and Model Fit

Because we did not know a priori whether all of the GCM's parameters (witnesses' competences, witnesses' response tendencies, and item difficulties) would be necessary to describe the response data in our experiment adequately, we computed the DIC badness-of-fit criterion for four versions of the GCM (see Table 1). GCM3 is the most general version. GCM2*g*, in contrast, assumes homogeneous response tendencies, whereas GCM1 assumes both homogeneous response tendencies and homogeneous item difficulties. Finally, GCM2δ is based on the assumption of homogeneous item difficulties.

When we analyzed all 30 participants jointly, the most complex version of the GCM (GCM3) had the lowest DIC value, thus indicating the best balance of model fit and number of parameters.[3] Therefore, we performed all subsequent analyses based on the GCM3.

Prior to estimating the parameter values, we also assessed the model fit. For this purpose, we inspected the posterior predictive probability of the GCM3, sometimes referred to as the Bayesian *p* value (e.g., Carlin & Louis, 1996). Its interpretation is similar to that of the classic *p* value in frequentist model assessment (cf. Gelman, Meng, & Stern, 1996). The posterior predictive probability was .38 in our analyses, revealing no significant misfit.

Manipulation Check

To assess the success of the video information manipulation, we compared the proportion of correct responses based on the true answer key in the two information conditions. As expected, participants in the high-information condition produced significantly more correct responses ($M = 0.79$, $SD = 0.04$) than did participants in the low-information condition ($M = 0.71$, $SD = 0.04$), $t(28) = 5.25$, $p < .001$, $d = 1.92$. The overall proportion of correct responses was rather high, indicating that the items were fairly easy on average ($M = 0.75$, $SD = 0.25$), with difficulties ranging between 0.07 to 1.00.

Competence and Response Tendency Estimates

To evaluate the GCM parameters, we first computed the competence and response tendency estimates of the 2-HTM. Note that

[3] DIC values within two to three DIC units of the lowest value may be considered as viable alternatives (Karabatsos & Batchelder, 2003).

Table 1
DIC Values for Several Versions of the GCM (N = *30*)

GCM version	Parameters	DIC
GCM3	Competence, response tendency, item difficulty	2779.86
GCM2g	Competence, item difficulty	2781.32
GCM1	Competence	3248.90
GCM2δ	Competence, response tendency	3257.01

Note. Model versions are arranged by DIC value in ascending order. The lower the DIC value, the better the balance between model fit and model parsimony. DIC = Distance Information Criterion; GCM = general Condorcet model.

whenever we refer to 2-HTM estimates of witness competence and response tendency, we use the true answer key that can be determined objectively based on the content of the videos shown to our participants. Because the 2-HTM competence estimates correlate perfectly with the individual's proportion of correct answers (for an explanation, see the Appendix), the result of the *t*-test comparing mean competence estimates (see Table 2) in the high-information with those in the low-information condition was the same as in the manipulation check reported previously. As expected, response tendencies were unaffected by the video condition, $t(28) = 1.30$, $p = .21$, $d = 0.47$.

Having established that the video condition influenced both the proportion of correct responses per witness and the 2-HTM competence parameters as expected, we analyzed the effect of the video conditions on the GCM competence parameters. Note that whenever we refer to GCM parameter estimates subsequently, we ignore the true answer key and treat the true answers as an unknown latent variable instead. We performed model-based analyses simultaneously for all 30 participants to compute estimates for the GCM parameters.[4] The mean GCM competence estimates across participants (see Table 2) were significantly higher in the high-information than in the low-information condition, $t(28) = 6.47$, $p < .001$, $d = 2.36$. Surprisingly, the GCM response tendency estimates also differed significantly between video conditions, $t(28) = 2.76$, $p = .010$, $d = 0.90$. As expected, the competence estimates of the GCM and the 2-HTM correlated highly ($r = .84$, $p < .001$) across both conditions. As shown in Figure 2, competence was fairly homogeneous despite the attempt to increase heterogeneity. Similarly, 2-HTM and GCM response ten-

Table 2
Means (and Standard Errors) of the 2-HTM and the GCM Parameter Estimates by Video Condition

Video condition	Competence	Response tendency
2-HTM		
High information	0.57 (0.02)	0.58 (0.02)
Low information	0.42 (0.02)	0.61 (0.02)
GCM		
High information	0.76 (0.02)	0.48 (0.03)
Low information	0.59 (0.03)	0.59 (0.02)

Note. GCM = general Condorcet model; 2-HTM = two-high threshold model.

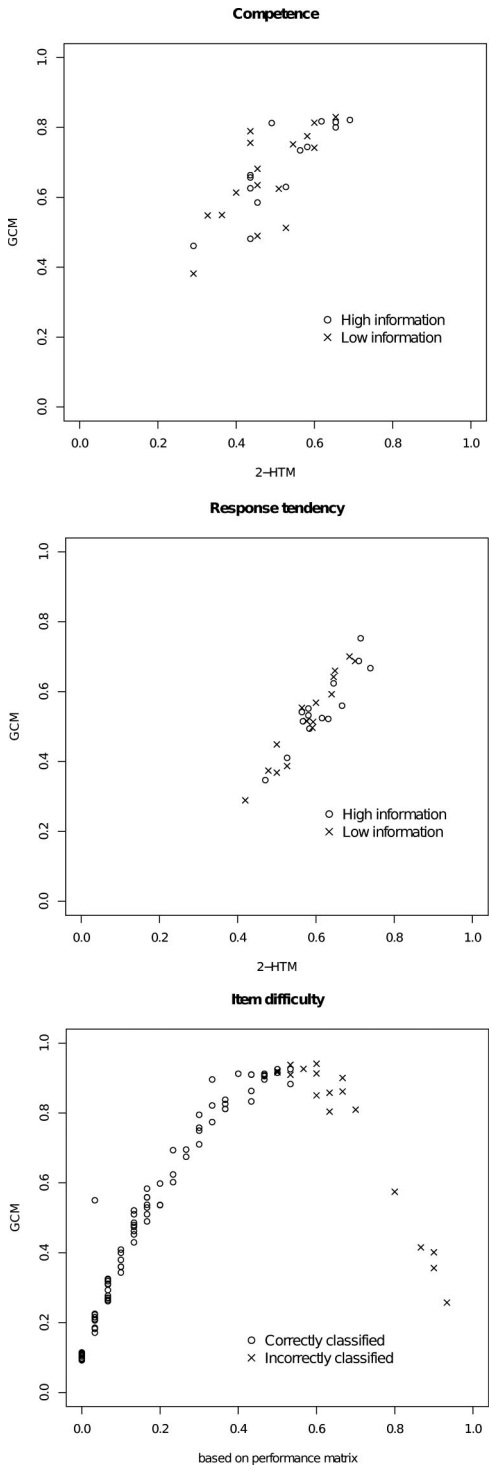

Figure 2. Relationship between GCM parameter estimates and estimates based on the true answer key by video condition/classification accuracy.

[4] The number of iterations employed by the MCMC algorithm was set to 10,000, including 1,000 burn-in trials (that are not included in the analyses but serve to overcome the influence of starting values) to achieve an acceptable trade-off between precision and computational effort.

dency estimates correlated significantly ($r = .94$, $p < .001$; Figure 2) across conditions.

Item Difficulty Estimates

We also compared item difficulty estimates based on the GCM with true item difficulties, as estimated from the proportion of correct responses per item. The comparison revealed a strong and statistically significant correlation ($r = .71$, $p < .001$). However, as can be seen in Figure 2, this relationship was quadratic in nature. The GCM estimated very difficult items as being easy. This estimation error probably results from the way item difficulties and answer keys are implemented in the GCM3. Assuming, for example, that the majority of witness responses for item k is $X_{ik} = 1$, then the true difficulty of item k would be high if $Z_k = 0$ (i.e., $X_{ik} = 1$ is incorrect), and it would be low if $Z_k = 1$ (i.e., $X_{ik} = 1$ is correct). However, if the GCM3 misestimates the answer key of item k because of sampling error (i.e., $\hat{Z}_k \neq Z_k$), a large estimation error for δ_k is implied and contributes to the quadratic relationship evident in Figure 2. This interpretation is corroborated by the fact that an exclusion of the misclassified items ($\hat{Z}_k \neq Z_k$) from the correlation analysis yields an almost perfect linear relationship between the item difficulty estimates and the true item difficulties ($r > .99$, $p < .001$).

Answer Key Estimates

Using the responses of all 30 participants, the answer key estimate based on the GCM correctly captured 83.63% of the 110-item answer key. Conversely, the majority rule correctly classified 80.90% of the answer key. It should be noted that the majority rule is not defined when both responses are chosen equally often. In these cases, the majority rule answer key estimate must be selected randomly. We counted each of these cases as .5 correctly estimated answer keys for the majority rule. This is in line with the expected value for random selections of dichotomous answer keys. The difference between GCM and majority rule of about 3% corresponds to four more answers being estimated correctly. Thus, the model-based analysis not only resulted in a high proportion of the answer key being correctly estimated overall but also outperformed the majority rule.

Despite this favorable result, the difference between GCM and majority rule performance appears small, if measured in terms of the proportion of correct item classifications. Does this mean that not much is gained by using the GCM instead of the simple majority rule in our case? This conclusion would be premature. An alternative index of the efficiency of both methods is the confidence associated with a specific answer key estimate. To reiterate, the MCMC method estimates the posterior distribution of the parameters given the data. Hence, we can use the posterior marginal probability of the $\hat{Z}_k$ estimate as a measure of the confidence we can place in this estimate. For example, posterior probabilities of .51 versus .99 for $Z_k = 1$ would both result in the answer key estimate $\hat{Z}_k = 1$ for item k, but the former would indicate low confidence in this estimate, whereas the latter would indicate high confidence. Similarly, a confidence measure for the majority rule can be derived by calculating the proportion of witnesses that agree with the majority response. Again, if 51% versus 99% of the witnesses agree with $Z_k = 1$, this would both result in the answer key estimate $\hat{Z}_k = 1$ according to the majority rule, but the former would clearly indicate much weaker confidence in this estimate compared with the latter. In our experiment, the mean confidences across items were .96 and .81 for the GCM answer key estimates and the majority rule, respectively, indicating that the GCM estimates are associated with much higher confidence compared with those derived from the majority rule. Thus, there is a strong benefit in terms of confidence in the answer key estimates when using the GCM method.

Group Size and GCM Performance

Thus far, we evaluated the GCM approach using all 30 participants. As revealed in the previously described survey of 773 Australian students (Paterson & Kemp, 2006), about half of the crimes are witnessed by at least four people (a median of 3 cowitnesses plus the reporting witness). To determine whether the GCM is able to render adequate results for smaller samples, or whether a certain minimum number of witnesses is required to produce stable results, we evaluated the GCM estimates as a function of the number of witnesses. Previously, it has been shown that the sample size required to classify items decisively into latent answer key categories depends on (a) the desired posterior confidence in the classifications, and (b) the average competence of the sample (see Batchelder & Romney, 1988, and Romney et al., 1986, for an overview). However, it is unclear how this translates to the conditions underlying the present experiment.

We repeatedly and randomly drew group sizes of $n = 4, 5, \ldots, 15$ from the participant pool of 30 individuals. We drew each group size 2,500 times (with replacement), irrespective of the video condition. Due to this large number of drawings, standard errors become very small. Hence, we abstained from using statistical tests in this analysis.

As illustrated in Figure 3, correlations of competence and response tendency estimates of the GCM and the 2-HTM were generally high and increased with witness group size, as expected. In other words, estimates without knowledge of the true answer key closely resembled estimates based on the true answer key, especially when the number of eyewitnesses was large.

Additionally, we computed the proportion of correct answer key estimates for the GCM and the majority rule as a function of group size. Whereas the majority rule performed on a relatively constant level of accuracy across sample sizes, the GCM performance increased with group size, consistently outperforming the majority rule (see Figure 4). Importantly, when the number of witnesses was even, the majority rule rendered fewer accurate results than the GCM. This is a result of both responses being chosen equally often (i.e., a standoff situation) and the subsequent random selection of the answer key estimate.

Discussion

When President Kennedy was shot, more than 500 people witnessed the event and were subsequently interviewed. Because the presumed assassin died soon after the incident, eyewitness accounts provided the main source of information in the investigation. One major question was the location from which the shots had been fired. Although nearly half the eyewitnesses interviewed

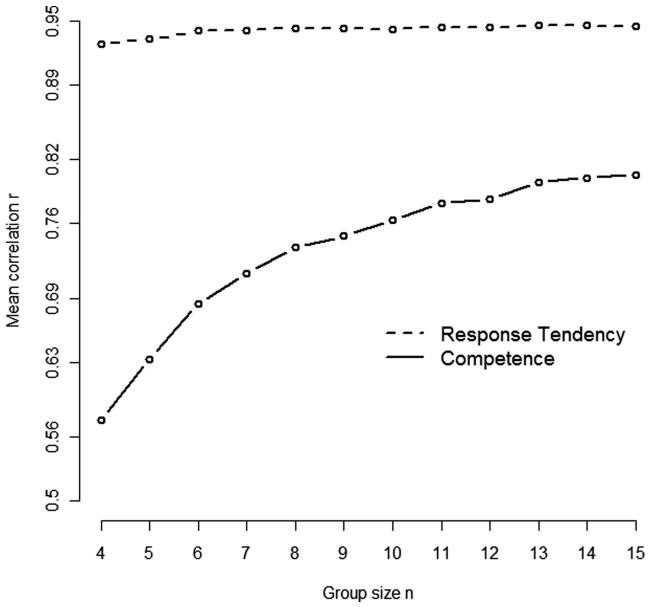

Figure 3. Mean correlations of response tendency parameter estimates based on the 2-HTM using the true answer key and the GCM ignoring the true answer key (dashed curve), and mean correlations of competence parameter estimates based on the 2-HTM using the true answer key and the GCM ignoring the true answer key (solid curve), separately for different numbers of independent eyewitnesses (group sizes $n = 4, 5, \ldots, 15$).

believed the shots came from the School Book Depository, and many people even stated seeing somebody with a rifle on the sixth floor of this building, many other people, most of whom were standing right next to the School Book Depository itself, believed that the shots originated from a railroad bridge nearby: "When the shots were fired, many people near the Depository believed that the shots came from the railroad bridge over the Triple Underpass or from the area to the west of the Depository" (President's Commission on the Assassination of President Kennedy, 1964, p. 71). Other people thought that the shots came from the vicinity of Elm and Houston, the area around the presidential limousine, or the trees north of Elm Street (p. 76).

In its report, the commission concluded that the School Book Depository was the assassin's location. Given the absence of knowledge about the actual chain of events, President Kennedy's assassination constituted a typical eyewitness situation with conflicting testimonies. Employing the GCM might have been helpful to derive more accurate estimates of witnesses' competences and the answer key. However, given the unavailability of the GCM, the police had to rely on intuitive judgments of eyewitnesses' competences or the less accurate majority rule.

In this paper, we introduced CCT and its formalization for dichotomous items, the GCM, as possible tools to assess witness accuracy and to derive a more accurate picture of a crime based on witness testimony. In an eyewitness recognition experiment, 30 participants watched a video depicting a heated quarrel between two people and then answered questions about this event. Results obtained with the GCM closely mirrored the true attributes of witnesses and items, especially for large numbers of witnesses.

Two findings are particularly important. First, the GCM competence estimates closely resembled the witnesses' actual accuracy. Thus, the GCM was able to identify those witnesses whose testimonies were most informative. Typically, researchers and people in the criminal legal system use self-reported confidence ratings to evaluate the quality of a witness's testimony. However, the strength of confidence ratings as an index of testimony accuracy largely depends on the circumstances under which one gathers the ratings (Brewer & Palmer, 2010; Brewer & Weber, 2008; Brewer & Wells, 2006). Therefore, confidence ratings are not always indicative of a witness's competence. Studies employing binary or multiple-choice recognition questions, after each of which participants give a confidence rating, found average confidence-accuracy correlation coefficients ranging from $r = -.01$ to $r = .77$ (Hollins & Perfect, 1997; Luna & Martín-Luengo, 2012; Roebers, 2002; Smith et al., 1989). In contrast, we found correlations between the GCM-based competence estimates and actual witness accuracies, as assessed with the 2-HTM, ranging between .58 and .81 across different sample sizes, indicating that the GCM may provide more valid measures of witness competence than self-reported confidence ratings.

The second important finding is that the GCM outperformed the majority rule in estimating the answer key, that is, the actual description of the crime, regardless of the number of witnesses. The larger the sample size, the more pronounced the advantage of the GCM over the majority rule was, both in terms of the percentage of correct item classifications and—even more so—in terms of the confidence associated with these classifications. Given the fact that the conditions were quite favorable for the majority rule in our experiment, because the overall level of accuracy was high and many participants were highly competent, this result is all the more remarkable. We therefore maintain that the GCM helps to overcome common problems in eyewitness testimony.

These two main findings provide a convincing basis for the GCM's potential application to real-life witness testimonies. Es-

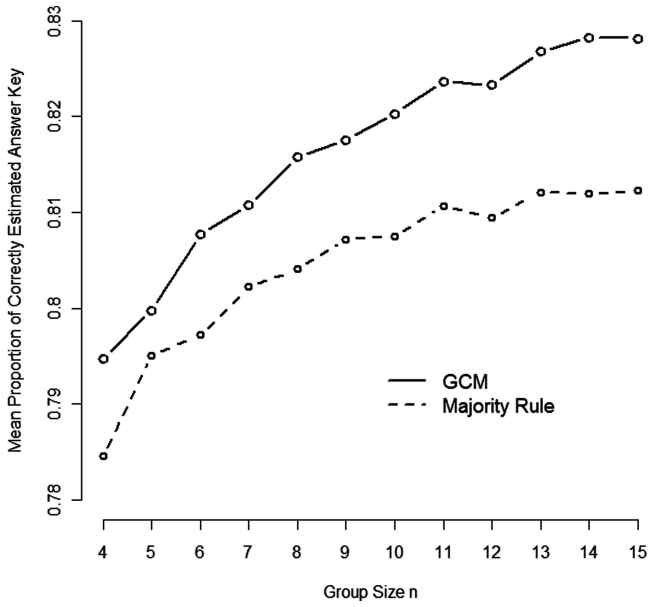

Figure 4. Mean proportion of match between the answer key estimates based on the GCM (solid curve) and the majority rule (dashed curve), respectively, and the true answer key as a function of the number of independent eyewitnesses (group sizes $n = 4, 5, \ldots, 15$).

pecially for large groups of witnesses, the GCM facilitates the incorporation of all witness statements and helps derive an estimate of what is likely true. It does so by reliably determining an individual witness's competence. This estimate may help a police officer evaluate a witness's statement. Moreover, with a larger proportion of the answer key being estimated correctly, one approaches the true chain of events, thereby enabling a better reconstruction of crime details. In the present study, the GCM outperformed the majority rule by correctly classifying 3% more of the items, or four more items, than did the majority rule. Four items may make a crucial difference, depending on their content. For example, correctly determining the ethnicity of a culprit may make an important difference to a case. In other cases, this small advantage of the GCM may be less important. However, in a more heterogeneous sample, the difference between the majority rule and the GCM is likely to be more pronounced. Especially when the majority is much less competent than the minority, the GCM may be far more accurate than the majority rule. Despite these qualities, little is known about the utility of the GCM in legal contexts. More research is needed, especially on the influence of different distributions of witness competences and item difficulties, but also on the robustness of the method against small to moderate violations of the independence assumptions underlying the GCM.

Despite these caveats, there are two major contributions of the present work to psycholegal research. First, in psychology and law, research on groups of eyewitnesses is largely limited to the investigation of cowitness talk and its potential downsides (Wright, Memon, Skagerberg, & Gabbert, 2009; Wright & Schwartz, 2010). In contrast, CCT focuses on advantages associated with groups of eyewitnesses testifying independently. Second, model-based analysis reflects a new trend in eyewitness research. Recently, the focus has been shifting toward more sophisticated research methods as the demand for more theoretically based research grows (Brewer & Wells, 2011; Sporer, 2008). However, there is still a gap between basic and applied research (Deffenbacher, 2008; Lane & Meissner, 2008) with "the pendulum . . . swinging too far in the applied direction" (Bornstein & Meissner, 2008, p. 734). Whereas some researchers suggest a "middle lane approach" (Lane & Meissner, 2008, p. 779), others go as far as stating the "importance (necessity) of computational modeling for eyewitness identification research" (Clark, 2008, p. 803). An advantage of formal models over nonformalized theories is their precision (Bjork, 1973), which helps objectify decision making in research (Hintzman, 1991). Only recently, Farrell and Lewandowsky (2010) noted that formal models facilitate scientific reasoning as researchers are "forced to specify all parts" of a theory (p. 330). As formal models are precise and aim at explaining underlying mechanisms, they facilitate the application of laboratory findings to real-world problems (Harley, Dillon, & G. R. Loftus, 2004). Correspondingly, formal models have numerous practical applications. For example, in an attempt to incorporate identification judgments from multiple witnesses, Clark and Wells (2008) modeled the memory processes of a group of witnesses (see also Clark, 2003). Another example is the recent work of Dunn and Kirsner (2011), who used formal models to identify and integrate relevant information from survivors to determine the locations of two ships that sank off the West Coast of Australia in November 1941 following a firefight in World War II. However, the methods of Clark and Wells (2008) and Dunn and Kirsner (2011) are specific for the analysis of eyewitness identification in lineups and location judgments, respectively. In the present article, we discuss model-based integration of witness reports on many details of a complex event for the first time.

Despite the advantages of applying CCT to the assessment of witness reports, we must acknowledge limitations to the study and the underlying theory in the context of witness research. The present experiment employed dichotomous items only. One might argue that this does not relate well to a "real" interrogation situation. Importantly, the model can be adapted to fit more complex data, such as multiple-choice questions or noncategorical data. Thus far, such versions of the model have been less thoroughly investigated. Before applying more complex versions of the GCM to real-world data, the model itself must be validated and tested in different contexts. Moreover, additional assumptions must be met, and little is known about potential consequences of their violation. For example, one prerequisite for multiple-choice items is the equity in a priori attractiveness of the answer options (cf. Batchelder & Romney, 1988). That is, any of the presented options is chosen with the same probability if the correct response is unknown to the respondent. For example, if a witness to President Kennedy's assassination had not known the number of shots fired, any single number of shots suggested by the police should have been perceived by the witness as being just as likely as all other numbers suggested. It is yet unclear whether violations of this rather implausible assumption would invalidate estimates of the GCM parameters derived from multiple-choice items.

An important and yet unstudied danger to the performance of the GCM is the presence of stereotypes and schemas—configurations in the data that are systematic but do not relate to competence. In such cases, stereotypical knowledge might be mistaken for competence in analyzing consensus. Hence, witnesses' responses based on stereotypes are assigned more importance when estimating the answer key than their competence would warrant. The issue of stereotypes has been investigated intensely in the context of witness memory research (cf. Allport & Postman, 1947; Peters, Jelicic, & Merckelbach, 2006; Shechory, Nachson, & Glicksohn, 2010; Stalans, 1993; Tuckey & Brewer, 2003). For high presentation pace, as in our experiment, stereotype-consistent information may be more readily stored and retrieved (Dijksterhuis & Van Knippenberg, 1995). However, in recognition memory research, schemas and scripts have also been found to impair memory accuracy for schema- and script-typical information (Erdfelder & Bredenkamp, 1998). Most importantly, when attention is low and distraction is high, which is frequently the case in witness situations, stereotyping strongly influences memory performance (Sherman, Groom, Ehrenberg, & Klauer, 2003; Sherman, Macrae, & Bodenhausen, 2000), leading to reporting of false memories (Tuckey & Brewer, 2003). To test the robustness of the GCM and compare it with that of the majority rule, further research using both stereotypical and nonstereotypical material is necessary.

Other obvious dangers to GCM validity in eyewitness testimony are leading questions and cowitness talk. Essentially, an individual such as a police officer can greatly impact witnesses' responses by asking leading questions (e.g., E. F. Loftus & Palmer, 1974). Cowitness talk is problematic for similar reasons. A common reaction to the observation of an unusual event is that one feels the need to discuss the observed event with another witness. If hundreds of people watch their president drive past in a motorcade,

and, all of a sudden, they hear shots and see the president slump to his side, many would be confused about what they had just seen. Hence, they would likely discuss what they saw with someone standing near them. These discussions can have distorting effects on memory (for an overview, cf. Wright et al., 2009). For the GCM, several witnesses simultaneously altering the memory of their observation due to the conversation based on *incorrect* information would clearly be problematic. Cowitness talk could easily induce correlations between witnesses' responses across items that are not indicative of their truth. Research on this topic is necessary to determine the impact of cowitness talk on the GCM-based estimates. Again, however, it should be noted that stereotypes and cowitness talk might also impair other aggregation procedures for multiple eyewitness testimonies such as the majority rule.

Despite these limitations, the present work theoretically underpins and improves an issue that is central to witness research but has so far been neglected: How can we learn best from multiple independent witness testimonies? Formal modeling promises significant advances over the status quo in answering this question. For dichotomous data, the GCM is a valuable and accurate method to reconstruct an unknown target event from multiple witness reports. Hence, the GCM provides eyewitness research with a valuable addition to the eyewitness accuracy assessment toolbox.

References

Allport, G., & Postman, L. J. (1947). *The psychology of rumor*. New York, NY: H. Holt.

Aβfalg, A., & Erdfelder, E. (2012). CAML – Maximum likelihood consensus analysis. *Behavior Research Methods, 44,* 189–201. doi:10.3758/s13428-011-0138-0

Batchelder, W. H., Kumbasar, E., & Boyd, J. P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology, 22,* 29–58. doi:10.1080/0022250X.1997.9990193

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103–112). Greenwich, CT: JAI.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53,* 71–92. doi:10.1007/BF02294195

Batchelder, W. H., & Romney, A. K. (1989). New results in test theory without an answer key. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 229–248). Berlin, Germany: Springer.

Batchelder, W. H., Strashny, A., & Romney, A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. *Advances in Social Computing, Lecture Notes in Computer Science, 6007,* 98–107. doi:10.1007/978-3-642-12079-4_15

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4,* 370–374. doi:10.1111/j.1745-6924.2009.01140.x

Bjork, R. A. (1973). Why mathematical models? *American Psychologist, 28,* 426–433. doi:10.1037/h0034623

Bornstein, B. H., & Meissner, C. A. (2008). Basic and applied issues in eyewitness research: A Münsterberg centennial retrospective. *Applied Cognitive Psychology, 22,* 733–736. doi:10.1002/acp.1478

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15,* 77–96. doi:10.1348/135532509X414765

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indices of memory processes not just markers of accuracy. *Applied Cognitive Psychology, 22,* 827–840. doi:10.1002/acp.1486

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

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

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 587–606. doi:10.1037/a0015279

Campos, L., & Alonso-Quecuty, M. L. (1999). The cognitive interview: Much more than simply try again. *Psychology, Crime & Law, 5,* 47–59. doi:10.1080/10683169908414993

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. London, UK: Chapman & Hall.

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17,* 629–654. doi:10.1002/acp.891

Clark, S. E. (2008). The importance (necessity) of computational modeling for eyewitness identification research. *Applied Cognitive Psychology, 22,* 803–813. doi:10.1002/acp.1484

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior, 32,* 406–422. doi:10.1007/s10979-007-9115-7

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review, 102,* 396–408. doi:10.1037/0033-295X.102.2.396

Dando, C., Wilcock, R., & Milne, R. (2009). The cognitive interview: The efficacy of a modified mental reinstatement of context procedure for frontline police investigators. *Applied Cognitive Psychology, 23,* 138–147. doi:10.1002/acp.1451

Deffenbacher, K. A. (2008). Estimating the impact of estimator variables on eyewitness identification: A fruitful marriage of practical problem solving and psychological theorizing. *Applied Cognitive Psychology, 22,* 815–826. doi:10.1002/acp.1485

Dijksterhuis, A., & Van Knippenberg, A. (1995). Memory for stereotype-consistent and stereotype-inconsistent information as a function of processing pace. *European Journal of Social Psychology, 25,* 689–693. doi:10.1002/ejsp.2420250607

Dripps, D. A. (1999). Miscarriages of justice and the constitution. *Buffalo Criminal Law Review, 2,* 635–680.

Dunn, J. C., & Kirsner, K. (2011). The search for HMAS Sydney II: Analysis and integration of survivor reports. *Applied Cognitive Psychology, 25,* 513–527. doi:10.1002/acp.1735

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology, 217,* 108–124. doi:10.1027/0044-3409.217.3.108

Erdfelder, E., & Bredenkamp, J. (1998). Recognition of script-typical versus script-atypical information: Effects of cognitive elaboration. *Memory & Cognition, 26,* 922–938. doi:10.3758/BF03201173

Erdfelder, E., Cüpper, L., Auer, T.-S., & Undorf, M. (2007). The four-states model of memory retrieval experiences. *Zeitschrift für Psychologie/Journal of Psychology, 215,* 61–71. doi:10.1027/0044-3409.215.1.61

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making, 6,* 7–22.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science, 19,* 329–335. doi:10.1177/0963721410386677

Fischer, G. H., & Molenaar, I. (Eds.). (1995). *Rasch models - Foundations, recent developments, and applications*. New York: Springer.

Fisher, R. P., Milne, R., & Bull, R. (2011). Interviewing cooperative witnesses. *Current Directions in Psychological Science, 20,* 16–19. doi:10.1177/0963721410396826

Frenda, S. F., Nichols, R. M., & Loftus, E. F. (2011). Current issues and advances in misinformation research. *Current Directions in Psychological Science, 20,* 20–23. doi:10.1177/0963721410396620

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1985). Eyewitness memory enhancement in the police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology, 70,* 401–412. doi:10.1037/0021-9010.70.2.401

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1986). Enhancement of eyewitness memory with the cognitive interview. *The American Journal of Psychology, 99,* 385–401. doi:10.2307/1422492

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–807.

Harley, E. M., Dillon, A. M., & Loftus, G. R. (2004). Why it's difficult to see in the fog: How contrast affects visual perception and visual memory. *Psychonomic Bulletin & Review, 11,* 197–231. doi:10.3758/BF03196564

Hilbig, B. E. (2012). How framing statistical statements affects subjective veracity: Validation and application of a multinomial model for judgments of truth. *Cognition, 125,* 37–48. doi:10.1016/j.cognition.2012.06.009

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Hillsdale, NJ: Lawrence Erlbaum.

Hollins, T. S., & Perfect, T. J. (1997). The confidence-accuracy relation in eyewitness event memory: The mixed question type effect. *Legal and Criminological Psychology, 2,* 205–218. doi:10.1111/j.2044-8333.1997.tb00344.x

Janssen, J., Kirschner, F., Erkens, G., Kirschner, P. A., & Paas, F. (2010). Making the black box of collaborative learning transparent: Combining process-oriented and cognitive load approaches. *Educational Psychology Review, 22,* 139–154. doi:10.1007/s10648-010-9131-x

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika, 68,* 373–389. doi:10.1007/BF02294733

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). The cognitive interview: A meta-analysis. *Psychology, Crime & Law, 5,* 3–27. doi:10.1080/10683169908414991

Lane, S. M., & Meissner, C. A. (2008). A "middle road" approach to bridging the basic-applied divide in eyewitness identification research. *Applied Cognitive Psychology, 22,* 779–787. doi:10.1002/acp.1482

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

Lindsay, R. C. L., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66,* 79–89. doi:10.1037/0021-9010.66.1.79

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7,* 560–572. doi:10.1016/0010-0285(75)90023-7

Loftus, E. F. (1996). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning & Verbal Behavior, 13,* 585–589. doi:10.1016/S0022-5371(74)80011-3

Luna, K., & Martín-Luengo, B. (2012). Confidence-accuracy calibration with general knowledge and eyewitness memory cued recall questions. *Applied Cognitive Psychology, 26,* 289–295. doi:10.1002/acp.1822

Milne, R., & Bull, R. (2002). Back to basics: A componential analysis of the original cognitive interview mnemonics with three age groups. *Applied Cognitive Psychology, 16,* 743–753. doi:10.1002/acp.825

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16,* 519–533. doi:10.1016/S0022-5371(77)80016-9

Münsterberg, H. (1908). *On the witness stand. Essays on psychology and crime*. New York, NY: Doubleday, Page.

Pansky, A., Koriat, A., & Goldsmith, M. (2005). Eyewitness recall and testimony. In N. Brewer, & K. D. Williams (Eds.), *Psychology and law: An empirical perspective* (pp. 93–150). New York, NY: Guilford.

Paterson, H. M., & Kemp, R. I. (2006). Co-witness talk: A survey of eyewitness discussion. *Psychology, Crime & Law, 12,* 181–191. doi:10.1080/10683160512331316334

Peters, M. J. V., Jelicic, M., & Merckelbach, H. (2006). When stereotypes backfire: Trying to suppress stereotypes produces false recollections of a crime. *Legal and Criminological Psychology, 11,* 327–336. doi:10.1348/135532505X74055

President's Commission on the Assassination of President Kennedy. (1964). *Report of the President's Commission on the Assassination of President Kennedy*. Washington, DC: U.S. Government Printing Office. Retrieved from http://www.archives.gov/research/jfk/warren-commission-report/letter.html

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, UK: Nielsen & Lydiche.

R Development Core Team. (2011). The R-project for statistical computing. Retrieved from http://www.r-project.org/

Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38,* 1052–1067. doi:10.1037/0012-1649.38.6.1052

Romney, A. K. (1999). Culture consensus as a statistical model. *Current Anthropology, 40,* 103–115. doi:1086/200062

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist, 31,* 163–177. doi:10.1177/000276487031002003

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88,* 313–338. doi:10.1525/aa.1986.88.2.02a00020

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

Schmechel, R. S., O'Toole, T. P., Easterly, C., & Loftus, E. F. (2006). Beyond the ken? Testing jurors' understanding of eyewitness reliability evidence. *Jurimetrics, 46,* 177–214.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 641–649. doi:10.1037/0278-7393.29.4.641

Shechory, M., Nachson, I., & Glicksohn, J. (2010). Effects of stereotypes and suggestion on memory. *International Journal of Offender Therapy and Comparative Criminology, 54,* 113–130. doi:10.1177/0306624X08322217

Sherman, J. W., Groom, C. J., Ehrenberg, K., & Klauer, K. C. (2003). Bearing false witness under pressure: Implicit and explicit components of stereotype-driven memory distortions. *Social Cognition, 21,* 213–246. doi:10.1521/soco.21.3.213.25340

Sherman, J. W., Macrae, C. N., & Bodenhausen, G. V. (2000). Attention and stereotyping: Cognitive constraints on the construction of meaningful social impressions. *European Review of Social Psychology, 11,* 145–175. doi:10.1080/14792772043000022

Simons, D. J., & Chabris, C. F. (2011). What people believe about how the memory works: A representative survey of the U.S. population. *PloS ONE, 6,* e22757. doi:10.1371/journal.pone.0022757

Smith, V. L., Kassin, S. M., & Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology, 74,* 356–359. doi:10.1037/0021-9010.74.2.356

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50. doi:10.1037/0096-3445.117.1.34

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64,* 583–639. doi:1369-7412/02/64583

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22,* 737–757. doi:10.1002/acp.1479

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

Stalans, L. J. (1993). Citizens' crime stereotypes, biased recall and punishment preferences in abstract cases: The educative role of interpersonal sources. *Law and Human Behavior, 17,* 451–470. doi:10.1007/BF01044378

Tuckey, M. R., & Brewer, N. (2003). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied, 9,* 101–118. doi:10.1037/1076-898X.9.2.101

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. doi:10.1037/h0020071

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10,* 156–172. doi:10.1037/1076-898X.10.3.156

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist, 31,* 178–193. doi:10.1177/000276487031002004

Weller, S. C., Romney, A. K., & Orr, D. P. (1987). The myth of a sub-culture of corporal punishment. *Human Organization, 46,* 39–47.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence. Improving its probative value. *Psychological Science in the Public Interest, 7,* 45–75. doi:10.1111/j.1529-1006.2006.00027.x

Wise, R. A., & Safer, M. A. (2004). What US judges know and believe about eyewitness testimony. *Applied Cognitive Psychology, 18,* 427–443. doi:10.1002/acp.993

Wright, D. B., Memon, A., Skagerberg, E. M., & Gabbert, F. (2009). When eyewitnesses talk. *Current Directions in Psychological Science, 18,* 174–178. doi:10.1111/j.1467-8721.2009.01631.x

Wright, D. B., & Schwartz, S. L. (2010). Conformity effects in memory for actions. *Memory & Cognition, 38,* 1077–1086. doi:10.3758/MC.38.8.1077

Appendix

Relation Between Accuracy and Competence

Observation.

If $p\ (Z = 1) = .5$, then the proportion of correct responses, p_i, is an affine function of the 2-HTM competence parameter D_i. Hence, p_i and D_i correlate perfectly across informants.

Proof. If $P_Z = p(Z = 1)$, then according to the 2-HTM,

$$p_i = P_Z H_i + \left(1 - P_z\right)(1 - F_i), \tag{A1}$$

where H_i and F_i are the hit and false alarm probabilities for informant i, respectively. If $P_Z = .5$, it follows from Equation A1 that

$$p_i = .5\ (1 + H_i - F_i). \tag{A2}$$

Using Equation 1 it follows from Equation A2 that the accuracy p_i is an affine function of D_i:

$$p_i = .5\ (1 + D_i). \tag{A3}$$

This completes the proof.

Received July 29, 2011
Revision received June 25, 2012
Accepted July 9, 2012 ■