Photo Arrays in Eyewitness Identification Procedures:

Follow-up on the Test of Sequential versus Simultaneous Procedures (Study One)

and

An Experimental Study of the Effect of Photo Arrays on
Evaluations of Evidentiary Strength by Key Criminal Justice Decision Makers (Study Two)

REPORT UNDER REVISION

Karen L. Amendola, Ph.D.
Maria D. Valdovinos, B.A.
Meghan G. Slipka, M.A.
Earl Hamilton, M.A.
Mary Sigler, B.A.
Adam Kaufman, B.A.

POLICE FOUNDATION
1201 Connecticut Avenue, N.W.
Suite 200
Washington, DC 20036
www.policefoundation.org

Chief James Bueermann (ret.)
President

February 25, 2014

This project includes a follow-up of the 2011 American Judicature Society's Eyewitness ID Field Studies Project and was supported by generous grants from the *Laura and John Arnold Foundation*, the *Open Society Foundations*, and the former *JEHT Foundation*.

The findings and discussion herein represent those of the authors and not necessarily those of the funding sources.

Table of Contents

List of Tables and Figures

TABLES

FIGURES

ACKNOWLEDGMENTS

As with any large-scale project, so many people contributed to these studies and the development of the *Evidentiary Strength Scale* (Amendola & Slipka, 2009) used herein, that it would be impossible to mention all of them. Nevertheless, a number of people were instrumental in the project from its conception to completion. First and foremost we would like to acknowledge those organizations and individuals who provided financial and programmatic support for this comprehensive effort, as the projects described herein, would not have been possible without them. We are very thankful to the *JEHT* Foundation. While no longer in operation, *JEHT* and its staff provided resources to numerous organizations focused on justice-related issues. We are particularly grateful to Scott Matson, M.A. and now Senior Policy Advisor at the U.S. Department of Justice for his initial guidance in our effort. Next, we greatly appreciate the fact that the *Open Society Foundations* (OSF) recognized the importance of this work already in progress, and committed to provide funding to see that it was completed. We especially thank the oversight provided by Terrance Pitts, J.D., our program officer at the Justice Fund at OSF. Finally, we are sincerely grateful to the Laura and John Arnold Foundation as they made it possible for us to conduct our study in Austin, Texas. These organizations represent the best of the best in terms of their commitment to justice system improvements nationwide and abroad.

Because our experimental study was carried out in Travis County, Texas, we would like to extend a heartfelt thank you to the staff and leadership of the Travis County, Texas District Attorneys' Office, and in particular, District Attorney Rosemary Lehmberg, who afforded us the opportunity to work with her agency to ensure we could carry out the study. She carefully guarded confidential data and information, while allowing us access to staff and case file information necessary for this effort. It is not easy for any agency head to take on the task of supporting scientific research for the sake of improving field practices. Her willingness to participate in this effort was the essential element in making this project a success. Former ADA Claire Dawson-Brown, also of the Travis County DA's Office, lent tremendous support to this effort by providing necessary case information, providing assistance throughout our three weeks on site in Austin, and by recruiting police investigators, prosecutors, defense attorneys, and judges from the immediate area to serve as our study participants/subject matter experts. While we are highly grateful to all of those individuals in Austin, our confidentiality agreement does not allow us to mention each by name, but we thank them anonymously. We are also very grateful for the insights provided by both First Assistant District Attorney John Neal and ADA Bryan Case in the Travis County DA's office.

In addition, we acknowledge Chief Art Acevedo of the Austin Police Department who willingly participated in our study and provided the necessary staff and commitment to ensure project success. We also thank Assistant Chief David Carter, Commander Julie O'Brien, Lieutenant Mark Spangler, and Commander Troy Gay in providing us with the much-needed data, information and rapid response to all inquiries.

Numerous individuals in all of the project sites supported these efforts substantively. In San Diego, Chief William Lansdowne and key staff of the San Diego Police Department including Captain James Collins, Lieutenant Kevin Ammon, and Karen Goodman were very supportive of our efforts, as was Deputy District Attorney Tia Quick and the San Diego District Attorney's Office.

In Mecklenburg County, NC we are grateful for the support of Chief Rodney Monroe of the Charlotte Mecklenburg Police Department (CMPD) as well as former Chief Darrel Stephens (ret.), who both provided necessary data and guidance. We also appreciate the assistance provided by Major Kevin Wittman (ret.) and Major Paul Zinkann III, both of the CMPD. We wish to acknowledge the support of the Mecklenburg County District Attorney's Office and ADA Bart Menser for assisting us with conducting the pilot test. Finally, all present and former police investigators, prosecutors, defense attorneys, and judges from the surrounding area who participated as case evaluators for our pilot test, are sincerely thanked for their terrific insights into our *Strength of Evidence Scale*, and the process of conducting consensus discussions. While our confidentiality agreement does not allow us to name them individually, we nevertheless thank those whose participation was invaluable to subsequent study efforts.

We also thank Lisa Judge, Principal Assistant City Attorney of the Tucson City Prosecutor's Office for providing us with case disposition data for the cases in that site.

We are pleased to acknowledge the tremendous effort of our partners, the American Judicature Society (AJS) and the Innocence Project who supported us in carrying out an evaluation of outcomes from the Wells, Steblay, and Dysart (2011) study of sequential and simultaneous presentation methods. Former AJS project manager Dani (Danielle) Mitchell was an enormous help in assuring data convergence between our project and that of AJS. Her tireless efforts to support our project, especially the pilot test we conducted in Charlotte, were exceptional at every step. We would especially like to thank Barry Scheck, JD, Co-Founder of the Innocence Project, whose dedication to equal justice and steadfast commitment to evidence-based practices is unwavering. His efforts to ensure we conducted an objective and comprehensive examination were unmatched. In addition, Executive Director Maddy deLone at the Innocence Project provided us with extensive program support and demonstrated a unique commitment to this project every step of the way by helping us in troubleshooting issues as those arose. Importantly, we are very grateful to the efforts of Professor Gary L. Wells, PhD, Nancy K. Steblay, PhD, and Jennifer E. Dysart, the scientists who conducted the AJS National Eyewitness Identification Field Studies. We especially thank Gary for his leadership of that effort and his assistance in helping us to reconcile issues that arose in our study. In addition, we are very grateful to Nancy Steblay who worked with our team to ensure that all necessary data were provided to support our efforts.

Many individuals worked with us to develop the *Evidentiary Strength Scale,* indeed, too many to name. Most notably was James M. Doyle, J.D. a current fellow at the National Institute of Justice, who helped us to gain perspective on evidence and evidentiary standards, and to think more broadly about these issues. In addition, Dr. Jon M. Shane of John Jay College of Criminal Justice and former police captain in Newark, NJ provided us with considerable information about the context of policing. We also thank the following individuals for their participation in this effort: Captain Sam Cogen of the Baltimore (MD) City Sheriff's Office; Rachel E. Cogen, District Court Judge in Baltimore (MD); Anne M. Corbin, JD; Judge Gregory Jackson (DC); Christopher Ortiz, PhD, City of Glen Cove (NY) Police Department; Chief LeRoy O'Shield (ret.); Professor Alfred Slocum (Rutgers School of Law); Lieutenant LaTesha Watson of the Arlington (TX) Police Department; Judge William Walls (NJ District Court); Assistant Prosecutor Gwendolyn Williams, Newark (NJ). We also thank all those who participated in completing our survey to help generate data useful in our establishing content-oriented validity for the *Strength of Evidence Scale.*

Furthermore, we gratefully acknowledge Jon Gould, J.D., Ph.D. of American University in Washington (DC) and Julia Carrano, J.D. (formerly of American University) for their reviews and commentary on the *Strength of Evidence Scale* and involving us in their work on wrongful convictions. We would like to make a special note of appreciation to Christine C. Mumma, J.D. of the NC Commission on Actual Innocence and Adjunct Professor at the UNC School of Law (NC) for her many insights into this effort. Special thanks to both Patricia Riley and Renata Cooper, Special Counsel to the U.S. Attorney, United States District Court (DC) and Kristine Hamann, J.D., currently a visiting fellow at the Bureau of Justice Assistance of the U.S. Department of Justice, for providing insights into the challenges of prosecutors, guidelines, and best practices.

The authors of this report would like to especially thank Professor Nancy Steblay (Augsburg College), Emily West, Ph.D. (Research Director at the Innocence Project), and Professor David Weisburd (George Mason University) among others for their thoughtful review, comments, and recommendations for improving this report. The Principal Investigator also wishes to thank Professor Weisburd for his ongoing mentorship and contributions to the Police Foundation.

Finally, the PI extends a warm thank you to Chief Jim Bueermann (ret.), President of the Police Foundation for providing an environment where scientific research, innovation, and evidence-based practices are recognized as paramount to the field of policing. Finally, the PI thanks former President, Hubert Williams who believed strongly in applying research to practice and was therefore supportive of this important effort.

Photo Arrays in Eyewitness Identification Procedures

Executive Summary

Eyewitness identification procedures have been the subject of much scientific and public policy controversy in recent years (Clark, 2012; Cutler & Kovera, 2008; Garrett, 2008). A recent experimental field test conducted by the American Judicature Society (AJS) in four sites, was designed to test the effect of police departments' methods for presenting photo arrays of suspects on the identification choices made by victims and other eyewitnesses (Wells, Steblay, & Dysart, 2011). The presentation methods compared in that study included the simultaneous procedure (all photos of potential suspects are shown at one time) versus the sequential approach (showing photos of potential suspects one at a time). Known as the American Judicature Society National Eyewitness Identification Field Studies (hereinafter referred to as *the AJS' EWID Field Studies)*, that effort stands as the first multi-site field-test of sequential and simultaneous presentation methods employing a range of procedures (known as the *Greensboro Protocols*[1]) to reduce bias and standardize lineup protocols.

Despite the fact that there are numerous other influences on the reliability, validity, and accuracy of victims/witnesses such as viewing opportunity, characteristics of the offender, characteristics of the crime, and time lag since the crime, etc., the findings demonstrated that *the sequential method resulted in significantly fewer misidentifications* (picking a "filler" over the police-identified suspect) though fewer overall picks (Wells, Steblay, & Dysart, 2011), findings consistent with much of the accumulated evidence from laboratory studies.

Nevertheless, little is known about whether that finding (the simultaneous presentation

[1] The *Greensboro Protocols*, as they are known, represent a series of reforms for guiding the design and rigor of future field studies as decided by a group of lawyers, prosecutors, police investigators, and leading behavioral scientists in the field of eyewitness identification that convened in Greensboro, NC in September 2006.

method leads to more filler picks) actually makes a difference in terms of the dispositional outcomes of those cases. Indeed, there are many other things including other forms of evidence that lead to case dispositions. As such, the Police Foundation research team conducted a follow-up study of the Wells, et al. (2011) study to examine the relationship between presentation methods and case outcomes. Referred to as *"Photo Arrays in Eyewitness Identification Procedures,"* the goals of this study were to examine the extent to which photo array presentation methods and their outcomes (pick types) were associated with case dispositions (adjudicated guilty or not[2]) across the three field sites, as well as evidentiary strength as rated by police investigators, prosecutors, defense attorneys, and judges in Austin, Texas. The latter is particularly important as these criminal justice practitioners are responsible for interpreting evidence and facilitating outcomes for the more than 90% of cases that are adjudicated without juries; primarily, those cases that result in plea deals or, on occasion, bench trials.

One of the important elements of the studies we present in this paper, was the development of the *"Evidentiary Strength Scale"* (Amendola & Slipka, 2009, see **Appendix A**), an objective and scientifically derived instrument used by key criminal justice decision makers in this study to rate evidentiary strength in the cases. Evidence of the scale's reliability and initial content-oriented validity were collected through a series of procedures[3] prior to the implementation of these studies and is the subject of a separate manuscript (Amendola, forthcoming). The inclusion and almost exclusive input of police, prosecutors, defense attorneys, and judges (as key subject matter experts) in the development of that instrument, renders it particularly relevant to real-world cases and criminal justice proceedings. While the instrument

[2] Adjudicated guilty means by plea or finding, not adjudicated means insufficient evidence and/or not prosecuted.

[3] Scale development is described elsewhere (Amendola, forthcoming). In addition, the scale has since been used in other research by Gould, J.B, Carrano, J., Leo, R. and Young, J. *Predicting Erroneous Convictions: A Social Science Approach to Miscarriages of Justice*, Final Report, Washington, D.C.: U.S. Department of Justice, National Institute of Justice, December 2012, Grant No. 2009-IJ-CX-4110.

was informed by other research and scientific knowledge about instrument development and validation, the content was that of the subject matter experts alone. Rated evidentiary strength was used as a proxy for ground truth (actual guilt or innocence), as it is clear that case dispositions may result from extra-evidentiary factors, such as the willingness of victims and/or witnesses to testify (e.g. witness intimidation); the exclusionary rule (that may prevent a guilty person from being convicted); personal biases and/or the "stories" or hypotheses about the suspect's involvement and the events of the crime constructed by prosecutors (that may cause an innocent person to be convicted); and alternative theories presented by defense attorneys (again, that may result in putting a guilty person back on the streets).

In addition, we present a second experimental study in Austin, Texas, the site from the *AJS' EWID Field Studies* that generated the greatest number of photo arrays in which we examined the impact of including photo arrays and the pick types made on evaluators' ratings of the overall evidentiary strength of the cases, as well as any specific type of evidence in the case. This was achieved by providing one group of evaluators the complete cases, and the other group with the same cases in which the photo arrays and pick types were fully redacted.

Findings

The presentation method did not relate to the case dispositions, suggesting that the method may not make a difference in terms of case outcomes. The Wells et al. (2011) finding that simultaneous photo arrays lead to substantially more filler picks, ultimately did not matter in these cases in terms of outcomes. In essence, because the misidentifications were those in which known innocents (fillers or "foils") were selected, it may indicate that "filler picks" are not necessarily representative of the more consequential error of picking a suspect in a lineup where the actual perpetrator is missing. Practitioners and scientists alike have acknowledged and asserted that fillers picked from lineups are very unlikely to be prosecuted, as the fillers are individuals whom, with exceedingly few exceptions, are “known innocents.” At the same time, it is important to note that as a result of an increasing number of cases in which DNA evidence has exonerated individuals erroneously convicted by eyewitness evidence, or perhaps with other advances and/or increasing public pressure over time, many policy and practice changes have occurred in prosecutors’ offices, as well as police departments regarding the unreliability of eyewitness id evidence, and its relative importance in establishing guilt or innocence.

In terms of case outcomes, the results of this study show that the evidentiary strength ratings were virtually the same (no statistically significant differences) across the groups who had the photo identification information and those who did not, when considering cases in which fillers were picked and in which no picks were made. And while the evidentiary strength in the cases that resulted in guilty outcomes were rated higher when suspects were picked, that only mattered for the cases in which the evidence was already particularly strong.

Our examination of observational data from the Wells et al. (2011) study revealed a strong association between lineup choices and case dispositions in that a greater proportion of cases with guilty findings were associated with suspect picks, as compared to those in which no picks were made or fillers were picked. Conversely, among cases that were not adjudicated, a greater proportion of them were associated with no picks and filler picks as compared to suspect picks, as might be expected. However, while these relationships were strong, they do not reflect a cause and effect relationship. Indeed there are many other factors, beyond photo array results, that influence case dispositions and these factors, e.g. other case evidence, may actually explain the outcomes of the cases, as is elaborated upon below.

To the extent that our rating scale differentiated adjudicated guilty and non-adjudicated cases in the sample, it demonstrates that the rating scale is valid as a proxy for ground truth, lending to the validity of the instrument in accurately representing case strength. Indeed, the mean evidentiary strength ratings (on a 5-point scale) for those cases adjudicated guilty versus not prosecuted were well over 4 and well below 3, respectively.

One of the key findings from this study was that the inclusion of a photo array in a case does not appear to have a significant influence on the overall ratings of evidentiary strength by key criminal justice decision makers. Indeed, for the cases in which photo lineup information (including pick type) was provided to the evaluators (the “yes” experimental condition), the evidentiary strength ratings were statistically equivalent to those from evaluators to whom no information about a photo array or its outcomes was provided (the “no” experimental condition). This indicates that the pick of a suspect is not likely to be the source of the increased ratings of overall case strength; instead, evaluators saw those cases as stronger despite the inclusion of a photo array. The one exception to that however, was that for those cases that had been adjudicated guilty,[4] evaluators assigned higher ratings when they knew there was a suspect pick (mean of 4.33), but *only* in those cases where the evidence was particularly strong despite the lineup (3.99 on a 5 point scale), and thus would have likely also resulted in conviction. In other words, the only time that suspect picks affected case dispositions, at least in Austin, was when the cases were already strong enough to be prosecuted at the outset.

Essentially, these findings suggest that the police in Austin most likely had identified the actual perpetrators and not the “wrong person,” given the strength of other corroborating evidence in the cases. Additionally, they suggest that the inclusion of a photo array does not

[4] Despite the fact that evaluators did not know how the cases had been adjudicated.

provide any additional meaningful benefit to the *evidentiary basis* for the case (neither strengthening or weakening it) in the eyes of police, prosecutors, defense attorneys, or judges than would be provided without a photo array, both a serendipitous and counter-intuitive finding. This key finding does not imply, however, that photo arrays are not diagnostic among police as photo arrays may have some investigative importance. Indeed, police may use them as a tool to help guide their investigations.

More research is needed, however, to examine whether policies or procedures in investigative units specify the need for a documented justification for including a potential suspect/confirmed suspect in a lineup or photo array. In this study, the researchers observed some cases in which the rationale for the inclusion of a particular suspect in a photo array was not documented in the case file, and was not readily apparent. This does not necessarily mean the investigators did not have valid reasons, simply that they were not always specified in the case file. Without a documented justification, it may be that the administration of an array or lineup is premature; if a suspect is picked, it may lead an investigator in one particular direction (i.e., put too much weight on the suspect pick, despite the known problems with eyewitness identification reliability) while at the same time, "ruling out" another viable suspect.

Furthermore, the fact that the picks do not provide any incremental value in most cases does not mean that they would not strengthen the police and prosecutor "stories" in court cases (heard by juries). There is no doubt that when a victim in a courtroom points to the suspect being tried and says, "it was him," that it has a profound effect on the jury (or any observer for that matter) in favor of the prosecution's case. Indeed, research with jurors/juries on the role of eyewitness id information has regularly shown it to have significant biasing effects on juries (Bodenhausen, 1990; Chapadelaine & Griffin, 1997; Kerr et al., 2008).

The key here is that the drama-induced impact is not necessarily reflective of ground truth. In other words, the "story" of the case may be improved when a suspect is picked from a photo array (even when other witnesses or victims do not pick the suspect or pick a filler) or worsened when fillers are picked or no one is picked (in favor of the defense). Nevertheless, key decision makers in our study were not necessarily strongly influenced by the photo array outcomes in interpreting evidence and its strength in connecting the suspect to the crime.

Indeed, an anecdotal finding by researchers in this study (both in the pilot and full study) was that police, prosecutors, and defense attorneys typically refer to "cases" as the entirety of the case they would present if heard by a jury, i.e. the evidence, as well as the context and prosecutor proposed theory/story of the crime, or the plausible alternative explanations offered by the defense, and not necessarily the objective evidence alone. This is the reality of the adversarial system, but when no courtroom story lines are required (as is evident in the vast majority of cases that result in plea agreements), it is very important that the key decision makers are able to interpret the evidence in an objective manner in order to ensure justice. It is for this reason that we were also concerned about whether the knowledge of a suspect or filler pick would bias the interpretation of other case evidence by police, prosecutors, defense attorneys, or judges.

However, we did not find a significant biasing effect of suspect picks on interpretation of other case evidence, suggesting that criminal justice decision makers can sufficiently separate different types of evidence, lending validity to the "*Evidentiary Strength Scale*" (Amendola & Slipka, 2009). This instrument shows promise as a tool for prosecutors (and potentially others) to separate the individual case facts from the context or "story" about the case, which should lend validity to the accuracy of their interpretations of evidentiary strength in delivering just outcomes. There was evidence of strong consistency of ratings among the evaluators in this data set, suggesting that various evidentiary factors can indeed be assigned values and relative weights (Amendola, forthcoming).

While it may be surprising that photo array outcomes did not even bias ratings of "identification information,"[5] it does underscore the fact that cases often have a range of identification information that allows them to connect suspects to the crime, rendering the result of a photo array a relatively unimportant factor among them. These other factors considered in the identification information category include: a) clothing, tattoo, hair styles, and other perpetrator descriptions that can help identify the suspect; b) details of the crime obtained through the investigation (e.g. finding a stolen item on a suspect, etc.); c) witness id information (e.g., detailed account of incident is given by witness consistent with other evidence, etc.); d) third party/ complainant information (e.g. pawn shop owner knows suspect and verifies he/she came in with stolen property, third party statement implicating the suspect, etc.); and e) circumstances surrounding arrest (e.g. suspect hiding near crime scene, etc.); f) co-conspirator flips, thereby implicating suspect; and g) anonymously provided information. Identification information, therefore, may be stronger, more reliable, or more important to criminal justice decision makers than photo arrays and their results, or it may simply stand on its own without the need for a photo array. Importantly, all types of decision makers ranked the identification information among the most important of the six categories of evidence (police investigators = #2, prosecutors, defense attorneys, and judges = #1). Similarly, all but judges ranked the physical evidence among the most important type of evidence,[6] although judges seemed to think that characteristics of witnesses and victims were more important than physical evidence.

[5] "Identification Information" was defined by the participants in the instrument development phase of the project as "*Independent corroboration of information linking the suspect to the particular incident, regardless of source.*"

[6] There were six distinct categories of evidence as determined by criminal justice decision makers; highest rank means either a "1" or a "2."

There were some distinct differences in the way that criminal justice practitioner types evaluate evidentiary strength. Judges, for example, rated suspect histories as stronger in implicating the suspects than did prosecutors or defense attorneys, but not police. It is possible that judges become more convinced over time that criminal background and/or gang affiliation of suspects makes them more likely guilty than do other groups, suggests greater skepticism on their part about the ability of individuals to change. Interestingly, police and judges rate the physical evidence higher than defense attorneys when a photo array is present. This is probably because defense attorneys generally benefit their clients more from knowing that photo arrays are unreliable, and are therefore less likely to, for example, allow a suspect pick to increase their ratings of physical evidence. And, judges rate the physical evidence more strongly than the prosecutors when no photo arrays were provided to either group, perhaps suggesting the prosecutors' are more influenced by photo arrays.

Perhaps surprisingly, police rated suspect statements lower than did prosecutors or defense attorneys. When an ID was present, prosecutors rated the witness credibility as higher. With regard to victims, defense attorneys tended to rate the victim's credibility as higher than the police when no photo array information was provided. This may indicate that prosecutors benefit more than others from believing the witnesses. Police officers also appear more skeptical with regard to the evidentiary value of suspect statements than defense or prosecutors; indeed, they rated witness characteristics weaker than did judges or defense attorneys. These findings are perhaps attributable to their general cultural tendency toward skepticism.

Extensive scientific findings over the past four decades on eyewitness unreliability, despite our best efforts to minimize errors and improve reliability of administrative procedures, and inclusive of the findings presented herein, suggest a different course for the future. The fact that photo arrays in this study did not add to the case's overall interpreted evidentiary strength (except in cases that already had particularly strong evidence), or the strength of any specific category of evidence, suggest that continued use of lineup/photo array procedures may not actually increase the ability to detect truth or sufficiently improve the evidentiary basis for cases beyond that provided by other evidence. Scientists have not heretofore examined the validity of eyewitness identification as an indicator of ground truth, despite the aforementioned, well documented problem of unreliability and misidentification.

Our study provides some reason for the criminal justice community to question whether or how the use of photo arrays benefits justice. In light of two facts – many people have been exonerated by DNA evidence and the primary cause of the wrongful convictions appears to have been the exclusion of the actual perpetrators from the lineups — it is important to consider both the relative importance and utility of lineups in achieving justice.

Certainly, many changes have occurred in prosecutors' offices (and probably many police departments) regarding the use of eyewitness id evidence over the past few decades, often requiring significant corroboration of a suspect pick in lineup procedures. In a vast majority of wrongful conviction cases (where the DNA or other evidence exonerated the suspect), it has been shown that the identification made by the witness or victim, was the only piece of evidence. Indeed, the Travis County District Attorney's office noted in 2012 that ID-only cases do not provide sufficient justification for prosecution. Assuming this is not likely true in every jurisdiction, despite today's knowledge of the fallibility of witness and victim identifications, the fact that eyewitness misidentifications have led to wrongful convictions should, at a minimum, raise questions about the use of eyewitness id without other strong corroborating evidence to accompany it, if not to fully re-examine its use at all as a material factor in cases. The fact that the case dispositions in the Austin cases were largely attributed to other case evidence, prevented the miscarriage of justice in both potential wrongful convictions and failing to convict the guilty.

Future studies should be conducted to determine if this effect is representative of other agencies. Additionally, an examination of the policies among prosecutors' offices should be conducted to assess the extent to which corroboration of suspect picks is necessary for prosecutions to proceed. Also, archival analysis of criminal cases should be done to assess the extent to which justifications are provided for including individuals in photo arrays.

There are some policy implications associated with these findings. Police agencies should examine the evidence about photo arrays and explore new methods for improving investigative procedures so as to not over-emphasize photo arrays, given their limitations in improving the evidentiary strength of cases. Police agencies should also train their officers and investigators regarding the limited utility and limitations of lineup procedures, as well as encourage the collection of, and emphasis on, physical evidence and other forms of identification information in order to minimize reliance on lineups (live and photographic). Police departments may also benefit from implementing policies that require clear documentation of investigators' justifications for including potential suspects in lineups and photo arrays so that they are not done prematurely or lead to an overly narrow investigative focus. Finally, participants in the criminal justice system should [continue to] emphasize corroboration when relying on lineups.

Photo Arrays in Eyewitness Identification Procedures

Over the past several decades, a significant body of research has examined the reliability and accuracy of eyewitness identification in criminal cases (Clifford & Scott, 1978; Cutler, Penrod, Stevens, & Martens, 1987; Deffenbacher, Bornstein, Penrod, & McGorty, 2004; Sporer, Penrod, Read, & Cutler, 1995; Wells, 1978), particularly as a result of early laboratory findings by Loftus and colleagues that eyewitness memory was often unreliable or inaccurate (Loftus, 1975; Loftus, Miller & Burns, 1978; Loftus & Palmer, 1974). Indeed, much of the accumulated evidence to date has been drawn from the science of cognitive and social psychology, including studies of human memory, decision-making, and social influence processes. As researchers have explored the issue, they have identified a number of factors influencing the accuracy and reliability of eyewitnesses and/or victims of crimes, many of which have been grouped under the categories of estimator and system variables. Estimator variables consist of those factors specific to the witness or the crime scene that affect eyewitness memory (e.g. features of the perpetrator or lighting at the crime scene, etc.) whereas system variables are those controllable factors within the legal system which may affect eyewitness memory, such as a police officer's questioning style, photographic versus physical line up presentation, pre-lineup instructions, etc. (Wells, Memon, & Penrod, 2006; The Innocence Project[7], 2014).

One major research focus in the scientific exploration of the accuracy and reliability of eyewitness identification has been on the presentation methods used in photo arrays (a form of lineup using photographs from a variety of sources). The traditional method, known as the simultaneous presentation method, involves presenting photos as a group, typically with six

[7] The Innocence Project, as mentioned in this report, refers to the New York City-based non-profit legal organization committed to exonerating the wrongly convicted through the use of DNA testing, and to reforming the criminal justice system to prevent future injustice, affiliated with the Cardozo School of Law at Yeshiva University.

photos (informally referred to as a “six pack”) or nine photos) with pictures appearing side by side typically across a few rows. The other method, now adopted in many jurisdictions, is known as the sequential presentation method. It involves showing photos one at a time, in sequence. Substantial scientific evidence has mounted on both of these methods; however, a series of recent issues has resulted in significant controversy over which approach produces more accurate and reliable results (Mecklenburg, 2006; Mecklenburg, Bailey, & Larson, 2008; Malpass, 2006; Mickes, Flowe & Wixted, 2012; Steblay, 2011; Steblay, Dysart, Fulero & Lindsay, 2001; Wells, Steblay, & Dysart, 2011; Wixted & Mickes, 2012). One of the more recent issues is the examination of diagnosticity of various presentation methods. Much like diagnosticity ratios, the traditional method relies on measures of probative value. More recently, however, Wixted and Mickes (2012) have asserted that these forms of analysis are theoretically less relevant, and that receiver operating characteristic analysis (consistent with theories of recognition memory, particularly signal detection theory) are better equipped to address superiority effects in light of balancing false positives and true positives.

At the heart of these discussions and complications regarding simultaneous and sequential procedures is a fundamental question as to whether actual perpetrators are present or absent in the lineups, and the importance of liberating witnesses from the assumption that the perpetrator is even in the lineup (Malpass, 2006; Malpass, & Devine, 1981; Steblay, 1997; Wells et al., 2011). Indeed, Cowdery (2005) aptly noted that among the reasons for wrongful convictions in the UK, is the sometimes narrow focus of the police and prosecutors on one suspect while ignoring other suspects. Additionally, an important and defining aspect of real criminal investigations is that in most circumstances, the police in fact do not know whether the person they suspect of committing a crime is actually the person who committed it (Clark &

Tunnicliff, 2001, p. 200). In a policy review conducted by Malpass (2006), the author summed up this controversy quite succinctly noting that:

> *"The utility of simultaneous and sequential lineups is responsive to two factors external to their actual performance; the values that are placed on the various eyewitness identification outcomes and the a priori probability that the police have been able to place the actual criminal in the identification procedure"* (p. 415).

Laboratory versus Field Studies

In order to gain a greater understanding of the impact of lineup presentation methods, a variety of approaches have been implemented in laboratory settings, and to a lesser degree, field settings. It may seem that the key challenge associated with the appropriate interpretation and translation of research into practice in this area of inquiry has been our inability to gain the unique benefits of both laboratory studies (knowledge of ground truth and experimental control) and field studies (real-world stakes and nuances).

In the late 1990s and early to mid 2000s, legal and psychological scholars, policy makers, criminal justice decision makers (police, prosecutors, judges, etc.) and other skeptics alike, began to question the applicability of a body of literature that was based solely on laboratory studies to real world settings and situations. Indeed, both forms of science are valid, but each with unique benefits and challenges (Chae, 2010; Konecni & Ebbesen, 1986; Wagstaff et al., 2003; Wright & McDaid, 1996). While laboratory studies raise questions of external validity, field studies using actual police case files lack the experimental control of laboratory studies. Nevertheless, field studies represent what some have characterized as a "degree of realism and a range of variables impossible to simulate in a laboratory setting" (Tollestrup, Turtle, & Yuille, 1994).

Laboratory studies in eyewitness identification are often conducted with staged crimes where the perpetrators are always known to the researchers allowing them to test the effects of manipulated variables on ground truth or in other words, a “known” perpetrator. Lab studies also allow for significant controls to be exerted, thereby limiting the influence of other factors on the manipulated variable(s). Many legal practitioners and researchers themselves, however, have justifiably raised questions regarding the limitations of laboratory studies, including their inability to mimic reality in many respects, the low stakes associated with wrong choices, and the application of those findings to real-world settings. Indeed, as Tollestrup et al. (1994) noted some 20 years ago, the historical over reliance on findings from laboratory research has left the field of eyewitness memory open to challenges of external validity (Clark & Tunnicliff, 2001; McCloskey & Egeth, 1983; McKenna, Treadway, & McCloskey, 1992; Yuille & Wells, 1991).

Laboratory research testing of the sequential versus simultaneous debate has also come under fire for a number of other reasons, including the same foils design issue (Clark & Tunnicliff, 2001), the target to foils shift phenomenon (Clark & Davey, 2005), bias in single versus double blind administration (Greathouse & Kovera, 2009), and a number of other issues related to the building and discrediting of evidence speaking to either a sequential lineup superiority effect or a simultaneous lineup superiority effect (Carlson, 2008; Lindsay & Wells, 1985; Malpass, 2006; Memon & Gabbert, 2003; Mickes et al., 2012; Steblay, Dysart, & Wells, 2011). Nevertheless, while the bulk of the scientific literature informing the sequential versus simultaneous debate continues to come largely from laboratory settings or meta analysis of these laboratory findings, questions of external validity and applicability to the real world have remained to some degree.

Even though field studies are limited in many respects as described previously, they also present an opportunity to test variables of interest in real-world settings and thereby, an opportunity to overcome the problems inherent in laboratory studies (Schacter et al., 2008;

Wagstaff et al., 2003; Wells et al., 2000; Wells et al., 2006), despite their inability to establish ground truth in studies of eyewitness identification, an issue that will be addressed in the present study. Nevertheless, until very recently there had not been a robust test of the sequential versus simultaneous procedures in the field.

Test of the Sequential versus Simultaneous Method

In response to calls for a more robust field study, the American Judicature Society (AJS) implemented the *"Eyewitness Identification Field Studies"* (hereafter referred to as the *AJS' EWID Field Studies*) designed to compare sequential and simultaneous presentation methods in multiple field sites (Wells, et al., 2011). That study consisted of a series of controlled field experiments and was informed both by project partners[8] who collaborated with the AJS, as well as the *Greensboro Protocols*[9] to establish greater scientific control in the course of conducting a field study. The *AJS' EWID Field Studies* represented perhaps the first time social scientists, prosecutors, criminal defense attorneys and the law enforcement community had come together in a systematic fashion to inform the testing of new ways to improve the reliability and credibility of eyewitness evidence. Wells et al. (2011) implemented that experiment in four sites: Charlotte-Mecklenburg County, North Carolina; Tucson, Arizona; San Diego, California; and Austin (Travis County), Texas.

In the *AJS Field Studies* test of the sequential versus simultaneous variable, all factors other than the presentation method were held constant[10] following the guidelines established in the Greensboro Protocols. In other words, at the field sites, the protocol required standardized

[8] Center for Problem Oriented Policing, Police Foundation, and Center for Modern Forensic Practice at the John Jay College of Criminal Justice (no longer in operation).

[9] The *Greensboro Protocols* represent a series of reforms for guiding the design and rigor of future field studies as decided by a group of lawyers, prosecutors, police investigators, and leading behavioral scientists in the field of eyewitness identification that convened in Greensboro, NC in September 2006.

[10] With some exceptions regarding blinded versus double blind procedures, see Wells, et al. (2011).

instructions administered via a laptop presentation mode, ensured that all lineup administrations were double blind[11], and also required the collection of confidence statements. The lineup presentation method itself – sequential versus simultaneous – was randomly assigned by computer for each lineup immediately prior to viewing; therefore, it (the treatment condition) was also blind to lineup administrators. While sample sizes were substantially lower in three of the four sites as compared to that of Austin, Texas, there were different reasons for this.[12] Nevertheless, as planned, data were combined across sites to increase the overall sample size, increase statistical power, and to create greater ability to generalize findings regionally and nationally.

The investigators in the *AJS' EWID Field Studies* found that the sequential procedure yielded a slightly higher rate of identification of the suspect than for the simultaneous condition (Wells et al., 2011) although this difference did not reach statistical significance. However, in analyzing the filler (eyewitness pick of a "known innocent") identification rates, the researchers found that the simultaneous condition yielded a much higher rate of filler picks than did the sequential condition (18.1% for simultaneous compared to 12.2% for sequential), a finding that was indeed statistically significant, and which represented almost a 50% higher rate of false identifications for the simultaneous presentation method. The research team also found that there were no reductions in identifications of suspects using the sequential procedure, despite the fact that its critics have been concerned about reductions in accurate identifications. Wells, et al.

[11] A double blind condition is one in which the lineup administrator does not know who the suspect is. In the AJS field studies, a lineup was deemed not to be double-blind if the administrator acknowledged knowing the image of the suspect, the case detective was the administrator, or the detective commented in the record that the lineup was not performed double blind.

[12] In NC, state law changed during the experiment requiring all agencies to use double-blind sequential methods. In Tucson, AZ, data were being collected as part of a related study conducted by Nancy Steblay in collaboration with Lisa Judge of the Tucson City Prosecutor's Office and the methods and sample sizes were deemed appropriate for inclusion in the AJS Field Studies. In San Diego, CA low participation resulted from compatibility between the photo data bank and the lineup presentation software that significantly impeded full data collection.

(2011) acknowledged, however, that laboratory studies have generally shown a reduction in suspect picks using a sequential procedure, adding that the fact their studies did not show the same effect could be explained by differences in the protocols used in their study not present in laboratory studies (e.g., a "not sure" option was provided, witnesses were permitted to review the lineup a second time – a "second lap," and others).

Wells and colleagues (2011) noted that: "later articles will continue to extract additional new findings from this data set" (p. 17). While the release of additional findings has not yet occurred, in this follow-up study we do track the cases to examine the relationship between presentation methods and case outcomes, if any.

STUDY ONE:

Examining the Relationship between the AJS' EWID Field Studies and Case Outcomes

This observational study consists of an examination of archival data from the *AJS' EWID Field Studies* (Wells et al., 2011) and follow-up data collection on the dispositions of cases across three of the four study sites. In addition, because case dispositions may not always accurately reflect actual guilt or innocence, the second and more robust follow-up analysis will consist of an examination of the relationship between presentation methods/associated pick types and evidentiary strength ratings made by police investigators, prosecutors, defense attorneys, and judges — also herein referred to interchangeably as (key) criminal justice decision makers, raters, evaluators, or practitioners — in Austin (Travis County), TX, one of the four study sites used in the Wells et al. (2011) study. That examination is expected to provide some understanding regarding the accuracy and validity of the identifications made by witnesses and victims.

Part A: Analysis of Case Outcomes *("Tracking Case Dispositions")*

With regard to case dispositions of the *AJS' EWID Field Studies'* lineups, we address two key questions. First, what is the relationship between the lineup presentation method (sequential, simultaneous) and the case dispositions, if any? Importantly, as the *AJS' EWID Field Studies* showed, the presentation method affected the type of picks made by the witnesses/victims. While random assignment of sequential or simultaneous presentation methods allowed for an independent assessment of the impact of presentation method on the pick types in the *AJS' EWID Field Studies,*[13] the relationship between the presentation method and the case disposition is highly dependent on other factors beyond photo array results. As such, we would not expect there to be a direct relationship between the lineup presentation method and the case disposition independent of the outcome of the photo array (suspect pick, no pick, filler pick), although that relationship should be ruled out first.

The second key question related to tracking case dispositions is: What is the relationship between the pick types in the Wells et al. (2011) study, and the case dispositions? Clearly, many other case factors (other evidence and context) were likely present that may or may not have been known to the witnesses when they reviewed the photo array, and importantly, those case factors may have had more to do with the case outcomes than did the pick types (suspect, filler, or no pick). Additionally, it is possible that other case factors such as physical evidence may have outweighed lineup results thereby diminishing their importance in the case outcomes. As such, this analysis is observational, not causal in nature (i.e. it is not possible to determine which factors were most influential in the case dispositions, nor the relative weight of each in

[13] This assertion presumes that the range of case conditions and estimator variables were not different across groups. While we are not aware of whether an analysis was done to examine the extent to which the case characteristics were randomly distributed across the groups, random chance would suggest that the groups (sequential and simultaneous) had equal distributions of various case conditions and estimator variables.

determining case dispositions due to the nature of these data and the design of both the *AJS' EWID Field Studies* and the present study). Nevertheless, it was important that we establish the relationships between the presentation methods, pick types, and case dispositions.

Method

In order to ensure that the cases associated with the lineups from the *AJS' EWID Field Studies* (Wells et al., 2011) had reached disposition, we required that at least one year pass since the lineups were presented.[14] In order to assess the relationship between lineup presentation methods and case dispositions, we conducted an archival analysis with data collected from the *AJS' EWID Field Studies* (Wells et al., 2011). We received disposition data from all four sites, and while the agencies were not able to provide us with dispositions for every case, we examined the data for all but one site.[15] Because the descriptions of the outcomes varied by agency, we were only able to categorize the dispositions as having been adjudicated guilty (by plea or judgment) or not prosecuted.[16] While the results of the Wells' et al. (2011) studies indicated a causal relationship between presentation methods (sequential, simultaneous) and selection outcomes (a suspect was picked, no one was picked, or a "filler" was picked), our examination of the relationships between those variables and case dispositions is not cause and effect. Instead, we analyzed the associative relationships between lineup presentation method and case dispositions as well as that between the pick type and case dispositions.

[14] Research team members and their partners agreed that a 12-month lag would, in most cases, be sufficient for the case to reach disposition. In fact, in most cases, the follow up period was significantly greater than 12 months, and other than those for which a status could not be determined or made available to the researchers, most cases had reached disposition by the time this analysis was conducted.

[15] Dispositions from Charlotte-Mecklenburg County were not used because the study was prematurely discontinued based on changes in state law mandating the double-blind sequential procedure for lineup presentation.

[16] Due to differences in disposition types across the three sites (Austin, San Diego, and Tucson) we were only able to report on whether cases were adjudicated guilty (by plea or judgment) or not prosecuted (not referred by police, not accepted for prosecution due to insufficient evidence, or other reason).

Results

Case Dispositions across Three Study Sites

This analysis included cases from Austin, Texas, the *AJS' EWID Field Studies'* site generating the greatest amount of data from lineups,[17] as well as San Diego and Tucson. The cases for which dispositions were reported by the agencies are presented in Table 1 below. As is shown in the Table, the adjudication rate among these cases is 38%, with Austin having the highest (48%) as compared to just 25% in Tucson and 21% in San Diego[18]. The adjudication rates appear much lower than the national average of 78% in state courts, where the vast majority of all felony convictions in the U.S. occur (BJS, 2003).[19] One possible explanation for the differences in conviction rates is that our data set primarily consisted of stranger crimes (suspect

Table 1. Number of Cases with Dispositions Provided by Research Site

Agency (Study Site)	N	Adjudicated	Not Prosecuted	Total
Austin, Texas	143 (61%)	67 (47%)	76 (52%)	**143** (100%)
San Diego, California	24 (10%)	5 (21%)	19 (79%)	**24** (100%)
Tucson, Arizona	69 (29%)	17 (25%)	52 (75%)	**69** (100%)
Total	**236** (100%)	**89** (38.1%)	**147** (62.3%)	**236**

[17] We captured case dispositions for only the subset of cases in Austin (there were 455 photo arrays administered in the Wells et al., 2011 study) that were selected for the "**Evidentiary Strength Study**" and the "An Experimental Study of the Effect of Photo Arrays on Evaluations of Evidentiary Strength by Key Criminal Justice Decision Makers" (both described later in this report) so as not to over-represent the effects obtained from Austin in this dispositional analysis.

[18] Could possibly be the result of the sample size of just 29 cases.

[19] Drawn from a geographically representative sample of the approximately 3,100 counties in the U.S. in 2000.

and victim unknown to each other),[20] whereas in non-stranger crimes, the victim or witness provides the name of the perpetrator and his/her relationship to the victim, rendering a lineup unnecessary. Also, estimates from other studies more closely align with the data in our study. For example, one study demonstrated that only 37% of rape cases are prosecuted (Kilpatrick, Resnick, Ruggiero, Conoscenti, & McCauley, 2007), although our study had considerably fewer sexual assault cases (in Austin they were excluded altogether). Additionally, Garner and Maxwell (2009) reported that only 30% of domestic violence cases investigated by police led to the filing of criminal cases by prosecutors (a rate closer to that which we found).

Relationship between Photo Array Presentation Methods and Case Dispositions.

The data shown in Table 2 appear show that there are far more cases "not prosecuted" than "adjudicated." Nevertheless, the proportions of photo arrays from the Wells et al. (2011) study for which the suspects were adjudicated guilty or not prosecuted did not significantly differ by presentation method[21] (see Table 2). This finding confirmed our expectations that the method of presentation would not be directly associated with case outcomes because the dispositions are, in large part, dependent on other case factors.

Table 2. Frequencies of Case Dispositions by Lineup Presentation Methods

Disposition	**Sequential**	**Simultaneous**	**Total**
Not Prosecuted	59 (59%)	88 (64%)	**147** (62.3%)
Guilty (plea or judgment)	41 (41%)	48 (36%)	**89** (37.7%)
Total	**100** (42.4%)	**136** (57.6%)	**236** (100.0)

X^2 = .799 (1 df), *p* = n.s.

[20] We identified one exception to this in our examination; while Wells, et al. (2011) required that cases for inclusion be those where the suspect and victim were not known to each other, the officer reported that after running the lineup, it was discovered that the victim/witness knew the suspect in the case.

[21] We did not examine sentencing outcomes, however.

Relationship between Photo Array Pick Types and Case Dispositions. For this analysis, we examined the relationships between the pick types from the Wells et al. (2011) study and subsequent case dispositions. While it is not possible to say that the pick types led to the case dispositions, the data presented in Table 3 suggest strong associations between suspect picks and guilty dispositions. A greater proportion of cases in which suspects were picked were associated with guilty findings (68%), whereas for those in which fillers were picked or no picks were made, less than 27% each were associated with guilty findings.

Table 3. Frequencies of Case Dispositions by Pick Types

Case Disposition	**No Pick**	**Suspect**	**Filler**	**Total**
Not Prosecuted	89 (75.4%)	22 (31.9%)	36 (73.5%)	**147** (62.3%)
Guilty (plea or judgment)	29 (24.6%)	47 (68.1%)	13 (26.5%)	**89** (38.1%)
Total	**118** (100%)	**69** (100%)	**49** (100%)	**236** (100%)

X^2 = 38.429 (2 df) $p \leq .001$, Cramer’s V = .404 $p \leq .001$

While these observational data demonstrate a strong association (Cramer’s V = .40) between lineup choice and the disposition of the case, it should be underscored that these are not cause and effect relationships. The differences in proportions of pick types from the Wells, et al. (2011) study resulted from a witness or victim viewing a lineup (in one of two presentation formats); however, the dispositions are likely to have resulted from a host of other factors. Although the relationship between suspect picks and guilt appear consistent with our expectations of justice, it may not be because of obvious or intuitive reasons, i.e. a suspect picked in the case leads to guilty findings. We cannot know, for example, which pieces of evidence were most influential in determining the case outcomes, and cannot say that the pick types caused the case dispositions when relying solely on the data from the Wells et al. (2011) study. As such, it is not clear the extent to which the pick type (e.g. suspect pick) and the associated certainty level influenced the case disposition (e.g. guilty). It would be inaccurate, then, to conclude from these data that suspect picks ***lead*** to more guilty outcomes, or that filler picks and no picks will result in cases not being prosecuted in any given case. Other plausible explanations may exist and whenever there are other hypotheses about or evidence in the case beyond the photo array (as may be more typical today than 30 years ago), there could be other factors that outweigh the photo lineup result. As such, not only are the findings above not cause and effect, they do not provide information about the potential impact of other factors on case dispositions (the *why?*).

Furthermore, in the 32 percent of cases where suspects were picked but the cases were not adjudicated, there may be any number of plausible explanations for the outcomes as follow: a) even though the suspect was picked, the witness, victim, and/or police got it wrong (the lineup was absent the actual perpetrator) for a variety of reasons, or b) the suspect may actually be guilty, but there was no plea reached or the case could not be successfully prosecuted due to insufficient corroborating evidence, unwillingness of the victim/witness to testify, the exclusion of inculpatory evidence, etc.

Similarly, the fact that in almost 30 percent of cases with filler picks, the suspects were nevertheless adjudicated guilty could have resulted from many other factors including: a) even though a filler was picked, the actual suspect may still be guilty (the victim/witness made the ‘wrong’ pick; and b) the other evidence in the case may be so strong as to outweigh the ‘wrong pick’ by the witness/victims. Finally, similar circumstances could have explained the “no pick” adjudications, all suggesting a range of influences or errors that could stem from the: a) victims and/or witnesses (inability to recall events, not getting a good enough "look" at the perpetrator, unwillingness to testify, etc.); b) police (e.g., missing the actual perpetrator when constructing the lineup, feeling pressured to close cases, failing to fully investigate other plausible suspects, or potentially using procedures that may be overly suggestive or fail to minimize potential biasing effects); c) the legal system (adversarial roles of prosecution and defense, the exclusionary rule and inadmissibility of evidence, failing to fully protect victims from retaliation, etc.); d) prosecutors (being keen to get a conviction, failure to fully investigate or audit police investigations, or inadvertent failure to turn over Brady material); and e) defense attorneys (being keen to get their clients off despite their dual obligation of serving justice (Espinoza & Willis-Esqueda, 2008; Freedman, 1966; Hedding, 2002), among others. It should be noted that these potential errors or biases might be rare.

Nevertheless, since the *AJS' EWID Field Studies* demonstrated a linkage between presentation methods and pick types, one might expect that the case dispositions would be related to presentation methods when considering each pick type separately. It would also stand to reason that if the pick types resulted, at least in part, from the presentation method, then the presentation method *may* also indirectly relate to the case dispositions. However, our analysis suggests that this was not the case as shown in Table 4 below.

For cases in which a suspect could not be identified (**no pick**), there were no significant differences between proportions of suspects not prosecuted or adjudicated guilty by presentation method (73% sequential vs. 78% simultaneous for <u>not prosecuted</u> and 28% sequential vs. 22% simultaneous for <u>adjudicated guilty</u>). The same was true when **fillers were picked** (64% sequential vs. 77% simultaneous for <u>not prosecuted</u> and 36% sequential vs. 23% simultaneous

Table 4: Differences in Case Disposition Proportions within Pick Type by Presentation Method

Case Dispositions				
No pick made	**Sequential (seq.)**	**Simultaneous (sim.)**	**Total**	Chi square
Not Prosecuted	39 (72.2)	50 (78.1)	89 (75.4)	X^2 = .551(1) n.s.
Guilty	15 (27.8)	14 (21.9)	29 (24.6)	
Total	54 (45.8)	64 (54.2)	**118** (100)	
Suspect was picked				
Not Prosecuted	11 (34.4)	11 (29.7)	22 (31.9)	X^2 = .170(1) n.s.
Guilty	21 (65.6)	26 (70.3)	47 (68.1)	
Total	32 (46.4)	37 (53.6)	**69** (100)	
Filler was picked				
Not Prosecuted	9 (64.3)	27 (77.1)	36 (73.5)	X^2 = .848(1) n.s
Guilty	5 (35.7)	8 (22.9)	13 (26.5)	
Total	14 (28.6)	35 (71.4)	**49** (100)	

for adjudicated guilty), although the numbers of cases (n) within cells were particularly small. Finally, even when **suspects were picked**, there were no significant differences in the proportions of suspects not prosecuted versus adjudicated guilty by presentation method (34% sequential vs. 30% simultaneous for not prosecuted. and 66% sequential vs. 70% simultaneous for adjudicated guilty). These data suggest that the presentation method of the photo array is not associated with the case dispositions for any of the types of picks made.

While there was a slightly higher rate of not prosecuted cases when no picks were made within the simultaneous condition (as compared to the sequential condition), as well as a slightly higher rate of guilty verdicts for those in the sequential method (as compared to the simultaneous condition), these differences were not statistically significant. In addition, while there appears to be a slightly higher rate of convictions for those in which suspects were picked in the simultaneous method than as compared to the sequential method, this difference was also not statistically significant.

The many potential influences on case dispositions underscore the fact that case dispositions are not necessarily strong and conclusive representations of actual guilt; indeed, many guilty suspects do not end up being convicted for a variety of aforementioned reasons, and some innocent suspects get convicted (as we know from DNA and other case exonerations or subsequent findings of innocence). Due to these potential errors in the criminal justice system, the next section examines the use of a proxy for actual guilt or innocence (evidentiary strength as evaluated by police investigators, prosecutors, defense attorneys, and judges, independently and collectively) not tainted by some of the factors unrelated to actual guilt or innocence described above including failure of witnesses to testify, improperly obtained evidence, witness intimidation, the quality of legal counsel, statements made to defendants by police and prosecutors, despite actual guilt or innocence.

Part B: Analysis of Case Outcomes (*"The Evidentiary Strength Study*")

It has been argued here that case dispositions may not always be reflective of ground truth about guilt or innocence of suspects in cases, and certainly not in perpetrator-absent lineups. As such, in this part of the case outcome analysis, we assess the accuracy and validity of the photo array decisions made by witnesses and victims in the *AJS' EWID Field Studies* by relying on a proxy of ground truth – evidentiary strength and its relationship to case dispositions from the Wells et al. (2011) study.

Approximating Ground Truth

Despite the increased rigor of the methodological design afforded to the *AJS' EWID Field Studies* by the Greensboro Protocol, the inability to know ground truth about whether or not the suspect was indeed the perpetrator of the crime remained a central concern for the *AJS' EWID Field Studies* team and partners. Indeed, it has been demonstrated that witnesses/victims frequently selected a "known innocent" person in perpetrator-absent lineups in a laboratory study conducted by Finklea & Ebbesen (2007) suggesting a bias to make a selection in the laboratory setting. However, the realities of actual cases are often far more nuanced than they are in laboratory studies; hence, the argument that field studies are optimal in assessing outcomes in real world settings is weakened by this key limitation.

While it is true that field studies provide for "known innocents" as fillers in photo arrays, there is no way to validate whether the suspect identified by police is the actual perpetrator and thus, no way to know ground truth about the guilt or innocence of the suspects, saving only for conclusive DNA evidence which is available in only an estimated 10-20% of actual cases.[22] And despite the fact that it is believed DNA serves as "proof positive" of actual guilt or innocence, many cases contain DNA of the suspect but not the victim, or conversely, the victim but not the suspect rendering it not always conclusive. Additionally, researchers showed that among jurors who demonstrated a higher level of pro-prosecution bias, they overestimated the weight of weak DNA evidence (Smith & Bull, 2012). Finklea and Ebbesen (2007) found that when DNA was used as a validity check for ground truth, witnesses were inaccurate in just 5% of cases, and among the 95% of cases in which witnesses were accurate, two thirds of those were for sexual assaults. This higher accuracy rate may result from prolonged exposure in sexual assault cases

[22] See Myrna Raeder, J.D., L.L.M. (2012). Overturning wrongful convictions and compensating exonerees in Springer encyclopedia of criminology and criminal justice http://www.springer.com/social+sciences/criminology/book/978-1-4614-5689-6?detailsPage=authorsAndEditors

than in other briefer encounters such as purse snatchings, robberies, or assaults, or the availability of DNA in sex crimes.

Because we cannot be sure if the actual perpetrator is in the lineup in real world cases, the present study included the development and inclusion of a proxy for ground truth – strength of evidence (as evaluated by key criminal justice decision makers) – that could be used in all case types, not just sexual assaults or those with DNA evidence. While the idea of a proxy for ground truth is not novel, research on proxies for ground truth are scant, despite the fact that their use may be quite effective for overcoming the limitations of field studies. As a result, researchers continue to routinely call for more in-depth research examining the role of strength of evidence, and its measurement in decision-making (see e.g., Adams, 1983; Smith & Bull, 2012; Wells et al., 2006). While ratings of evidentiary strength have more traditionally been used by prosecutors to prioritize cases for prosecution, their use in approximating ground truth in eyewitness identification validation is very limited.

In one known study, Behrman and Davey (2001) attempted to overcome the limitation that ground truth is typically not readily available by examining eyewitness identification and connecting it to evidentiary strength. Following in the footsteps of Tollestrup et al. (1994)[23], these researchers attempted to alleviate the problem of perpetrator-absent lineups through the use of various categories of extrinsic incriminating evidence. Using their own system of categorization of case factors that had minimal or substantial probative value, these researchers examined 271 actual police cases for suspect identifications. In the archival analysis of these cases, suspect identifications were assessed for three different levels of extrinsic evidence that the criminal justice practitioners themselves defined; no extrinsic evidence (no evidence was

[23] Tollestrup et al. (1994) asserted that generating various evidentiary levels provided a solution to the problem of a certain unknown percentage of real world lineups not including the perpetrator of the crime.

recorded), evidence of minimal probative value (evidence was incriminating but not particularly strong), and evidence of substantial probative value (evidence was highly incriminating and thus strong). This particular categorization scheme was one of the limitations of their approach. Despite the rating category of "no evidence," the rating scale was essentially dichotomous in that evidence was characterized as either minimally probative or substantially probative, and therefore lacked variation, precision, and utility for other purposes. Indeed, as noted by Gould, Carrano, Leo and Young (2012) in a recent study, the quality of cases with extrinsic evidence used by Behrman and Davey (2001) varied drastically between those deemed by the researchers to have minimal or substantial probative value (p. 51). The assertion by Gould et al. (2012) infers both a lack of relative ability to discriminate among levels of evidentiary quality (i.e. beyond just "weak" or "strong") and utility of the Behrman and Davey (2001) categorization scheme, which is not likely reflective of actual cases. Another key limitation of that categorization scheme was that because it was developed by the researchers, as well as used by them in the rating of cases, it lacked necessary content-oriented validation from other experts.

DNA as a measure of eyewitness accuracy and approximation of ground truth. It has often been overlooked that evidence often plays a limited role in predicting judicial outcomes (see, e.g. LaFree, 1989; Peterson, Ryan, Houlden, & Mihalgovic, 1987) and that many extralegal factors influence the criminal justice process (Adams, 1983; Alderden & Ullman, 2012; Peterson et al., 1987), leading some to assert that "evidence plays an important but far from exclusive role in predicting judicial outcomes" (Peterson et al., 1987). While in an ideal world it could be assumed that a case disposition is an objective criminal justice system variable, it is clear that it is at least partly influenced by a number of variables not related to the strength of the case. For example, case dispositions may be influenced by legal realities described in the previous section (e.g. the exclusionary rule, the inability or unwillingness of witnesses to testify, insufficient representation, the plea bargaining process, etc.). As such, it can be argued that rated case strength is a suitable and potentially more accurate proxy for ground truth than case disposition given that it can be determined for all cases regardless of other influences on the case disposition.

In rating case strength, however, there is a wide range of evidence types potentially available, and some are better than others. In the literature on wrongful convictions, mistaken eyewitness identification is often cited as the most common culprit (Connors, Lundregan, Miller, & McEwen, 1996; Gross & Shaffer, 2012; Wells et al., 1998). Indeed, in nearly 75% of wrongful convictions overturned through DNA testing (312 as of December 31, 2013 according to the Innocence Project), mistaken eyewitness identification was found to have played a prominent role.[24] This has resulted in a vast body of research focused on identifying best practices for the collection and handling of eyewitness evidence at the investigator level (NIJ, 1999) and also on the evaluation of the strength and accuracy of eyewitness evidence in the courtroom, which has largely revealed a limited understanding of the complexities associated with eyewitness testimony by law enforcement, attorneys, judges, jurors, and the lay public (Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2006; Devenport, Penrod, & Cutler, 1997; Magnussen, Melinder, Stridbeck, & Raja, 2010).

In response to the controversy surrounding mistaken eyewitness identification, and its contribution to wrongful convictions, Finklea and Ebbesen (2007) set out to determine the true rate of eyewitness accuracy using DNA to establish ground truth regarding guilt or innocence of

[24] Although other contributing causes include not yet validated or improper forensics, false confessions/admissions, forensic science misconduct, government misconduct, unreliable informants and bad lawyering, The Innocence Project has identified eyewitness misidentification as the greatest single cause in wrongful convictions nationwide. For more information see http://www.innocenceproject.org

a defendant.[25] They found that in the sample of target present lineups (n = 173), witnesses were accurate in selecting the suspect 90.2% of the time (they rejected the lineup 8.7% of the time and selected a foil in 1.1% of the lineups). In the sample of target absent lineups (n = 10), the witnesses selected a suspect 100% of the time (even though absent from the lineup). Despite the high rates of eyewitness identification accuracy noted by the authors, perhaps the more interesting finding is the fact that in target absent lineups, witnesses still selected a suspect 100% of the time, indicating a biasing effect of simply having a lineup and suggesting that suspect pick rates are higher for lineups in which the police get it wrong.

Indeed, while we may be certain that the identification of a known innocent filler is clearly an error in the field, we can never be 100 percent sure that the identification of the suspect is correct, barring perhaps those cases with conclusive DNA evidence (and even then, there are exceptions). Although DNA would offer the most promise here, and has served as the basis for reversing convictions (312 DNA exonerations as of December 2013),[26] the proportion of cases for which DNA is present, collected, and processed, is estimated to be only about 10% and there are numerous reasons for this (Gardner & Anderson, 2012; Pratt, Gaffney, Lovrich, & Johnson, 2006; Samuels, Davies, & Pope, 2013). While Finklea and Ebbesen (2007) found that witnesses were accurate 95% of the time in selecting the suspect in cases with DNA evidence available (by using it as a proxy for assessing the accuracy of identifications) it should be noted that approximately two-thirds of the cases were rapes or other sex crimes; the nature of which is fundamentally different from say, a purse snatching on the street. Sex crimes, characterized by close contact between the perpetrator and victim, increase the probability that DNA will be transferred, and therefore collected (although not necessarily analyzed).

[25] DNA was used establish whether lineups were target present or target absent, although it was not always possible resulting in a sample of 71 unknown lineups.
[26] http://www.innocenceproject.org/

Furthermore, in describing the type of DNA samples obtained, Finklea and Ebbesen (2007) also noted a full spectrum of sample quality. Indeed, in their archival analysis some of the cases had DNA evidence matching a victim but not the suspect, while some had DNA matching a suspect. Other cases had DNA matching both victim and suspect while others still had sources of evidence with inconclusive results highlighting the fact that DNA evidence is still prone to bias and misinterpretation of the strength of its evidential standard (Smith & Bull, 2012). In addition to the limited cases where DNA analysis has been conducted, it is also unclear as to whether cases with DNA are also unique in other ways that could influence the outcomes.

Few field studies on eyewitness identification procedures have attempted to overcome the limitation that ground truth is typically not available. Those that have developed a form of proxy for approximating ground truth (e.g. Behrman & Davey, 2001; Smith & Bull, 2012) acknowledge that no matter how highly incriminating or strong the evidence is, perpetrator presence in a lineup can never be assured. Finklea and Ebbesen (2007) present a very strong case for using DNA as a proxy to establish ground truth and assess the accuracy of eyewitness testimony outside of the lab but even so, the fact that DNA is present in a small proportion of criminal cases limits its usefulness as a proxy for ground truth.

Measurement and interpretation of strength of evidence by key decision makers. The findings noted by Smith and Bull (2012) regarding bias in the interpretation of weak DNA evidence echo a particular focus by the National Research Council (NRC) Committee on Forensic Sciences regarding the adversarial process of the American judicial system and the ability of its members to accurately measure and interpret evidence as noted in the following:

> *"The adversarial process relating to the admission and exclusion of scientific evidence is not suited to the task of finding "scientific truth." The judicial system is encumbered by, among other things, judges and lawyers who generally lack the*

> *scientific expertise necessary to comprehend and evaluate forensic evidence in an informed manner…." (NRC Report on Forensic Science, 2009, p. 12).*

Recently, the NRC Committee on Forensic Sciences has focused on the mounting body of evidence indicating that there are biases and inconsistencies in the examination of various types of evidence (particularly forensic evidence), as well as the interpretation of the strength (and meaningfulness) of said evidence, interpretation of which is often left to the police, prosecutors, defense attorneys, judges and juries (Kassin, Dror, & Kuckucka, 2013; NRC, 2009; Smith & Bull, 2012; Smith & Bull, 2013). Given that the vast majority of case dispositions are arrived at through plea bargaining, it is important to make sure that those responsible for making initial charges (police), and offering or accepting pleas (prosecutors, defense attorneys, and judges) are able to accurately assess the strength of evidence in cases; yet, there is reason to believe that improvement is needed in this area.

It should be noted that while the importance of multiple perspectives in informing the criminal justice system has long been recognized[27] there continues to be a wide gap between what the scientific literature says and what practitioners believe, and the influence of those beliefs in informing their practice. For example, in a recent study exploring the disclosure of forensic evidence by police, Smith and Bull (2013) surveyed 398 experienced police interviewers regarding their use of and perceptions of the strength of various types of forensic evidence. The researchers found that the vast majority of these interviewers had not received any training on how to interpret or use such forensic information; yet, despite the lack of training, the perceived strength of the forensic evidence in a case was reported to affect some participants'

[27] In 1998, the National Institute of Justice of the U.S. Department of Justice assembled a multi-disciplinary working group comprised of both researchers and practitioners (law enforcement, prosecutors and defense attorneys) to make recommendations for the collection and preservation of eyewitness evidence. Released in October 1999, the report: "*Eyewitness Evidence: A Guide for Law Enforcement*," represents one of the first attempts at building a body of reference that incorporates both scientific and practitioners perspectives.

interview strategies and more specifically, the timing of the disclosure of such evidence during an interview.

Previously, Smith, Bull, and Holliday (2011) had found that mock jurors were able to correctly identify the strength of various types of evidence when they were not presented with other case information but found that the perceived strength of that same evidence was significantly inflated when presented in the context of a criminal case ("the story model") especially when the evidence was of a weak or ambiguous nature. These findings echo those from almost two decades earlier that juror interpretations of ambiguous evidence were highly influenced by their need for cognition, the presentation order of arguments (whether pre or post evidence presentation), and the abilities of both the prosecution and defense to persuade (Kassin, Reddy, & Tulloch, 1990). With particular regard to eyewitness evidence, this mounting body of evidence has also revealed a limited understanding of the complexities associated with eyewitness memory and testimony by key decision makers within the criminal justice system (law enforcement, attorneys, judges, and jurors) and the lay public, including students (Benton et al., 2006; Boyce, Lindsay, & Brimacombe, 2008; Devenport et al., 1997; Lindholm, 2008; Magnussen et al., 2010; Pozzulo, Lemieux, Wilson, Crescini, & Girardi, 2009; Wise & Safer, 2010).

Instrument development. In order to address limitations associated with prior forms of proxies for ground truth, and evidentiary strength rating systems, researchers (Amendola & Slipka, 2009) oversaw the development of a comprehensive, more sensitive instrument that was seen as a necessary part of the research design in the experiment described herein. It was clear that such an instrument should not be developed solely by the researchers and other social scientists, but rather be fully informed by actual criminal justice decisions makers in real cases, i.e. police investigators, prosecutors, defense attorneys, and judges from a range of jurisdictions. This approach was expected to lend face validity to the instrument, and be representative of the knowledge and information held by these critical actors in the criminal justice system.

As such, the development of the instrument was iterative in nature; for example, it relied upon more than one group to establish the content of the instrument, including the categories of evidence and their respective definitions, as well as specific types of information within them (e.g. Physical Evidence would be a category and DNA evidence would be one type within that category). In addition, it included objective reference points (exemplars) representing various levels of evidentiary strength along a 5-point continuum. Those categories of evidence, specific types, and exemplars were cross-validated through an iterative process in order to establish initial content validity of the rating instrument. The final revised instrument by Amendola and Slipka, (2009)[28] is provided in **Appendix A.**

The Strength of Evidence Scale (Amendola & Slipka, 2009) was informed in part by a particularly ambitious case evaluation system used by the Bronx County District Attorney (Bronx DA) for prioritizing cases for prosecution.[29] This system, developed in the 1970s, aimed to make the subjective assessment of evidentiary strength more objective (National Center for Prosecution Management & National District Attorneys Association, 1974, hereafter referred to as the NCPM and NDAA). As one of the first measures of case strength, it evaluated cases along four key case characteristics[30]: a) the nature of the crime charged, b) the gravity of the particular

[28] Revised and completed after multiple iterations and a pilot test conducted in Charlotte-Mecklenburg County, NC.

[29] The development of the case evaluation system for the Bronx District Attorneys Office was undertaken by the National Center for Prosecution Management (NCPM) and the National District Attorneys Association (NDAA) at the request of Mario Merola, District Attorney.

[30] The nature of the crime charged was determined by the grade of the felony involved; the gravity of the particular offense was determined primarily by the extent of personal injury and property loss or damage; propensity of commit violent crime was determined via the nature of background and prior criminal record; and case strength was determined primarily by facts, circumstances and available evidence.

offense, c) the use of the defendant's criminal history to determine propensity to commit violent crime, and d) the evidentiary strength of the case (Merola, 1982 as cited in Gould et al., 2012).

A primary limitation of using the same ambitious approach to develop a rating instrument for this study, is that the Bronx County DA case evaluation system attempted to over-mathematize all aspects of cases, and that despite its rigor and sophistication given its development some 40 years ago, it failed to address the relative level of subjectivity involved when evaluations are made by people. While aptly arguing that objective standards would vastly increase the utility and reliability of these systems for prosecution (NCPM & NDAA, 1974, p. 13) the Bronx County DA report did rightfully acknowledge that case evaluation systems would never be able supplant the individual case preparation and trial expertise of the individual prosecutor (NCPM & NDAA, 1974, p. 11). In the development of the system, however, it seems that the Bronx County DA failed to adequately acknowledge the limitations of individual differences in those making the evaluations or a need for training[31] or evaluator concurrence.

After the development, piloting, and use of *The Strength of Evidence Scale* (Amendola & Slipka, 2009) for the purposes of the experiment presented herein, Smith and Bull (2012) published an article based on a measure they developed to identify pretrial attitudes affecting juror perceptions of physical evidence. With the *Forensic Evidence Evaluation Bias Scale* (FEEBS), the authors hoped to: a) inform legal policy about the admissibility of ambiguous evidence, and b) contribute to the improvement of methods of presentation and explanation of physical evidence in the courtroom (p. 799). What the authors found was that DNA evidence of a

[31] While there were initial meetings with representatives from the Bronx County DA to specify the criteria and prosecution policy that was to form the basis for referral, coding for the nature of the offense, seriousness of the offender, and decision on whether the case should be referred to the Major Offense Bureau was conducted by the Chief of the Major Offense Bureau. A procedures manual for use of the case evaluation form was prepared for staff assigned to complete and review the case evaluation form. Overall, scores represented the subjective judgment of the Major Offense Bureau personnel and prosecutorial policy.

weak evidential standard that was presented to juror participants who demonstrated a higher level of pro-prosecution bias for forensic evidence was informed by that existing bias. Indeed, despite instruction on how to objectively evaluate physical evidence, it was almost impossible to remove the influence of bias in the interpretation of evidentiary strength for these jurors, leading them to perceive weak DNA evidence as being of a higher probative value than it actually was.

While subjective decisions can be made to be more systematic and objective through the development and calibration of an instrument, it is unrealistic to assume that everyone processes, combines, and interprets information (case facts) in the same manner; indeed, it would not likely mimic reality without some form of evaluator concurrence. As such, we utilized a multi-disciplinary approach in order to ensure that the instrument reflected the perspectives of all the major decision makers in the criminal justice system (police investigators, prosecutors, defense attorneys, and judges). An additional component of implementing the rating scale, was the provision of training and a calibration process among evaluators across disciplines within the criminal justice system. Furthermore, recognizing that even with instruction it is almost impossible to completely eliminate the influence of bias on rater assessments of evidentiary strength, a consensus exercise was built into the case evaluation process to accompany use of this rating scale. In these respects, our study attempted to create a more objective process for evaluating evidentiary strength.

Indeed, in much the same way as the purpose of the Bronx DA case evaluation system, the purpose for the "*Strength of Evidence Rating Scale*" developed for these studies of "Photo Arrays in Eyewitness Identification Procedures" was to overcome the limitation of perpetrator absence by attempting to approximate ground truth. Unlike the Bronx case evaluation system which relied heavily on its mathematical quantification of all aspects of a case, or Smith and

Bull's (2012) forensic evidence bias scale focusing on jurors and utilizing fictional case scenarios, the use of an alternate proxy for ground truth that relies on actual criminal justice decision makers' evaluations of evidentiary strength in actual police cases will provide information about the relationship between case outcomes and actual evidence. As such, while the proxy for ground truth developed and administered in this study is clearly not the first attempt at coming up with such a system, it may be the first that comes from multiple criminal justice perspectives.

Initial Validation of *The Strength of Evidence Scale* (Amendola & Slipka, 2009). The *Strength of Evidence Scale* improves upon previous research of measures of case and/or evidentiary strength that lacked content-oriented validity evidence (Merola, 1982; Behrman & Davey, 2001). It does this in three respects: 1) via the generation of exemplars that serve as objective scoring anchors; 2) via the engagement of experts in the scale development to establish evidence of content-oriented validity; and 3) via the use of five more specified levels of evidential strength. In addition, the use of the instrument by other researchers has set the stage for building validity evidence. Indeed, Gould et al. (2012) characterized the Police Foundation's Strength of Evidence Scale, as being "a more nuanced, objective, and applicable tool" (p. 51).

In order to establish validity of a measurement tool, it is important that the process include at a minimum, a content-oriented approach to validation. Typical of these approaches is the establishment of items, criteria, definitions, and standards representing the domain of interest via subject matter expert consensus. In the development of that instrument, we established groups of subject matter experts to provide initial input and then gathered additional data from others. Our process was thus iterative in nature; that is, it involved many repetitive steps to ensure that experts provided input, inter-rater agreement was established, item analysis was conducted, and that the first developmental process was tested and cross-validated on two additional groups. While initial reliability evidence was established, evidence on validity must accumulate over time through multiple methods and multiple studies, especially that which is predictive in nature. As a result, it is not possible to say that an instrument is in itself valid; instead, validity evidence is accumulated over time in terms of establishing the validity of a measurement tool for particular purposes. The resulting instrument for evaluating case strength is presented in **Appendix A.** Details regarding the process for establishing evaluation criteria, conducting reliability analysis, and collection of initial data useful for establishing validity are provided elsewhere (Amendola, forthcoming).

While the instrument may appear overly complex, it is actually a very simple 5-point Likert scale where a "5" means that the evidence is particularly strong in linking to the identified suspect, and a "1" means that the evidence is exceptionally weak in linking to the identified suspect. The scale requires ratings across six categories of evidence[32] and an overall evidentiary strength rating, but appears more complex, ironically by its simplified use of exemplars representing consensus on what experts believed to best represent various points on the rating scale, as shown in **Figure 1** below. The purpose of the exemplars was to help guide the rater in determining the appropriate strength of a particular piece of evidence in the case that was being rated. The use of a Likert rating scale without this specificity would minimize its accuracy. This example comes from the final rating instrument used in Austin, and represents the second of a number of evidence types within the category of "physical evidence." The description of the type of evidence (surveillance tapes, etc. is shown on the left) followed by the five point rating scale, "anchored" by examples of the various points on the scale (based on the average values

[32] The categories of evidence consisted of physical evidence, suspect statement information, suspect history information, victim characteristics, witness characteristics, and identification information.

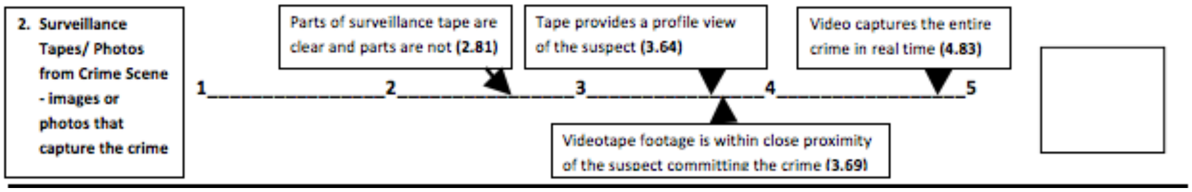

2. Surveillance Tapes/ Photos from Crime Scene - images or photos that capture the crime
Parts of surveillance tape are clear and parts are not (2.81)
Tape provides a profile view of the suspect (3.64)
Video captures the entire crime in real time (4.83)
1
2
3
4
5
Videotape footage is within close proximity of the suspect committing the crime (3.69)

Figure 1. Example of evaluation form for rating a type of evidence

provided by dozens of police investigators, prosecutors, defense attorneys, and judges). Anything in this evidence type that was weaker than these examples, would fall below the lowest exemplar (here that was rated a 2.81). The box on the right is where an evaluator would register a score if this type of evidence were available, or would have entered not applicable (n/a) if there was no evidence.

This instrument will likely increase accuracy and meaningfulness through the inclusion of: a) standardized definitions of the categories of evidence, b) descriptions of various types of evidence that would be subsumed in one of the six broader categories of evidence (for example, "surveillance tapes/photos from crime scene" as shown in above example), and c) the use of exemplars that provide reference points ("*anchors*") on the rating scale to increase objectivity. Without these, a simple 1 to 5 rating scale would be subject to potentially strong variations in the meaning or definition of different types of evidence, a potential halo effect or bias of specific evidence in interpreting other evidence, and potentially strong variations in interpretation of the strength of a specific case fact, piece of evidence, or other information where no standards had been established. Nevertheless, the utility of this rating instrument is limited by the need to provide training to case evaluators and to calibrate its use within a particular rating group.

In this study we relied on key criminal justice decision makers to assess the strength of evidence; a deviation from the past research along three lines of inquiry: 1) assessing jury decision making (e.g. Smith & Bull, 2012); 2) examining evidentiary strength for prioritization of cases for prosecution (e.g. NCPM & NDAA, 1974); and 3) relying on researchers to evaluate evidentiary strength (e.g. Behrman & Davey, 2001). This alternative approach was seen as necessary given the fact that juries are only involved in a small proportion of cases; indeed, it is estimated that no more than 5 -10% of cases are decided by jury trials (Durose & Langan, 2003; Oppel, 2011; Pastore & Maguire, 2003; Peterson et al., 1987) and the other approximately 90-95% of all cases are resolved through the plea bargaining process (Glater, 2008).

The focus on police, prosecutors, defense attorneys, and judges to evaluate evidentiary strength in this study underscores the fact that these key decision makers have perhaps the greatest effect on case dispositions through plea-bargaining. Even when pleas are not reached, these individuals have significant influence on the way in which evidence is communicated in court. While there have been a number of key studies that have focused on the role these influential actors play in the criminal justice system, these studies have largely tended to look at these actors individually rather than collectively (Alderden & Ullman, 2012; Bushway & Redlich, 2011; Frederick & Stemen, 2012; Jacoby, Mellon, Ratledge, & Turner, 1982; Peterson, Hickman, Strom & Johnson, 2013; Smith & Bull, 2013).

Method

Site Selection

The "*Evidentiary Strength Study*" was conducted in Austin (Travis County), Texas, the site in the *AJS' EWID Field Studies'* (Wells et al., 2011) from which the bulk of the data were generated. Using the "*Strength of Evidence Scale*" and the associated training and consensus building process described previously, a panel of criminal justice decision makers rated the evidentiary strength of a subset of cases in which lineups were administered in the Wells et al.

(2011) study. Austin was selected as the experimental site not only because the greatest number of photo arrays were administered there, but also because restricting our study to one robust site for the outcome evaluation allowed us to exert greater experimental control, minimizing effects associated with potential site differences, and thereby increasing statistical power.[33] Practitioners were asked to evaluate the strength of evidence across six categories of evidence (physical, suspect statement, suspect history, victim characteristics, witness characteristics, and identification information) and to give an overall assessment of the evidentiary strength of the case (see **Appendix B)**.

Case Selection

The cases were selected from the overall pool of cases in the *AJS Field Studies* in which all the experimental protocols had been followed in phase one (n= 340), thereby also minimizing the potential for error associated with potential differences in protocol adherence. The cases included were criminal and primarily made up assaults and aggravated assaults, burglaries, robberies, and thefts. A diagram of the case flow for selection in both the "Evidentiary Strength Study" and the "Experimental Study of the Influence of Photo Arrays on Police Investigators', Prosecutors', Defense Attorneys', and Judges' Evaluations of Evidentiary Strength in Criminal Cases" (presented subsequently in this report) as described here is provided in Figure 2.

[33] The three other sites were excluded from this site for a variety of reasons. First, two sites (Charlotte, NC and San Diego, CA) had limited sample sizes. In Tucson, AZ a study had been underway for some time without District Attorney involvement in the AJS studies, and prior to the establishment of a methodology for the outcome analysis. Another key reason this study's PI implemented this study in Austin alone, is that a means for controlling site variance (if any) would have been necessary at the outset (i.e. a blocked-randomized design would have allowed for better statistical control), and the limited sample sizes obtained in the Wells, et al. (2011) study across the remaining sites would make the design of a new experiment less robust. While cost would have increased if the study was done in the three remaining sites, this was not a deciding factor in the decision to implement the study in Austin alone; the sample sizes were small enough to be an inefficient (and possibly ineffective) use of resources to address the questions and a single study in Austin was therefore seen as the most reasonable, pure, and simple approach.

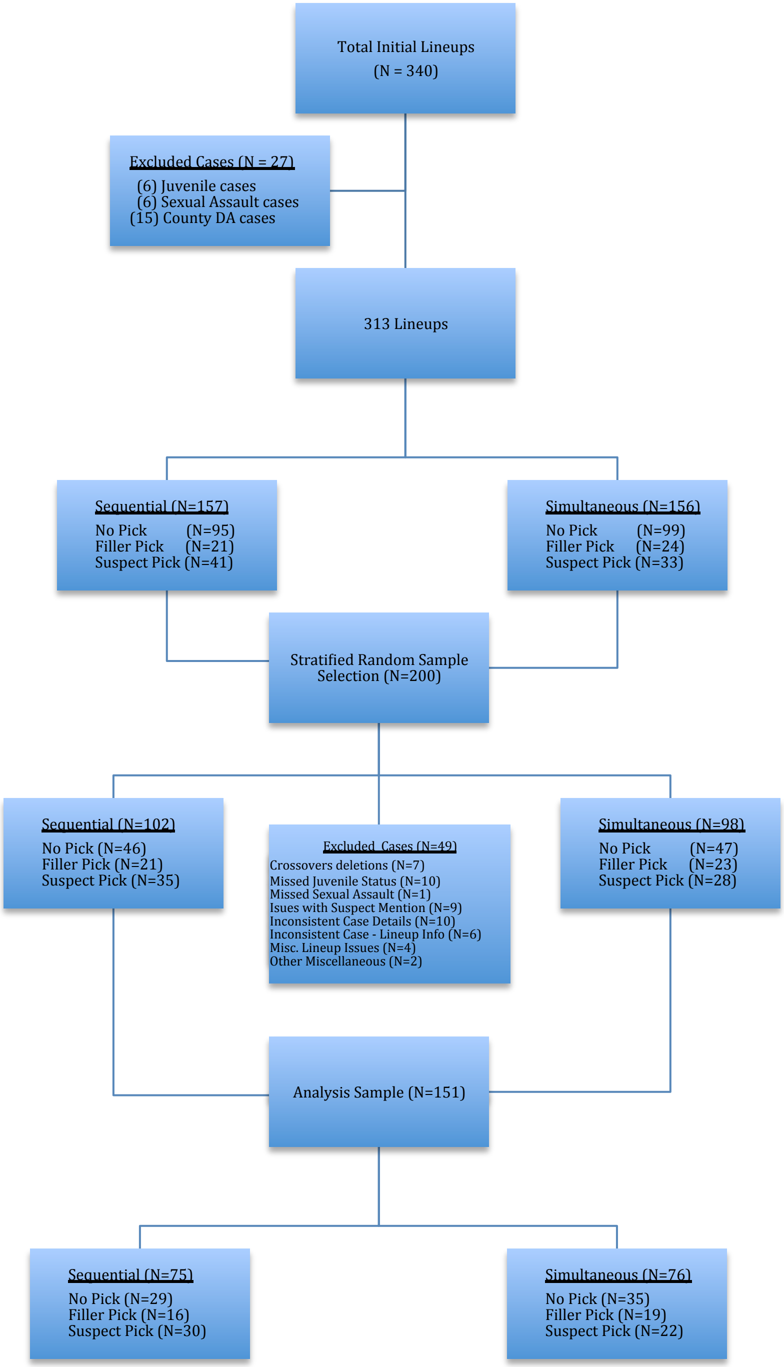


Figure 2. Case Selection Process

From the Wells et al. (2011) study lineups, we eliminated non-pristine lineups[34] resulting in an initial sample of 340 lineups. Next, due to state law in Texas, and instructions from the District Attorney's Office, we also eliminated any cases involving juvenile suspects (n = 6) and lineups associated with cases that involved sexual assault (n = 6) resulting in 328 lineups that met the criteria of the agency and research team. Additionally, we eliminated cases that were referred to the county attorney's office (n = 15), resulting in a final sample of 313 eligible lineups. We conducted a power analysis that suggested a sample of 200 lineups would be more than sufficient to ensure a high level of power for the study. In order to obtain relative balance among the pick types for the experimental study and to maximize our sample size, we selected all lineups resulting in "filler picks" (n = 45) and "suspect picks" (n = 74), due to the limited number of cases for comparison, and to ensure sufficient power for assessing the impact of suspect picks, for a total of 119 lineup cases. Of the remaining more common outcomes "no picks" (n = 194) we used a random number generator to select 93 "no pick"[35] cases stratified within sequential and simultaneous procedures resulting in 212 lineups with the expectation that additional lineups might need to be excluded during the course of the study. Indeed, upon review of case details after the initial selection, additional lineups were found to be ineligible for inclusion by research staff (i.e., juvenile involvement, sexual assault, inconsistencies in case details, suspect not mentioned in case, and a number of other reasons). As a result, the number of no pick lineups was actually reduced to 64, and further reductions in the other two groups

[34] Cases were deemed to be non pristine by Wells et al. (2011) if one or more of the following conditions applied: 1) the lineup line administrator knew which person was the suspect, hence not double blind; 2) the eyewitness knew the suspect, hence not a stranger identification; 3) the identification decision of the witness could not be clearly determined, in other words, unclear if decision was no pick, filler pick or suspect pick; and 4) the witness saw the suspect photo after the crime but before viewing the lineup, hence not a first time viewing.

[35] We attempted to achieve a total of 74 "no pick" lineups to match the number of suspect picks, but knew that approximately 20% of the cases may become ineligible. Therefore, the inclusion of 93 cases would allow a buffer so that at least 74 cases would result.

resulted in 52 suspect pick lineups, and 35 filler picks rendering a final sample in Austin of 151 lineups for evaluation, sufficient to detect medium effect sizes, according to our power analysis.

The 151 lineups selected for evaluation in this project comprised 114 distinct police cases and 139 individual suspects. Given that the lineup was the unit of analysis for both the *AJS EWID Field Studies*, and this follow-up study, any given suspect may have appeared in one or more criminal cases or other lineups but the other lineups may not have been selected for this study. Of the 114 distinct police cases included in this study, 78 of those cases had respective District Attorney (D.A.) case files for which information was collected. The 114 police cases in Phase II consisted of a variety of felony crimes (see Table 5), although more than one crime type may have been present in a given case.

Table 5. Crime Types in 114 Police Cases

Crime Types	**Total n**	**Crime Types**	**Total n**
Aggravated robbery	36	Murder	4
Assault/aggravate assault	10	Robbery	8
Burglary/home invasion	14	Robbery by assault	14
Criminal mischief	1	Robbery by threat	2
Fraud/forgery	3	Stabbing	1
Harassment	1	Theft	20
		TOTAL	114

Participants. Case evaluators were selected from a recruited pool of 26 criminal justice decision makers (10 female and 16 male). In total, the pool of evaluators to choose from on any given day was actually 36,[36] as several individuals were able to serve in more than one capacity

[36] A number were recently retired, one was currently employed in a neighboring jurisdiction, and several defense attorneys and judges were still practicing on at least a limited (part time) basis.

on different teams.[37] While in a day, we needed just eight (8) participants (two each of police investigators, prosecutors, defense attorneys, and judges), this expanded pool of evaluators afforded the research team more flexibility to construct different teams each day, given the fact that not every rater was available for all of the scheduled evaluation sessions.

Training. Training was provided to the participating criminal justice evaluators to explain how the instrument was developed, what the exemplars (rating scale anchors) represented, how they were derived, and how to rate each category of evidence independently. The latter was discussed at length in the training, as the goal of the evaluation was to minimize biases that may occur when the same piece of evidence was considered to fall into more than one category, or when the strength/weakness of any type of evidence influences the interpretation of other evidence. Because the evidence categories were established as independent with unique definitions and unique types of evidence within each category, raters were encouraged to first identify the appropriate category under which to evaluate a particular piece of evidence before assigning a score. The aforementioned training steps required a training block of approximately four to five hours.

The next step in the training was to have evaluators practice using the instrument on actual cases provided by an independent jurisdiction. This training began with a group session in which all of the case evaluators read the same case and came up with a rating. This was followed by a group discussion in which the variability in ratings was discussed in order to calibrate the ratings, so that all had an equal understanding of what constituted weak, moderate,

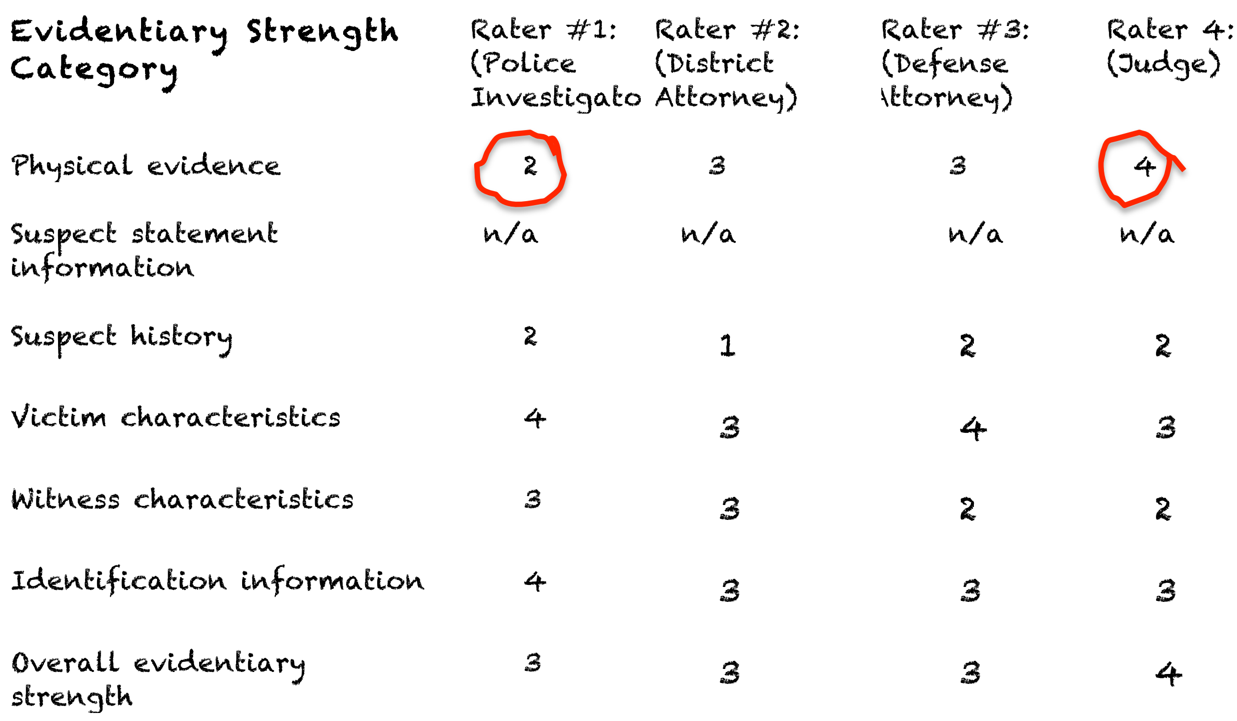
[37] Of the available raters, eight were able to conduct evaluations in more than one capacity given their previous experience serving the criminal justice system in different roles. For those individuals, they were asked to rate from the specific perspective they were assigned to that day. Those eight raters' roles consisted of: three (3) raters who were able to serve in the capacity of either district attorney or judge; two (2) raters were able to serve in the capacity of either district attorney or defense attorney; one (1) rater who was able to serve in the capacity of either defense attorney or judge; and two (2) raters who were able to serve in the capacity of either district attorney, defense attorney or judge.

and strong evidence, as well as how to arrive at a category score and overall case rating score. Suffice it to say, this process was not mathematical in nature; one piece of evidence could be so strong (e.g. DNA) that an overall case rating could be 5 even in the absence of any other evidence. This process was expected to result in a restriction of range in scores, but at the same time, a more objective rating of case strength based on agreement from the instrument development teams and the rater groups on what actually represents strong or weak evidence. The remainder of the two-day training was spent evaluating 4-5 additional cases and conducting consensus discussions so that raters could best prepare for rating actual cases in groups of four.

Once the training was completed, the research team coordinated with the case evaluators to establish teams and corresponding meeting dates to conduct the evaluations. Key criminal justice decision makers were provided with case files associated with a particular suspect associated with a lineup and case.

Study oversight and monitoring. Research team members were on site for the entire time during which ratings were conducted in the fall of 2012. The Principal Investigator and second author each oversaw the rating teams and assigned cases for each day, while a third team member ensured materials were sufficient for scoring and assisted in checking in the data at the end of each consensus session (also checking for missing data). Depending on the complexity of the case as estimated by the researchers, approximately two (2) to thirteen (13) cases were provided to evaluators in an eight-hour day.

Consensus process. After half of the day's cases had been rated by all individual evaluators (evaluators were provided with 'morning' and 'afternoon' cases) the researcher facilitated a consensus discussion that began with raters (one at a time) providing their scores for all six categories of evidence followed by their overall case strength rating (down a column) that were transferred to a white board by the researcher as shown below in **Figure 3**. The facilitator



Evidentiary Strength Category	Rater #1: (Police Investigato	Rater #2: (District Attorney)	Rater #3: (Defense Attorney)	Rater 4: (Judge)
Physical evidence	2	3	3	4
Suspect statement information	n/a	n/a	n/a	n/a
Suspect history	2	1	2	2
Victim characteristics	4	3	4	3
Witness characteristics	3	3	2	2
Identification information	4	3	3	3
Overall evidentiary strength	3	3	3	4

Figure 3. Example of ratings across raters for consensus discussion

and group reviewed the rows across, noting discrepancies of two points or more. The research protocol required that when such a discrepancy was found between any two evaluators within the team, or when the raters differed in their belief that a certain type ofevidence was present or not, a facilitated discussion among evaluators was necessary. The purpose of this was not to force raters to come up with the same scores[38]; indeed it was clear that none of these practitioners could be "bullied" by their peers. Instead, the purpose was to ensure that all raters had seen

[38] This would indeed be a problem for the research team as well, as this could have led to greater restriction of range, thereby limiting variability and the ability to detect differences in the analysis. The study already had some built-in restriction in range as a result of using a calibrated and anchored rating scale, as well as training evaluators to rate cases in one specific category.

and/or considered all evidence thoroughly because of the limited time allotted to review the case (which would not necessarily be the case if the evaluators were working in their formal capacities). Finally, discussions were allowed when raters simply wanted to discuss the case with others to clarify information or help interpret the meaning of evidence. These discussions, however, were not allowed during the individual rating process, as raters were required to retain the scores of the their evaluations independent of any discussion, or influence by other evaluators. As shown in **Figure 3**, the two ratings for physical evidence by the police investigator and the judge were two points apart, and therefore required discussion to determine the reasons for the differences. If after the discussion both raters wanted to keep their original scores, they were able to do so. However, if either one changed his/her rating, that person was required to note this on his/her final rating form, and check the box that best explained their reason for the change (see **Appendix B**).

Results

Evidentiary Strength Ratings by Presentation Method, Pick Types, and Judicial Outcomes

Operating on the assumption that ratings of evidentiary strength, as described herein, are a better proxy for ground truth than case dispositions, the examination of the relationship between those ratings and the presentation methods, pick types, and case dispositions, will allow for a more detailed assessment of the accuracy and validity of the picks made by the witnesses and victims in the Wells, et al. (2011) study. In order to validate the likely accuracy of the pick types by witnesses in that study, we address the following three questions:

1. Are ratings of evidentiary strength in this follow-up study associated with the presentation methods used in the Wells et al. (2011) study?

2. Are evidentiary strength ratings associated with case dispositions in Austin?

3. Are the overall evidentiary strength ratings in this follow-up study associated with the pick types made in the Wells et al. (2011) study?

Presentation Methods and Evidentiary Strength

In our study, criminal justice decision makers (police, prosecutors, defense attorneys, and judges) were not made aware of case dispositions or photo array presentation methods when reviewing the evidence[39] associated with suspects from the lineups by Wells et al. (2011). As such, we did not expect the presentation methods to have a strong association with the ratings of evidentiary strength. However, since the Wells et al. (2011) study indicated that presentation method impacts the type of pick being made, we did provide evaluators with information about the pick types (suspect, filler, or no pick) via the lineup software printout showing the photos on a single page (vertically with all six photos listed down the page), along with the associated names, the police-identified suspect, and which person, if any, was picked from the lineup. There was also information included in many of the officers' reports about the various lineups run and the results.[40]

As we expected, the evidentiary strength ratings did not significantly differ across the two photo array presentation methods within pick types for either the prosecuted or not prosecuted cases (see Table 6 below). While it appears that the simultaneous method is associated with higher scores when suspects are picked for the non-prosecuted cases, that difference was not significant.

[39] Any reference to presentation methods or case dispositions in the police reports or in the case file presented to evaluators was redacted or removed.

[40] The officers' narratives may have contained information about other suspects in the case as well, and/or other lineups run with other suspects and their results, even if those suspects were not part of our sample

Table 6. Differences in Evidentiary Strength by Presentation Method within Pick Types

Case Dispositions	**Sequential (seq.)**	**Simultaneous (sim.)**	**T-Test**
No pick made			
Not Prosecuted	1.92	2.12	n.s.
Guilty	4.40	4.36	n.s.
Suspect was picked			
Not Prosecuted	2.91	3.57	n.s.
Guilty	4.29	4.38	n.s.
Filler was picked			
Not Prosecuted	2.58	2.25	n.s.
Guilty	4.19	4.42	n.s.

Relationship between Evidentiary Strength Ratings and Case Dispositions

The purpose of this section is to examine the relationship between the overall evidentiary strength ratings within pick types by case dispositions to determine whether or not the stronger cases were associated with guilty outcomes and the weaker ones with suspects not being prosecuted. As shown in Table 7, the highest scoring cases (regardless of suspect pick type) on average were associated with "adjudicated guilty" outcomes in Austin. This, despite the fact that the evaluators had no idea if the cases were adjudicated guilty or not but saw all evidence in the cases. Specifically, those cases with higher evidentiary strength (mean of 4.34) were associated with the guilty verdicts and those with weaker ratings (mean of 2.50) were associated with the non-prosecuted cases, suggesting the police probably had the correct suspects, and that the prosecutors made accurate decisions; i.e. the system got it right.

Table 7. Overall Evidentiary Strength Ratings for Guilty vs. Not Adjudicated Cases within Pick Types

Photo array decision by victim/witness	Adjudicated Guilty (1 – 5)	Not Adjudicated (1 – 5)	Finding and significance
No pick made	4.38	2.03	t(60)=11.25 $p \leq .001$
Filler pick made	4.32	2.34	t(25)=6.331 $p \leq .001$
Suspect pick made	4.33	3.14	t(45)=5.407 $p \leq .001$

Overall Evidentiary Strength by Pick Types and Case Disposition

When considering the strength of cases across photo array pick types among those in which suspects were adjudicated guilty, there were no differences as shown in Table 8 below. This means that for the cases with guilty outcomes, the evidence was equally as strong across all pick types. The fact that the evidentiary strength did not vary across pick types among the adjudicated guilty cases also suggests that the suspect picks did not add anything to the interpreted case strength.

Table 8. Mean Overall Evidentiary Strength Ratings by Pick Type and Disposition

Disposition	No pick	Filler pick	Suspect Pick	F, p	Eta squared (effect size)	Post hoc comparisons with significant differences
Adjudicated Guilty	4.38	4.32	4.33	F(2,62) = .077 n.s.	--	--
Not Prosecuted	2.03	2.34	3.14	F(2,69) = 7.097 p < .01	.170 (small)	no pick vs susp p < .01

Among the non-adjudicated cases, however, the mean evidentiary strength ratings were higher when suspects were picked (3.14) as compared to those in which no pick was made (2.03). This difference is important when considering that both the instrument (scores ranging from 1 – 5) and the training provided to the evaluators were clear in terms of the meaning of scores less than three or greater than three. Specifically, scores *below* three (3) were described as "evidence and/or information that is weak" (in terms of connecting to the suspect) whereas assigning scores *greater than* three (3) meant "evidence and/or information that is strong." However, this finding may suggest that even when the other case evidence is weak, key criminal justice decision makers interpret the evidentiary strength higher when a suspect is picked. Nevertheless, the mean rating of 3.14 for the suspect-pick cases is closer to the mean of the weak, non-prosecuted cases of 2.5 (mean difference of .64) than to the mean of the strong adjudicated guilty cases of 4.34 (mean difference of 1.20), suggesting that those cases would not likely have been prosecuted anyhow.

As observational data, the findings above do not tell us whether or not the raters were influenced by the suspect picks in making their overall rating, or if the cases associated with suspect picks were also associated with stronger overall evidence at the outset. In addition, because these results were based on group means, we also cannot interpret findings from any particular cases in which the scores and outcomes were anomalous, without a qualitative, case-by-case analysis. The findings thus far do not allow us to delve deeply enough into the questions of the influence of photo arrays on key criminal justice decision makers, or the accuracy of any particular pick by witnesses or victims in those cases. Therefore, in order to gain a better understanding of the relationships between the presentation methods and pick types to the evidentiary strength ratings and case dispositions, we conducted an experiment to test the differences in evaluations of evidentiary strength in which cases were reviewed by two groups, one that had the photo array information and pick type and the other for whom that information was redacted. Conclusions and recommendations of these studies will be provided at the end of this report subsequent to the second study presented below.

STUDY TWO:

An Experimental Study of the Effect of Photo Arrays on Evaluations of Evidentiary Strength by Key Criminal Justice Decision Makers

This experimental study was conducted in order to examine more information about a subset of cases from the Wells et al. (2011) study of the impact of photo array presentation methods on the pick types (suspect pick, no pick, filler pick) made by victims and witnesses. However, the design of this experiment allowed us to examine a range of other questions relevant to the impact of photo arrays on key criminal justice decision makers (police investigators, prosecutors, defense attorneys, and judges), and thereby add to what we, as eyewitness id researchers, can contribute to the public policy.

Therefore, there were four goals of this experiment: a) to conduct a qualitative case-by-case review of The Evidentiary Strength Study (reported in the previous chapter) to examine anomalies and outliers (i.e. cases in which the evaluators with the photo array information gave scores inconsistent with the pick types);[41] b) to determine if decisions made by victims and/or witnesses in photo array procedures had a biasing effect on key criminal justice decision makers' interpretations of evidentiary strength; b) to assess the added benefit of photo arrays, if any, on interpretation of evidentiary strength by these decision makers; and d) to evaluate whether police investigators, prosecutors, defense attorneys, and judges may differ in how they interpret the relative strength of various types evidence. Therefore, the basic research questions in this study are as follow:

1. Were the pick types from the Wells et al. (2011) study seemingly accurate given the ratings of evidentiary strength?
2. In what way do suspect picks affect the ratings of evidentiary strength, if at all?

[41] These data anomalies triggered more in-depth examinations of the possible explanations for score differences as a means for validating the accuracy of the pick types and adjudication decisions from the Wells et al. (2011) study.

3. Does knowledge that a suspect was picked lead to biases in interpreting other case evidence?
4. Do police investigators, prosecutors, defense attorneys, and judges differ in their evaluations of evidence in criminal cases, and if so, how?

Method

The methods used for this experimental study were precisely the same as those used for the ***Evidentiary Strength Study,*** but with the inclusion of an experimental component. That is, the site selection, case selection, participants, training, oversight, and consensus process were almost the same as those from that study (which was, for all intents and purposes, embedded within this experiment) with some exceptions. For example, the experiment required that two teams of raters comprised of one each of the respective evaluator groups (e.g. police, prosecutor, etc.) were assigned the same cases in different workrooms and sometimes on different days. However, the two experimental conditions were informally counterbalanced within each team on any given day so as to shield the groups from the purposes of the study.

The teams of case evaluators were provided with case files stripped of case dispositions, and other necessary data, so as not to influence their determination of the case strength. Also, cases rated with the photo array did not include information about whether the case was presented sequentially or simultaneously. Specifically, researchers ensured that all participants rating "yes" cases (with photo array) regardless of presentation method, included just one page in the case file with the six photos presented vertically. This page indicated the actual suspect, and the one picked (if any) so as not to give away that a procedure was sequential or simultaneous.[42] These cases also included the officers' full reports about the photo array procedure with

[42] When a photo array was presented simultaneously in Austin, the case file included the actual picture of the photo array with suspects shown across two rows horizontally. For a sequential procedure, however, there were six separate photos. The photo array software, however, provided a report for the police file, which was the same for either condition showing six photos presented vertically and indicating the actual and selected photo, if any.

information such as why the photo array was run, and if there were multiple photo lineups, whereas that information was redacted from the other group.

Procedure

All of the photo array cases from the "Evidentiary Strength Study" (see preceding chapter) were assigned to two groups as follows:

1. The first group was provided with the cases ***inclusive of the photo array and associated pick type*** (but not the presentation method). This group's data served as the basis for the "Evidentiary Strength Study" described earlier in this report.

2. The second group examined the same cases, however, *photo array information was redacted from the case* (including case details about the photo array, the photo array printout and associated pick types).

Because all of the cases were assigned to both groups (exact matches of the cases, except for the manipulated variable "***photo array information*****")**, random assignment was unnecessary. Alternatively we randomly assigned both types of cases (those ***with*** the photo array and result and those ***without*** the photo array and result) to two groups, whose members changed each day based on their availability to participate that day. We believed this approach was both appropriate and necessary as these groups (hereafter referred to as "teams") would likely have been keyed-in to the subject matter of the study had they examined only cases with id information or without id information on any particular day, or across the entire sample of cases. Had one group received all of the cases with the photo arrays redacted, the participants would have – no doubt – questioned why none of their cases had photo arrays. Similarly had one group always received the photo array cases, they would likely be on-alert that the purpose of the study had to do with photo arrays, as in real world settings, various actors in the criminal justice regularly see cases where the case facts and procedures vary from case to case.

Part A:
Experimental Results

In this section, we present the findings of our experiment in which we examined differences in evidentiary strength ratings both for the overall case, and for each specific category of evidence (six total). The results are based primarily on comparisons of the mean evidentiary strength ratings for the two experimental groups by both presentation method and pick types. The statistical tests indicate the probability of obtaining a mean difference between the two groups by presentation methods and pick types. Our alpha level for rejection of the null hypothesis was set at $p < .05$. Missing data were excluded from the analysis on a case-by-case basis, so our *n* for any statistical tests includes all of the valid cases in the dataset. The analytic approach used was an analysis of variance (ANOVA) as a test of statistical significance that assesses whether differences in the means of the groups can lead us to reject the null hypothesis which assumes that the means of the population from which they are drawn are the same. Throughout our discussion of the findings of this study, we present eta squared (η^2)—the proportion of the total sums of squares that is accounted for by the between sums of squares—as the effect size to measure the magnitude of the differences. We relied on Cohen's (1988) criteria for interpreting the magnitude of the effects as small ($\eta^2 = 0.01$), medium ($\eta^2 = 0.059$), and large ($\eta^2 = 0.138$).

When comparing those criminal justice decision makers that had information about the photo array and its outcome to those who did not, the overall evidentiary strength ratings are not significantly different as shown in Table 9. While it appears that having knowledge of a suspect pick results in a slightly higher rating than those without, that difference was not statistically

Table 9. Mean Ratings of **Overall Evidentiary Strength** by Knowledge of Photo Array by Pick Type

Photo array decision by victim/witness	**Knowledge of photo array used and outcome across rater types**	**No knowledge of photo array across all rater types**	**Finding and significance**
No pick made	2.81	2.85	n.s
Filler pick made	2.75	2.87	n.s
Suspect pick made	3.93	3.64	n.s

significant and therefore not meaningful[43]. Indeed, it appears that the suspect-pick cases were stronger than the other pick types at the outset. Therefore, it is appropriate to conclude that other information is responsible for differences in outcomes, not specifically the fact that the suspects were picked. Similarly, there are no significant differences across raters from both conditions for filler picks or no picks, thereby not significantly hindering the case. These findings suggest that having a photo array in a case (despite the pick type made) does not add anything significant to the interpretation of the overall evidentiary strength than had no lineup been run at all.

Furthermore, as also shown in the Evidentiary Strength Study (described earlier in this report), cases with suspect picks have higher evidentiary strength ratings than those in which fillers were picked or no picks were made. While one may have drawn conclusions that those higher scores resulted from the inclusion of the suspect pick, the results from this experimental analysis disprove that hypothesis; indeed, we found that the higher evidentiary strength ratings were present despite the suspect picks. As such, the inclusion of positive photo array results

[43] This finding means that while the mean difference in the rating is one that is so small and could likely be attributed to chance.

indicates that suspect picks do not add anything substantive to the case in the eyes of key criminal justice decision makers.

Influence of Photo Arrays on Evidentiary Strength Ratings for Specific Types of Evidence

The following series of analyses explore the extent to which ratings of any one category of evidence are influenced by knowledge of a suspect pick (as well as a filler pick or no pick) in the photo array, by comparing those evaluators who had the array information to those who did not.

Physical evidence. As shown in Table 10, the physical evidence ratings are not substantially improved by having a photo array. Additionally, our results indicate that cases in which suspects were picked were associated with stronger physical evidence at the outset, as is demonstrated by the fact that those evaluators not knowing about the suspect picks, rated those cases higher than the cases with no picks or filler picks and the fact that the ratings between the two groups did not significantly differ for suspect picks. We can also conclude that when suspects are picked in photo arrays, that finding does not result in a bias in the interpretation of the physical evidence.

Table 10. Mean Ratings for Strength of **Physical Evidence** by Knowledge of Photo Array within Pick Types

Photo array decision by victim/witness	Knowledge of photo array used and outcome across all rater types	No knowledge of photo array across all rater types	Finding and significance
No pick made	2.43	2.51	n.s.
Filler pick made	2.30	2.51	n.s.
Suspect pick made	3.27	3.09	n.s

Identification Information. The across group means of strength of identification information evidence ratings for raters with and without photo array data are presented in Table 11. Again, there were no significant differences across the experimental conditions, indicating that even though photo arrays, especially those with suspect picks, are a subset of the category of identification information, they do not seem to add anything meaningful to the interpretation of that category of evidence. Importantly, while people tend to think that the pick of a suspect from a lineup is the identification information in the case, the other identifying information in these cases with suspect picks accounts for the higher scores; i.e., strong identification information was present despite inclusion of a photo array (3.62 for cases with suspect picks versus 2.91 and 2.94 respectively for no picks and filler picks). These findings are consistent with the physical evidence findings, indicating no incremental benefit of including photo arrays, even when the suspect is picked in the case. Again there are many forms of identification information linking suspects to crimes, without a need for a photo array.

Table 11. Mean Evidentiary Strength Ratings for **Identification Information** by Knowledge of Photo Array

Photo array decision by victim/witness	**Knowledge of photo array used and outcome across all rater types (1 - 5)**	**No knowledge of photo array across all rater types (1 - 5)**	**Finding and significance**
No pick made	2.75	2.91	n.s.
Filler pick made	2.75	2.94	n.s.
Suspect pick made	3.87	3.62	n.s

Suspect Statement Information. Knowledge of the photo array, even when suspects were picked, did not result in higher ratings of the suspect statement (see Table 12), indicating no

Table 12. Mean Ratings of Strength of **Suspect Statement** by Knowledge of Pick Type

Photo array decision by victim/witness	**Knowledge of photo array used and outcome across all rater types (1- 5)**	**No knowledge of photo array across all rater types (1 - 5)**	**Finding and significance**
No pick made	2.92	3.00	n.s
Filler pick made	2.56	2.97	n.s.
Suspect pick made	3.18	3.23	n.s.

biasing effect of photo arrays on interpretation of suspect statements. This finding also indicates that key criminal justice decision makers evaluate suspect statements independently; indeed, there is no meaningful difference in case ratings based on the pick types or knowledge of the photo array.

Suspect history. As shown in Table 13, there were also no significant differences across the two experimental groups when considering the evidentiary strength ratings for suspect history. Again, it appears that there is no biasing effect of suspect picks on the interpretation of suspect statements. This finding suggests that key criminal justice decision makers evaluate

Table 13. Mean Ratings of Strength of **Suspect History** by Knowledge of Pick Type

Photo array decision by victim/witness	**Knowledge of photo array used and outcome across all rater types (1- 5)**	**No knowledge of photo array across all rater types (1 - 5)**	**Finding and significance**
No pick made	2.68	2.78	n.s.
Filler pick made	2.76	2.75	n.s.
Suspect pick made	3.21	3.12	n.s.

suspect histories independently and that there is virtually no difference in them based on the pick types made or the knowledge of the array.

Victim characteristics. As shown in Table 14, there are no differences across the experimental groups with regard to evidentiary strength ratings for victim characteristics (e.g. credibility, etc.), even when a suspect is picked, indicating no biasing effect on the interpretation of victim credibility, etc. This finding also indicates that key criminal justice decision makers evaluate victim characteristics independent of pick types. There also do not appear to be substantively higher scores for victim characteristics in the cases in which suspects were picked versus those in which no picks were made or fillers were picked.

Table 14. Mean Evidentiary Strength Ratings for **Victim Characteristics** by Knowledge of Photo Array within Pick Type

Pick Type	**Knowledge of photo array used and outcome across all rater types (1- 5)**	**No knowledge of photo array across all rater types (1 - 5)**	**Finding and significance**
No pick made	3.25	3.39	n.s.
Filler pick made	3.42	3.51	n.s.
Suspect pick made	3.64	3.55	n.s.

Witness characteristics. Finally, as shown in Table 15, there were no observed differences across ratings among those with or without photo information with regard to the ratings of witness characteristics; therefore, knowledge of photo array outcomes do not seem to impact interpretations of witness characteristics suggesting that evaluators were able to independently evaluate witnesses. It also suggests that knowledge of suspect picks does not bias the interpretation of witness characteristics. As with victim characteristics, there also do not appear to be substantively higher scores for witness characteristics for the cases in which suspects were picked versus when no picks were made or fillers were picked.

Table 15. Mean Ratings of **Witness Characteristics** by Knowledge of Photo Array by Pick Type

Photo array decision by victim/witness	Knowledge of photo array used and outcome across all rater types (1- 5)	No knowledge of photo array across all rater types (1 - 5)	Finding and significance
No pick made	3.29	3.36	n.s.
Filler pick made	3.41	3.41	n.s.
Suspect pick made	3.61	3.49	n.s.

Part B:
Impact of Photo Array Knowledge on Evidentiary Strength by Case Dispositions

Among the not prosecuted cases, the evidentiary strength ratings did not vary when evaluators were provided with photo arrays and pick types made (see Table 16). This suggests suspect picks did not improve particularly weak cases, nor did the no-pick or filler-pick cases further diminish an already weak case.

Table 16. Overall Evidentiary Strength Ratings by Knowledge of Lineup Decision for Cases ***Not Prosecuted***

Photo array decision by victim/witness	Knowledge of photo array across all rater types (1- 5)	No knowledge of photo array across all rater types (1 - 5)	Finding and significance
No pick made	2.03	2.04	n.s
Filler pick made	2.34	2.56	n.s.
Suspect pick made	3.14	2.90	n.s.

Within the adjudicated guilty cases, however, the evidentiary strength ratings did vary between treatment conditions (with photo array or not) but only for cases in which suspects were picked. However, while raters with photo array data gave significantly higher overall evidentiary strength ratings when suspects were picked as compared to the group that did not know the suspects were picked (mean = 4.33 vs. 3.99, $p \leq .05$), this was only true among those cases for which the evidence was already particularly strong (see Table 17).

Table 17. Overall Evidentiary Strength Ratings by Knowledge of Lineup Decision for Cases ***Adjudicated Guilty***

Photo array decision by victim/witness	**Knowledge of photo array used and outcome across all rater types (1- 5)**	**No knowledge of photo array across all rater types (1 - 5)**	**Finding and significance**
No pick made	4.38	4.51	n.s
Filler pick made	4.32	4.35	n.s.
Suspect pick made	4.33	3.99	t(64) = 2.275, $p \leq .05$

However, when comparing mean ratings of evidentiary strength for the cases which had been adjudicated guilty or not, there were strong differences as shown in Table 18 below. Across all cases, regardless of pick type or knowledge of the photo array, the non-prosecuted cases significantly differed from those in which suspects were adjudicated guilty, suggesting that police and prosecutors in Austin made the proper decisions with regard to inclusion of the correct suspects and in terms of prosecuting those in which the cases had a significantly stronger evidentiary basis, and not prosecuting those for which the overall evidence, even when suspects were picked, was weak. Ultimately, in this study neither photo array presentation methods nor pick types differentiated between adjudications or impacted evidentiary strength ratings.

Table 18. Mean Overall Evidentiary Strength Ratings By Judicial Outcomes within Pick Type

Pick type	**With or without photo array**	**Adjudicated Guilty**	**Not Adjudicated**	**t-test and *p* value**
No pick	With photo array	4.38	2.03	*t* (60) = 11.255 p ≤ .001
	No photo array	4.51	2.04	*t* (60) = 12.623 p ≤ .001
Filler pick	With photo array	4.32	2.34	*t* (25) = 6.331 p ≤ .001
	No photo array	4.35	2.56	*t* (25) = 5.607 p ≤ .001
Suspect pick	With photo array	4.33	3.14	*t* (45) = 5.407 p ≤ .001
	No photo array	3.99	2.90	*t* (45) = 4.093 p ≤ .001

Part C:
Qualitative Case Analysis

It was important to this study that we also conduct a qualitative case analysis because the experimental findings were based on group averages, and therefore cannot tell us about what may have happened in any individual case. In addition, because errors such as wrongful convictions may only be present in a small number of cases (in fact we do not and cannot know the actual rate), we conducted an assessment between the groups with the identifications and those without to identify any cases in which scores were anomalous with the pick types (e.g. a case with a suspect pick received a lower score by those who knew the suspect was picked than for the group who had no information about a photo array).

In order to gain a better understanding of the specific case ratings across both treatment groups, we examined three possible patterns of differences: a) proportions of cases in which the scores from the group who had the photo array were ***higher*** than for those who did not, regardless of pick type; b) cases in which the scores from the group who had the photo array were ***the same as*** those who did not, regardless of pick type; and c) cases in which the scores from the group who had the photo array were ***lower*** than for those who did not, also regardless of pick type. This allowed us to identify anomalous data, e.g. a case with a suspect pick in which those with the photo array actually had lower scores than those who evaluated the case without any id information. We were then able to conduct a qualitative, case-by-case assessment of the reasons for score anomalies.

Results

The initial set of results show that there were a number of expected findings, but also a number of anomalous findings (see Table 19). If knowledge that a suspect was picked influences the evaluators in terms of the strength of the case evidence, then we would expect the scores to go up, as they did in two-thirds of the cases with suspect picks. Similarly, we would expect scores to go down when a filler was picked or no one was picked, which they did in

Table 19. Impact of Photo Array on Ratings of Overall Evidentiary Strength across Pick Types

	Suspect Pick	**Filler Pick**	**No Pick**	**Total**
Scores went up *(score with id > score without id)*	35 (67%)	8 (23%)	19 (30%)	**62 (41%)**
Scores did not change *(score with id = score without id)*	8 (15%)	11 (31%)	14 (22%)	**33 (22%)**
Scores went down *(score with id < score without id)*	9 (17%)	16 (46%)	31 (48%)	**56 (37%)**
Total	**52 (34.4%)**	**35 (23.2%)**	**64 (42.4%)**	**151 (100%)**

almost half the cases. However, if there was no influence of the picks on the cases, or there was ambiguity associated with the picks, we would expect the ratings to stay the same, which was the case in about 22% of cases, especially for filler picks.

Cases with No Score Changes When Photo Array Was Included

In twenty-two percent (22%) of cases in which the photo array and associated information were provided to the evaluators, that information had no impact on the overall evidentiary strength rating. Among these cases, 42% were for no pick cases, 33.3% were for filler picks, and 24% were for suspect picks, indicating that in almost a fourth of cases where suspects are picked, it does not substantively add to the interpretation of evidentiary strength. Surprisingly, raters did not reduce the case scores when fillers were picked in one third of the cases, perhaps indicating that the case evidence was not degraded by the knowledge that a non-suspect was picked by the victim or witness.

Cases in which Scores Decreased When the Photo Array was Included

In 37% of the cases in which evaluators had knowledge of the id, the scores went down, primarily when no picks were made or filler picks were made (55% and 29% respectively), as might be expected. This finding suggests that when lineups are run without success (the suspect is not picked), key criminal justice decision makers tend to give lower scores than those for which there was no photo array. This suggests that the inclusion of an unsuccessful photo array may cause key actors in the criminal justice system to discount the other evidence pointing to the suspect. Surprisingly, ratings went down in 16% of cases in which a suspect was picked. A detailed analysis of these nine cases revealed that for most of the cases (n = 7), the reasons for the score reduction were clear (see below). However, we could not establish a reason for one, and for the other, although the score for the suspect pick went down, both the groups gave this case a high score (over 4.0 out of 5.0). These explanations are provided below:

1. This was a sequential lineup where the witness/victim picked the suspect, but then asked to see the photos a second time, and in the second lap, the suspect was not picked.
2. In two cases, the witness/victim made two picks from the lineup, a suspect and a filler, but the data were recorded as a suspect picks.
3. A co-conspirator at the scene was identified by the victim, but the victim was not sure which one of the two had done the shooting.
4. In this case, the victim was described as "slow," and there was another suspect in the case.
5. There was a second lineup run for this suspect in which the suspect was not picked.
6. The victim was untruthful and therefore considered unreliable (the report and statement were inconsistent).

Cases in which Scores Increased When the Photo Array was Included

In just over 40% of cases in which photo arrays were provided to the case evaluators, the case scores went up, and among those, two-thirds resulted from suspects being picked, suggesting a strong influence of suspect picks on evaluations of overall case strength. Surprisingly, cases with no suspect picks accounted for about 30% of the increases in scores; yet, in about 79% of those cases, researchers could not find an explanation, suggesting that maybe just having a photo array, regardless of anyone being picked, may lead evaluators to think the case is stronger than it actually is. However, this assertion should be interpreted with caution since: a) these are relatively small numbers (n = 15), and b) we cannot determine the reasons for the increases from the case files. For the remaining 21% (n = 4) where there were explanations of why the no-picks resulted in higher scores that can be explained as follows:

1) The police report about the lineup procedure included information about the suspect's history, thereby accounting for the higher score.
2) While a filler was picked by the victim, it was determined that the victim was being untruthful, and two co-conspirators were identified by other witnesses in the case.
3) In two cases, there was a second witness who did identify the suspect.

In thirteen percent (n = 8) of cases in which fillers were picked by witnesses/victims but the scores went up, we could establish reasonable explanations for 75% (n = 6) of them as follows:

1) In two cases there were secondary lineups in which a suspect was picked, even though in one the witness/victim was only 50% sure.
2) The witness recognized the suspect as familiar, even though a filler was picked.
3) Another witness picked the other suspect who was also at the scene.
4) The officer noted in the case file about the lineup procedure, that the witness first picked the suspect but then was not sure and instead selected a filler with uncertainty.
5) Surprisingly, the officer noted that the filler picked by the witness/victim closely resembled the suspect in the lineup and thus believed that the suspect had been identified (in the simultaneous condition). Importantly, the prosecution did not move forward.

Part D: Differences Among Key Criminal Justice Decision Makers in Evaluating Evidentiary Strength

One of the purposes of the experimental study was to assess whether ratings of evidentiary strength would vary across different types of criminal justice practitioners, and whether that would vary by the independent variable (with photo array and outcome information or without). We conducted *t*-tests for the mean evidentiary strength ratings across groups. Throughout the discussion of findings in this section, we present Cohen's *d* as the effect size representing the magnitude of the differences.[44] The criteria for interpreting the magnitude of the effects are as follow: small effect (d = .20); medium effect (d = .50); and large effect (d = .80) see Cohen (1988).

For ratings of overall evidentiary strength, there was just one comparison that was significant as shown in Table 20. When interpreting the overall case strength it appears the judges had slightly higher evidentiary strength ratings than did those of defense attorneys, however the effect is considered small. The "no photo array" condition demonstrated no

[44] Cohen's *d* is the difference between sample means ($X^1 – X^2$) divided by the pooled standard deviation.

differences, so the judges appear slightly more likely to be influenced in their overall ratings when an id is included in the case.

Table 20. Group Differences in Ratings of ***Overall Evidentiary Strength***

Treatment Group	**Group 1**	**Group 2**	***p* value**	**Effect size**
With ID	Judge = 3.19	Defense = 3.02	.05	.21 (small)

In examining group differences for physical evidence, when the photo array result is provided, defense attorneys rate cases weaker than do police investigators or judges as demonstrated in Table 21. However, when considering cases where no photo array information is provided, judges rate the physical evidence stronger than do prosecutors.

Table 21: Group Differences in Ratings of ***Physical Evidence***

Treatment Group	**Group 1**	**Group 2**	***p* value**	**Effect size**
With ID	Defense = 2.61	Police = 2.76	≤ .05	.18 (small)
	Defense = 2.54	Judges = 2.72	≤ .05	.19 (small)
Without ID	Judges = 2.85	Prosecutors = 2.63	≤ .01	.28 (small)

In considering differences across rater groups with regard to the strength of the suspect statement in implicating him/her, police rated that evidence surprisingly weaker when the pick types were included in the case as compared to both defense attorneys and prosecutors (see Table 22). However when the evaluators examined cases without photo arrays, they did not differ significantly. Thus when a photo array is provided, defense and prosecutors may be slightly more likely to evaluate the suspects' statements as more incriminating.

Table 22. Group Differences in Ratings of ***Suspect Statement Evidence***

Treatment Group	Group 1	Group 2	*p* value	Effect size
With ID	Police = 2.74	Defense = 3.00	≤ .05	.25 (small)
	Police = 2.74	Prosecutors = 3.05	≤ .01	.35 (small)

When exploring the strength of the suspects' histories in being somewhat more incriminating, judges (with or without photo array information) tend to put more weight on the suspects' histories than do prosecutors or defense attorneys, as shown in Table 23.

Table 23. Group Differences in Ratings of ***Suspect History***

Treatment Group	Group 1	Group 2	*p* value	Effect size
With ID	Judges = 3.01	Defense = 2.66	≤ .001	.36 (small)
	Judges = 3.02	Prosecutors = 2.82	≤ .05	.22 (small)
Without ID	Judges = 3.08	Defense = 2.85	≤ .05	.22 (small)
	Judges = 3.11	Prosecutors = 2.82	≤ .001	.31 (small)

With regard to victim characteristics (veracity), only one difference was apparent; defense attorneys surprisingly gave more credence to victim characteristics as shown in Table 24 below and this was only true in the "no" condition (without a photo array).

Table 24. Group Differences in Ratings of ***Victim Characteristics***

Treatment Group	Group 1	Group 2	*p* value	Effect size
Without ID	Defense = 3.53	Police = 3.37	≤ .05	.19 (small)

When considering witness characteristics, however, an interesting pattern emerged. As shown in Table 25, when the photo array decision was provided to prosecutors, they rated the veracity of the witness as stronger than the police. In contrast, when considering the same case where no photo array information is provided, the police tended to rate witness characteristics as weaker than the judges or defense attorneys, suggesting possibly that police are less likely to be convinced by witnesses.

Table 25. Group Differences in Ratings of ***Witness Characteristics***

Treatment Group	**Group 1**	**Group 2**	***p* value**	**Effect size**
With ID	Prosecutors = 3.55	Police = 3.31	≤ .001	.28 (small)
Without ID	Police = 3.36	Judges = 3.52	≤ .05	.25 (small)
	Police = 3.33	Defense = 3.48	≤ .05	.17 (small

Finally, and perhaps most importantly, when considering the identification information as a whole, judges appear to consider this type of information more important (see Table 26).

Table 26. Group Differences in Ratings of ***Identification Information***

Treatment Group	**Group 1**	**Group 2**	***p* value**	**Effect size**
With ID	Judges = 3.18	Police = 3.04	≤ .05	.18 (small)
Without ID	Judges = 3.42	Police = 3.10	≤ .001	.31 (small)
	Judges = 3.42	Defense = 3.12	≤ .001	.34 (small)
	Judges = 3.42	Prosecutors = 3.20	≤ .01	.25 (small)

Indeed, when there is no photo array included in the case, they are more likely than all three groups to consider the other identifying information as critical, whereas when they have it, their rating of its strength is lower and there is just a slight difference between them and the police evaluators.

In sum, there are clearly some differences in the way different criminal justice practitioner types evaluate and interpret the strength of evidence in the cases. Most notably, judges appear to differ from the other types of raters (police, prosecutors, and defense attorneys) in that they tend to rate evidence as stronger. This is only slightly true for the overall case when compared to the defense (which is not at all surprising). However, there is an across the board affect with regard to identification information which demonstrates that judges are likely to have somewhat higher ratings than all other groups when there is no photo array provided. This is consistent with the earlier finding that judges rank identification information most important in their evaluations of cases. Higher evidentiary strength ratings for suspect histories were also common for judges compared to other groups. Perhaps judges become more convinced over time that past behavior predicts future behavior (which is indeed one of the main tenets of human behavior asserted by psychologists, based on its predictive strength). However, that impact may mean judges are more skeptical about the ability of individuals to change or less willing to consider each case's full merits when they are aware of a suspect history. It is important to note here that suspect histories as defined in this study consist not only of criminal history, but could include gang affiliation that may, in itself, be biasing. Judges also rate physical evidence higher when no photo information is included in cases, suggesting that perhaps they revert to physical evidence in absence of information about the pick type.

Other interesting findings include that when photo array pick types are present, prosecutors rate the veracity of witnesses more strongly than do police or defense attorneys. Perhaps, this is logical due to the general skepticism held by police, and that prosecutors benefit more than others from believing the witnesses. Another suggestive finding is that police and judges seem to be more convinced by the physical evidence when a photo array is present than are defense attorneys, probably because defense attorneys generally benefit their clients more from knowing that photo arrays are unreliable, and are therefore less likely to allow a suspect pick (e.g. to increase their beliefs about the physical evidence). Police officers also appear more skeptical with regard to the evidentiary value of suspect statements and witness characteristics (when no id is made), although not as much where victims are concerned (they did have a slightly lower score of victim characteristics than compared to defense), suggesting that perhaps police are skeptical when it comes to witnesses (likely based on experience), but not so much where victims are involved (as they are the ones for whom all members of the criminal justice system are working, despite the fact that the system itself does not always work to benefit them).

Limitations of the Present Study

The present study had a number of limitations as described. First and foremost, while the number of cases from the *AJS Field Studies* was substantial, the number of lineups and associated cases eligible for our study were limited. We chose to use only the 'pristine' cases as characterized by the study authors, so as to control for the effects of non-pristine lineups. Also, due to the great range of sample sizes across the sites, we selected to only use those from Austin, so as to minimize variability from potential site differences encountered in *AJS' EWID Field Studies* (Wells, et al., 2011) and furthermore, so as to not over generalize from a sample whose cases were drawn primarily from one site. Next, due to Texas law, we were not able to examine cases involving juveniles or those in which sexual assaults had occurred. While the resulting number of lineups (n = 151) was sufficient for our experiment, based on our power analysis, it was not sufficient for drawing conclusions about the potentially wrongfully convicted. While this was not a purpose of our study, it did minimize our ability to detect any pattern with regard to whether suspect picks in Phase I were associated with weak evidence. As a result, we did conduct a qualitative case analysis to identify anomalies in the cases.

Next, the fact that we restricted our experiment to Austin limits our ability to make any generalizations to other jurisdictions. This may be particularly important as in Austin, we learned that it is the policy of the Travis County District Attorney not to proceed with any cases in which the pick of a suspect from a photo array is the only evidence associated with the case. This may mean that in other jurisdictions, if this conservative standard is not applied, that the potential for getting it wrong is likely much greater. Nevertheless, photo arrays did not appear to add to the evidentiary strength except when suspects were picked in cases that were already particularly strong.

Similarly, we cannot say with certainty that the evaluators selected for our sample represent views that may be held by all persons working in the same capacity. While we did have diversity in our sample, a couple of things stand out as potentially affecting the generalizability of our findings. First, most of the evaluators we used were either retired or fairly advanced in their careers. This was necessitated by the fact that we wanted to ensure that the participants had not worked on these particular cases, and the need for availability on an ongoing basis over a period of several weeks, something not possible for current full time employed (especially prosecutors and police). The fact that most of our evaluators were very experienced in their respective fields may have provided some extra "wisdom" related to potential pitfalls or even foibles of their respective disciplines, and therefore, we cannot assume that same level of sophistication for lesser experienced police, prosecutors, or defense attorneys. Nevertheless, our findings may be conservative in terms of "getting it right," that is making appropriate and accurate determination of guilt or innocence.

The findings from this study do not allow us to determine whether the pick types from the photo array influenced the continuation of the case at any particular point in the criminal justice process; it may be that police are influenced by a suspect pick as a means for investigating the case, but that it becomes less important as other evidence accumulates. This may imply that if there is a biasing effect early on, and no other evidence is found, that the police may still feel strongly about referring it to the prosecutor. Our study used a retrospective approach in examining case strength after all of the evidence had been accumulated, and therefore we were not able to address this issue.

Another limitation of the study was that due to the need for multiple evaluators on a given day, some evaluators who had previously been in a role in the system, different than their current or most recent one (e.g. a prosecutor who has since become a judge), were asked to "step into the shoes" of their former role on any given day. While the individuals felt that they could speak from a different perspective on any given day, there is a possibility that separating out the totality of their various perspectives may not have been feasible. In particular, those that were judges were likely to have come from a different role at some point. Nevertheless, we identified some group differences despite this fact.

Additionally, the police investigators in our study were no longer serving in the capacity of an investigator (as we were not able to include those still active in the agency in an investigative role). As such, some of the investigators that had gone on to advanced roles in the agency and retired at those levels, also may not have been fully able to put themselves in the mindset of "just an investigator" given their new broader perspective. Indeed, the views by police may have been moderated by their age and experience more than would have been likely if they were currently in investigative roles.

Overall Results and Discussion

The controversy surrounding eyewitness identification procedures has been the subject of extensive scientific debate going back more than 100 years when Münsterberg's (1908) treatise, *"On the Witness Stand"* suggested the fallibility of eyewitnesses in recalling events, and was met with stark criticism from John Henry Wigmore,[45] as well as his scientific colleagues. In the 1970s, significant eyewitness identification research was initiated and some 40 plus years later, many controversies can still be found among the research and practices of eyewitness identification, despite the many contributions of science. Much of that focus has been on system and estimator variables.

Now in the 21st century, scientific advances have allowed modern society to get to "the truth" in cases with clear and convincing DNA evidence. This alone has led to hundreds of formerly convicted persons being exonerated and/or proved innocent despite serving years or even decades in prisons across the U.S. The focus on exonerating innocent persons should most certainly continue. Nevertheless, a significant proportion of these original convictions were based solely or predominantly on eyewitness identifications alone, raising significant concerns about lineup procedures. There has been extensive research on various aspects of eyewitness memory, influences on eyewitnesses during the crime, lineup and show-up procedures, and photo array methods. Indeed, several decades of research have informed our knowledge about

[45] See James M. Doyle (2005). True Witness: Cops, Courts, Science and the Battle Against Misidentification. New York: Palgrave MacMillan.

the unreliability of eyewitnesses and victims including those variables that cannot be improved upon by science and have alerted us to the need to improve the reliability of eyewitness identification and reduce bias in the administration of photo arrays and lineups. Yet surprisingly, little attention has been paid to the influence of lineups on the interpretation of case strength.

One major focus in the scientific exploration of the accuracy and reliability of eyewitness identification has been on the presentation methods used in photo arrays; the traditional simultaneous presentation method, and the sequential method, now adopted in many jurisdictions. Substantial scientific evidence has mounted on both of these methods; however, a series of recent issues has resulted in significant controversy over which approach produces more accurate and reliable results (Mecklenburg, 2006; Mecklenburg, Bailey, & Larson, 2008; Malpass, 2006; Mickes, Flowe & Wixted, 2012; Steblay, 2011; Steblay, Dysart, Fulero & Lindsay, 2001; Wells, Steblay, & Dysart, 2011; Wixted & Mickes, 2012).

The recent *AJS' EWID Field Studies* demonstrated that the sequential method of presentation resulted in significantly fewer misidentifications (Wells, et al., 2011), findings consistent with much of the accumulated evidence from laboratory studies. However, because the misidentifications were those in which known innocents (fillers or "foils") were selected, a key question emerging about this finding is the whether filler picks are representative of the more consequential error; picking a suspect when the actual perpetrator is not in the lineup. Practitioners and scientists alike have asserted and acknowledged that fillers picked from lineups are very unlikely to be prosecuted, as those individual are, with exceedingly few exceptions, "known innocents." At the same time, it is important to note that as a result of an increasing number of cases in which DNA evidence has exonerated individuals erroneously convicted by eyewitness evidence, or perhaps with other advances and/or increasing public pressure over time, many policy and practice changes have occurred in prosecutors' offices, as well as police departments regarding the unreliability of eyewitness id evidence, and its relative importance in establishing guilt or innocence. This has rendered some agencies less likely to prosecute on the basis of a victim/witness identification alone or to even put much weight on the identification as was shown in this study in Austin.

In this study, there were no significant differences in the proportions of cases adjudicated guilty versus not adjudicated based on the photo array presentation method, indicating that the presentation method may not matter in terms of case outcomes. Indeed, our study demonstrated that many other factors influenced the case outcomes.

The observational results revealed a strong association between outcomes of photo array process (i.e. lineup choices) and case dispositions in that a greater proportion of cases with guilty findings (68 percent) were associated with suspect picks, as compared to those in which no picks were made (25 percent) or fillers were picked (29 percent). Nevertheless, the pick types did not translate to inaccurate outcomes.

In our experimental study, there were no differences in evidentiary strength ratings for each pick type regardless of whether the photo array was included or not. This finding demonstrates that the pick types do not influence the interpretations of the overall case evidence or any of the specific categories of evidence. This means that although the presentation method leads to the pick type, the pick type does not, in turn, lead to the case disposition; instead, other factors account for the outcomes. However, there was an exception to this among the adjudicated cases, where those with the photo array information assigned ratings slightly higher than did those without the array (4.33 versus 3.99). However, it should be noted that the score (almost 4 out of 5) still represents a very strong case that would likely also lead to conviction. In essence, there is an impact of suspect picks on outcomes, but only for those cases that were particularly strong despite the photo array. When evaluators were not provided with information that a suspect was picked, they rated the case the same as those who were aware that the suspect was picked. In cases that were weak, the photo array had no bearing (regardless of pick type).

Our study demonstrated a strong distinction among cases with guilty findings versus those that were not prosecuted with regard to evidentiary strength. The mean evidentiary strength ratings for cases adjudicated guilty across pick types was 4.34 versus those not prosecuted of 2.50, suggesting that a case whose evidence rendered it a rating of 3.99 would be more likely to result in a conviction. Even when suspects were picked, the not-prosecuted cases averaged 3.14 whereas those adjudicated guilty averaged 4.33.

One of the key findings from this study was that the inclusion of a photo array in a case did not appear to have a significant influence on the overall ratings of evidentiary strength by key criminal justice decision makers. Indeed, for the cases in which photo lineup information (including pick type) was provided to the evaluators (the "yes" experimental condition), the evidentiary strength ratings were statistically equivalent to those from evaluators to whom no information about a photo array or its outcomes was provided (the "no" experimental condition). Surprisingly, this was true even for cases in which a suspect was picked. Furthermore, while cases in which suspects were picked were associated with higher ratings of overall evidentiary strength, this was also true in the condition where the knowledge of a suspect pick was hidden from evaluators. This indicates that the pick of a suspect is not likely to be the source of the increased ratings of overall case strength; instead, evaluators saw those cases as stronger despite the inclusion of a photo array. Essentially, this suggests that the police most likely got these suspects right (i.e. had the correct suspects), given the strength of other corroborating evidence in the cases.

The aforementioned findings suggest that the inclusion of a photo array does not provide any added benefit to the *evidentiary basis* for the case (neither strengthening nor weakening it) in the eyes of police, prosecutors, defense attorneys, or judges than would be provided without a photo array, both a serendipitous and counter-intuitive finding. When examining the relationship between pick types and evidentiary strength, there was no demonstrated bias of knowing the pick type on the ratings of evidentiary strength for no picks or filler picks; means for both groups—with and without id— were not significantly different, nor were they for suspect picks when the cases were not adjudicated. However, among adjudicated guilty suspects, the photo arrays resulting in suspect picks were rated significantly higher (mean of 4.33) than were those without the photo arrays (3.99), demonstrating a biasing effect of photo arrays when suspects were picked, but only when the cases were already particularly strong (mean of almost 4 or above on a 5-point scale). This means that for all intent and purposes, suspect picks enhance already strong cases, but have no meaningful impact on weaker cases (in this case less than 3 on a 5-point scale).[46]

This key finding does not imply, however, that photo arrays are not diagnostic among police as photo arrays may have some investigative importance; indeed, police may use them as an investigatory tool to help guide their investigations. More research is needed, however, to examine whether policies or procedures in investigative units specify the need for a documented

[46] The mean evidentiary strength ratings for cases adjudicated guilty across pick types were 4.34 vs. 2.50 for those not prosecuted, a finding that was statistically significant; $t(45) = 5.41$, $p \leq 001$. This suggests that a case with ratings of 3.99 would be more likely to result in a conviction based on these data. The means across guilty and not guilty adjudications of 4.34 and 2.50 suggest a 3.99 is more like a 4.34 (.35 mean difference) than a 2.50 (1.49 mean difference).

justification for including a potential suspect/confirmed suspect in a lineup or photo array. In this study, the researchers observed some cases in which the rationale for the inclusion of a particular suspect in a photo array was not documented in the case file, and was not readily apparent. This does not necessarily mean the investigator did not have a reason, simply, that it was not well justified in the case file. Without a documented justification, it may be that the administration of an array or lineup is premature; if a suspect is picked, it may lead an investigator in one particular direction (i.e. put too much weight on the suspect pick, despite the known problems with reliability of victims, witnesses, and the procedures themselves) while at the same time, "ruling out" another viable suspect.

Furthermore, the fact that the picks do not provide any incremental value in most cases does not mean that they would not strengthen the police and prosecutor "stories" in court cases (heard by juries). No doubt that when a victim in a courtroom points to the suspect being tried and says, "that was him," it has a profound effect on the jury (or any observer for that matter) in favor of the prosecution's case. Indeed, research with jurors/juries on the role of eyewitness id information has regularly shown to have significant biasing effects on juries (Bodenhausen, 1990; Chapadelaine & Griffin, 1997; Kerr et al., 2008). The key here is that the drama-induced impact is not necessarily reflective of ground truth. In other words, the "story" of the case may be improved when a suspect is picked from a photo array (even when other witnesses or victims do not pick the suspect or pick a filler) or worsened when fillers are picked or no one is picked (in favor of the defense). Nevertheless, key decision makers in our study were not necessarily strongly influenced by the photo array outcomes in interpreting evidence and its strength in connecting the suspect to the crime.

Indeed, an anecdotal finding by researchers in this study (during the pilot test in Charlotte and the study in Austin) was that police, prosecutors, and defense attorneys typically refer to "cases" as the entirety of the case they would present if heard by a jury, i.e. the evidence, as well as the context and prosecutor proposed theory and story of the crime and the plausible alternative explanations offered by the defense, but not necessarily the objective evidence alone. This is the reality of the adversarial system, but when no courtroom story lines are required (as is the case in the vast majority of cases that result in plea agreements), it is very important that the key decision makers are able to interpret the evidence in an objective manner in order to ensure justice.

We did not find any significant biasing effect of suspect picks on interpretation of other case evidence, again showing that photo arrays do not impact on how other evidence is interpreted. This finding also suggests that criminal justice decision makers can sufficiently separate different types of evidence, lending validity to the "*Evidentiary Strength Scale*" (Amendola & Slipka, 2009). This instrument shows promise as tool for prosecutors (and potentially others), to separate the individual case facts from the context or "story" about the case, which should lend validity to the accuracy of their interpretations of evidentiary strength in delivering just outcomes. There was evidence of strong consistency of ratings among the evaluators in this data set, suggesting that indeed, various evidentiary factors can be assigned values and relative weights (Amendola, forthcoming).

While it may be surprising that photo array outcomes did not even bias ratings of "identification information,"[47] it does underscore the fact that cases often have a range of identification information that allows them to connect suspects to the crime, rendering the result

[47] "Identification Information" was defined by the participants in the instrument development phase of the project as "*Independent corroboration of information linking the suspect to the particular incident, regardless of source.*"

of a photo array a relatively unimportant factor among them. These other factors considered in the identification information category include: a) clothing, tattoo, hairstyle and other descriptions of perpetrators; b) details of crime obtained through the investigation (e.g. finding a stolen item on a suspect, etc.); c) witness id information (e.g., detailed account of incident is given by witness consistent with other evidence, etc.); d) third party/ complainant information (e.g. pawn shop owner knows suspect and verifies he/she came in with stolen property, third party statement implicating the suspect, etc.); e) circumstances surrounding arrest (e.g. suspect hiding near crime scene, etc.); f) co-conspirator flips, thereby implicating suspect; and g) anonymously provided information.

The aforementioned finding implies that the identification information, other than the photo array and its result, may be stronger, more reliable, or more important to criminal justice decision makers, or that the other identification information may simply stand on its own without the need for a photo array, again, indicating that lineups (at least in Austin) do not add any useful information to the evidentiary basis for a case. Importantly, all types of decision makers ranked the identification information among the most important of the six categories of evidence (police investigators = #2, prosecutors, defense attorneys, and judges = #1). Similarly, all but judges ranked the physical evidence among the most important type of evidence[48] (police = #1, defense and prosecutors = #2), although judges seemed to think characteristics of witnesses and victims were more important than physical evidence.

There were some distinct differences in the way that criminal justice practitioner types evaluate the evidentiary strength. Judges, for example, rated suspect histories as stronger in implicating the suspects than did prosecutors or defense attorneys, but not police. It is possible

[48] There were six distinct categories of evidence as determined by criminal justice decision makers.

that judges become more convinced over time that criminal background and/or gang affiliation of suspects makes them more likely guilty than do other groups and suggests greater skepticism on their part about the ability of individuals to change. Interestingly, police and judges rate the physical evidence higher than defense attorneys when a photo array is present. This is probably because defense attorneys generally benefit their clients more from knowing that photo arrays are unreliable, and therefore are less likely to, for example, allow a suspect pick to increase their ratings of physical evidence. And, judges rate the physical evidence more strongly than the prosecutors when no photo arrays were provided to either group, possibly suggesting the prosecutors' are more influenced by photo arrays.

Perhaps, surprisingly, police rated suspect statements lower than did prosecutors or defense attorneys. When an ID was present, prosecutors rated the witness credibility as higher. With regard to victims, defense attorneys tended to rate the victim's credibility as higher than the police when no photo array information was provided. Perhaps, this may indicate that prosecutors benefit more than others from believing the witnesses.

Police officers also appear more skeptical with regard to the evidentiary value of suspect statements than defense or prosecutors. They rated witness characteristics weaker than did judges or defense attorneys. These findings are may be attributable to their general cultural tendency toward skepticism.

Conclusion

Given the extensive findings over the past four decades on the unreliability of eyewitnesses despite the best efforts to minimize errors and improve reliability of administrative procedures, these findings suggest potentially a different course for the future. The fact that photo arrays in this study did not add to the cases' overall interpreted evidentiary strength or the strength of any specific category of evidence, suggests that use of lineup procedures may not actually increase the ability to detect truth or substantially improve or impair the ratings of evidentiary strength of the case over what the other case evidence already provides.

Yet, scientists have avoided inquiries that would establish the validity of relying upon eyewitness identification as an indicator of ground truth, despite the well-established, aforementioned problems of unreliability and misidentification. Our study provides some reason for the criminal justice community to question whether or how the use of photo arrays contributes to meting out justice. In light of the fact that many individuals have been exonerated by DNA evidence and that the primary cause of the wrongful convictions was the exclusion of the suspect from the photo array and the ensuing wrong pick, it is important that we consider not only the relative importance of photo arrays, but also the utility of them in executing justice.

Certainly, many changes have occurred in prosecutors' offices (and probably many police departments) regarding the use of eyewitness id evidence over the past few decades, often requiring significant corroboration of a suspect pick in lineup procedures. In a vast majority of wrongful conviction cases (where the DNA or other evidence exonerated the suspect), it has been shown that the identification made by the witness or victim, was the only piece of evidence. Indeed, the Travis County District Attorney's office noted in 2012 that ID-only cases do not provide sufficient justification for prosecution. Assuming this is not likely true in every jurisdiction, despite today's knowledge of the fallibility of witness and victim identifications, the fact that eyewitness misidentifications have led to wrongful convictions should, at a minimum, raise questions about the use of eyewitness id without other strong corroborating evidence to accompany it, if not to fully re-examine its use at all as a material factor in cases. The fact that the case dispositions in the Austin cases were largely attributed to other case evidence prevented the miscarriage of justice in both potential wrongful convictions and failing to convict the guilty.

Future research should attempt to focus on other categories of evidence, particularly physical and forensic evidence, and the appropriate interpretation of that evidence by key criminal justice decision makers who are responsible for closing at least 90% of cases (without juries). One potential implication for police departments is that they continue to explore methods for improving investigative procedures so as to reduce the emphasis placed on photo arrays, given their limitations in improving the evidentiary strength of cases. Police agencies should also train their officers and investigators regarding the limited utility and limitation of lineup procedures, and encourage the collection of and emphasis on physical evidence, and other forms of identification information in order to de-emphasize the importance of lineups as critical to a case. Additionally, police departments may benefit from implementing policies that require clear documentation (in the case file) of investigators' justifications for including potential suspects in lineups and photo arrays so that they are not done prematurely or lead to an overly narrow investigative focus.

In the two studies we presented here, we conducted a follow-up on the Wells et al. (2011) study of sequential versus simultaneous presentation methods to determine whether the presentation methods were associated with differences in quantitative (case dispositions) and qualitative outcomes (evidentiary strength ratings). However, our results demonstrated that there is not a relationship between presentation methods and case dispositions or evaluations of evidentiary strength. This finding is not surprising given the fact that, at least among these cases in Austin, there were likely many influences on the case dispositions and evidentiary strength not the least of which are appropriately-derived confessions, physical evidence, and many other forms of identifying information, etc.

Nevertheless, we found that regardless of pick type (suspect, filler, no pick), ratings of evidentiary strength were significantly higher for those cases adjudicated guilty versus not prosecuted, suggesting that prosecutors appeared to have made the right decisions. Observationally, it appeared that suspect picks were associated with higher ratings and proportions of guilty outcomes; however, our experiment showed for the cases with guilty findings that included suspect picks, that the cases were particularly strong despite the suspect picks (those not knowing that a suspect was picked rated the cases very high, despite the fact that in those cases, the suspect picks may have provided more confidence).

When evaluators did *not* look at photo arrays and pick types, the ratings for those adjudicated guilty all averaged above 3.9 on a 5-point scale. However, for those not adjudicated, average evidentiary strength ratings were below 2.5 for the no pick and filler pick cases, as compared to 3.14 for suspect picks, suggesting that the overall evidence in those suspect pick cases was not strong enough to proceed with a prosecution; yet, another indication that the prosecutors made accurate decisions.

Most importantly, our experiment showed that including a photo array in cases had no meaningful impact on the evidentiary strength value of the case as a whole, or any particular category of evidence. As such, at least for those key criminal justice decision makers we relied upon in Austin, the photo arrays neither biased the interpretation of the other evidence in the cases, nor made a meaningful difference in how the case was interpreted, suggesting that it did not add any real value to the case nor to its outcomes. Cases that were particularly strong appeared to be so, despite the photo array raising the question of the relative importance of photo arrays in getting at truth regarding guilt or innocence. This study did not, however, look at the biasing effect of lineups on jurors. Because key criminal justice decision makers such as those included in our study handle about 95% of cases out of court, there is reason to believe that at least for cases in which plea deals are reached, that photo arrays will have little impact on case dispositions arrived at without juries. Indeed, the right to have one's case heard before a "jury of its peers," does not mean that the jury will "get it right" either.

Indeed, it appears that police departments, prosecutors, defense attorneys, and judges are becoming better versed in the scientific evidence related to eyewitness identification practices, and will continue to do so. Of course, this is not likely to be the case in all agencies, depending upon their leadership, political views, or unwillingness to acknowledge the importance of scientific evidence. It is important to note that while the evaluators in our study were likely representative of those in the Travis County area, this does not mean that they (or the Travis County criminal justice system as a whole) is representative of other jurisdictions nationwide. Therefore it is important that these findings be interpreted within that context.

At the same time, the presence of photo arrays did not appear to have biasing impacts on case evaluators, suggesting that at least for these evaluators, photo arrays did not take the place of other strong evidence in prosecuting the cases. These findings together suggest *no added benefit* to the evidentiary basis of the case by inclusion of a photo array— indeed a serendipitous finding. This finding does beg the rarely asked question: Are lineups alone necessary in identifying "truth" as demanded of our justice system? The fact that today, massive advances in physical and forensic evidence have occurred, may have contributed to our findings that identification information in a case tends to stand on its own without the need for corroboration from a witness or victim picking a suspect out of a lineup.

In conducting a qualitative case analysis we did reveal some important anomalies. Specifically, our quantitative case analysis revealed few anomalies with regard to cases being strengthened by suspect picks. Importantly, in about 30% of cases where no suspect was picked and over 20% of cases in which a filler was picked, evaluators increased their ratings of evidentiary strength. While in some cases, there were good explanations for these changes (especially for the filler picks), in most of the cases it seems that just the inclusion of a photo array, regardless of its result, led case evaluators to increase their ratings. Nevertheless, as shown in the experimental findings, these increases were not statistically meaningful in terms of their magnitude, yet it does speak to some of the idiosyncrasies associated with photo arrays.

Finally, where knowledge about errors in eyewitness procedures and memory are well established, these findings do not really add to those facts. Instead, in considering actual cases in the field, it appears that many of these problems were mitigated by sound law enforcement and prosecutorial practices with regard to the limited importance of photo arrays in moving cases forward. Our findings do, however, suggest a re-consideration of the benefit(s), if any, lineups provide over and above the more objective case evidence, as our study showed no added benefit beyond the other evidence. It is perhaps true that lineups may assist officers in facilitating an investigation, but in some cases that assistance may lead to tunnel vision with regard to a particular suspect thereby eliminating the actual perpetrator (although that did not appear to be true in any of the cases we evaluated in Austin). It is also clear that lineups add a dramatic effect in courtrooms, and that they may strengthen the theories of both the prosecution or defense (due to the adversarial process), but perhaps not in terms of ground truth as to actual guilt or innocence, thereby underscoring the need for corroboration when relying on lineups and photo arrays in cases to ensure justice.

Nevertheless, dispositions in Travis County, Texas were supported by evidentiary strength ratings; indeed, while there may certainly be influences on dispositions other than truth, the system as a whole seemed to get it right in Austin. Importantly, evaluators in this study were encouraged to consider all evidence present in the cases, even if it were not to be admissible in a court proceeding, thereby adding to the validity of the *Strength of Evidence Scale* as a closer proxy of ground truth.

The results of this study are particularly meaningful for a number of reasons. Firstly, the fact that 90 – 95% of cases are settled through plea agreements rather than through jury trials, suggests that these key decision makers are responsible for interpreting evidence and in facilitating justice, this study attempted to consider their interpretation of evidentiary strength as considerably more important than the almost exclusive focus on jury decision making in past research. Indeed, these criminal justice decision makers account for the majority of criminal justice outcome decisions in the system. Secondly, the fact that these evaluators were trained in this experiment to rate evidence in a singular category, in order to avoid biasing the interpretation of other evidence, and were "checked" in this process by other team members during consensus discussions regarding individual ratings, suggests that evaluators were able to separate case facts from context in order to arrive at more scientifically sound decisions.

Thirdly, all types of decision makers (regardless of having the photo id information or not) ranked the identification information among the most important of the evidentiary categories, however, our findings suggested that the identification information stood on its own without the addition of the photo arrays, suggesting limits to their utility. Similarly, all but judges ranked the physical evidence among the two most important, whereas judges seemed to think that characteristics of witnesses and victims were more important than physical evidence. Finally, these findings are important because little attention has been paid to how prosecutors and others consider evidence in their decisions and findings. Whereas research with juries on the role of eyewitness id information has regularly been shown to have significant biasing effects on juries as well (Bodenhausen, 1990; Chapadelaine & Griffin, 1997; Kerr et al., 2008) the same was not true for the police, prosecutors, defense attorneys, and judges in our study.

Recommendations for Practice and Future Research

Given the extensive findings over the past four decades on the unreliability of eyewitnesses, despite the best efforts to minimize errors and improve reliability of administrative procedures, these findings potentially suggest a different course for the future. Future research should attempt to focus on other categories of evidence, particularly, physical and forensic evidence, and the appropriate interpretation of that evidence by key criminal justice decision makers who are responsible for closing at least 90% of cases (without juries). One potential implication for police departments is that they continue to explore methods for improving investigative procedures so as to reduce the emphasis placed on photo arrays, given their limitations in improving the evidentiary strength of cases. Police agencies should also train their officers and investigators regarding the limited utility and limitation of lineup procedures, and encourage the collection of and emphasis on physical evidence and other forms of identification information in order to de-emphasize the importance of lineups as critical to a case. Additionally, police departments may benefit from implementing policies that require clear documentation (in the case file) of investigators' justifications for including potential suspects in lineups and photo arrays so that they are not done prematurely or lead to an overly narrow investigative focus. Finally, law enforcement personnel should [continue to] emphasize corroboration when relying on photo arrays or lineups in criminal cases.

References

Adams, K. (1983). The effect of evidentiary factors on charge reduction. *Journal of Criminal Justice, 11*, 525-537. doi:10.1016/0047-2352(83)90005-3

Alderden, M.A., & Ullman, S.E. (2012). Creating a more complete and current picture: Examining police and prosecutor decision-making when processing sexual assault cases. *Violence Against Women, 18*(5), 525-551. doi:10.1177/1077801212453867

Amendola, K.L. (2014). The development of an instrument to rate evidentiary strength in criminal cases. Unpublished manuscript, Police Foundation, Washington D.C.

Amendola, K.L., & Slipka, M.G. (2009). Strength of evidence scale. Unpublished instrument, Police Foundation, Washington, DC. Contained in Appendix A of this report.

Behrman, B.W., & Davey, S.L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25*(5)*,* 475-491. doi: 10.1023/A:1012840831846

Benton, T.R., Ross, D.F., Bradshaw, E., Thomas, W.N., & Bradshaw, G.S. (2006). Eyewitness memory is still not common sense: Comparing jurors, judges and law enforcement to eyewitness experts. *Applied Cognitive Psychology, 20*(1)*,* 115-129. doi:10.1002/acp.1171

Bodenhausen, G.V. (1990). Second-guessing the jury: Stereotypic and hindsight biases in perceptions of court cases. *Journal of Applied Social Psychology, 20*(13)*,* 1112-1121. doi:10.1111/j.1559-1816.1990.tb00394.x

Boyce, M.A., Lindsay, D.S., & Brimacombe, C.A.E. (2008). Investigating investigators: Examining the impact of eyewitness identification evidence on student-investigators. *Law and Human Behavior, 32*(5), 439-453. doi:10.1007/s10979-007-9125-5

Bushway, S.D., & Redlich, A.D. (2012). Is plea bargaining in the "shadow of the trial" a mirage? *Journal of Quantitative Criminology, 28*(3) 437-454. doi:10.1007/s10940-011-9147-5.

Carlson, C.A. (2008). *Distinctiveness in an eyewitness identification paradigm: Comparing simultaneous and sequential lineups.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. Retrieved from http://gradworks.umi.com/33/15/3315570.html

Chae, Y. (2010). Application of laboratory research on eyewitness testimony. *Journal of Forensic Psychology Practice, 10*(3)*,* 252-261. doi:10.1080/15228930903550608

Chapdelaine, A., & Griffin, S.F. (1997). Beliefs of guilt and recommended sentence as a function of juror bias in the O. J. Simpson trial. *Journal of Social Issues, 53*(3)*,* 477-485. doi:10.1111/j.1540-4560.1997.tb02123.x

Clark, S.E. (2012). Costs and benefits of eyewitness identification reform psychological science and public policy. *Perspectives on Psychological Science, 7*(3), 238-259. doi:10.1177/1745691612439584

Clark, S.E., & Davey, S.L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior, 29*(2), 151-172. doi:10.1007/s10979-005-2418-7

Clark, S.E., & Tunnicliff, J.L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior, 25*(3), 199-216. doi: 10.1023/A:1010753809988

Clifford, B.R., & Scott, J. (1978). Individual and situational factors in eyewitness testimony. *Journal of Applied Psychology, 63*(3), 352-359. doi:10.1037/0021-9010.63.3.352

Committee on Identifying the Needs of the Forensic Science Community, National Research Council. (2009). *Strengthening forensic science in the United States: A path forward.* (Report No. NCJ 228091). Retrieved from: https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

Connors, E., Lundregan, T., Miller, N., & McEwen, T. (1996). *Convicted by juries, exonerated by science: Case studies in the use of DNA evidence to establish innocence after trial.* (Report No. NCJ 161258). Washington DC: National Institute of Justice. Retrieved from: https://www.ncjrs.gov/pdffiles/dnaevid.pdf

Cowdery, N. (2005). Wrongful conviction and double jeopardy. *Judicial Officers' Bulletin, 17*(4), 27-30.

Cutler, B.L, Penrod, S.D., & Martens, T.K. (1987). The reliability of eyewitness identification: The role of system and estimator variables. *Law and Human Behavior, 11*(3), 233-258. doi:10.1007/BF01044644

Cutler, B.L., & Bull Kovera, M. (2008). Introduction to commentaries on the Illinois pilot Study of lineup reforms. *Law and Human Behavior*, *32*(1), 1-2. doi:10.1007/s10979-007-9120-x

Deffenbacher, K.A., Bornstein, B.H., Penrod, S.D., & McGorty, E.K., (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*(6), 687-706. doi:10.1007/s10979-004-0565-x

Devenport, J.L., Penrod, S.D., & Cutler, B.L. (1997). Eyewitness identification evidence: Evaluation commonsense evaluations. *Psychology, Public Policy, and Law, 3*(2-3), 338-361. doi:10.1037/1076-8971.3.2-3.338

Doyle, J.M. (2005). True witness: Cops, courts, science, and the battle against misidentification. New York, New York: Palgrave MacMillan.

Durose, M.R., & Langan, P.A. (2003). *Felony Sentences in state courts, 2000*. (Report No. NCJ 198821). Retrieved from Bureau of Justice Statistics website: http://bjs.gov/content/pub/pdf/fssc00.pdf

Espinoza, R.K.E., & Willis-Esqueda, C. (2008). Defendant and defense attorney characteristics and their effects on juror decision-making and prejudice against Mexican Americans. *Cultural Diversity and Ethnic Minority Psychology, 14*(4), 364-371. doi:10.1037/a0012767

Eyewitness misidentification. (n.d.). In The Innocence Project's understanding the causes (2014). Retrieved from http://www.innocenceproject.org/understand/Eyewitness-Misidentification.php

Finklea, K.M., & Ebbesen, E.B., (2007). *Eyewitness accuracy in the real world: DNA evidence as "ground truth" for eyewitness accuracy rates.* Unpublished manuscript.

Frederick, B., & Stemen, D. (2012). *The anatomy of discretion: An analysis of decision making - technical report* (Report for Award No. 2009-IJ-CX-0040). Washington DC: Vera Institute of Justice. Retrieved from: https://www.ncjrs.gov/pdffiles1/nij/grants/240334.pdf

Freedman, M.H. (1966). Professional responsibility of the criminal defense lawyer: The three hardest questions. *The Michigan Law Review Association, 64*(8), 1469-1484.

Gardner, T.J. & Anderson, T.M. (7th Ed.). (2012). *Criminal evidence: Principles and cases.* Belmont, California: Wadsworth, Cengage Learning.

Garner, J. H. & Maxwell, C.D. (2009). Prosecution and conviction rates for intimate partner violence. *Criminal Justice Review* 34(1), 44-79. doi:10.1177/0734016808324231

Garrett, B. (2008). Judging innocence. *Columbia Law Review, 108*(1), 55-142.

Glater, J.D., (2008, August 8). Study finds settling is better than going to trial. *The New York Times*. Retrieved from http://www.nytimes.com/2008/08/08/business/08law.html

Gould, J.B., Carrano, J., Leo, R., & Young, J. (2012). Predicting erroneous convictions: A social science approach to miscarriages of justice. (Report for Award No. 2009-IJ-CX-4110). Washington, DC: National Institute of Justice. Retrieved from: http://observer.american.edu/spa/djls/prevent/upload/Predicting-Erroneous-Convictions.pdf

Greathouse, S.M., & Bull Kovera, M. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior*, *33*(1), 70-82. doi:10.1007/s10979-008-9136-x

Gross, S.R., & Shaffer, M. (2012). *Exonerations in the United States, 1989-2012.* (Research Report). Retrieved from the National Registry of Exonerations website: https://www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf

Hedding, N. (2002). The fine line between strategic miscalculation and harmful error: Consequences and repercussions of legal malpractice to the criminal defense attorney. *Journal of Legal Advocacy and Practice, 4,* 222-234.

Jacoby, J.E., Mellon, L.R., Ratledge, E.C., and Turner, S.T. (1982). *Prosecutorial decision-making: A national study.* (Report for Award No. 79-NI-AX-0006 and 79-NI-AX-0034). Washington DC: National Institute of Justice. Retrieved from http://www.jijs.org/publications/prospubs/Decisionmaking.pdf

Kassin, S.M., Dror, I.E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2*(1)*,* 42-52. doi:10.1016/j.jarmac.2013.01.001

Kassin, S.M., Reddy, M.E., & Tulloch, W.F. (1990). Juror interpretations of ambiguous evidence: The need for cognition, presentation order, and persuasion. *Law and Human Behavior, 14*(1), 43-55. doi:10.1007/BF01055788

Kerr, N.L., Boster, F., Callen, C.R., Braz, M.E., O'Brien, B., & Horowitz, I. (2008). Jury nullification instructions as amplifiers of bias. *International Commentary on Evidence, 6*(1), 1554-4567. doi: 10.2202/1554-4567.1068

Kilpatrick, D.G., Resnick, H.S., Ruggiero, K.J., Conoscenti, L.M., & McCauley, J. *Drug-facilitated, incapacitated, and forcible rape: A national study.* (Research Report for Award No. 2005-WG-BX-0006). Charleston, South Carolina: National Crime Victims Research & Treatment Center. Retrieved from http://www.niccsa.org/downloads/elders/DRUGFACILITATEDINCAPACITATEDANDFORCIBLERAPE.pdf

Konecni, V.J., & Ebbesen, E.B. (1986). Courtroom testimony by psychologists on eyewitness identification issues: Critical notes and reflections. *Law and Human Behavior, 10*(1-2), 117-126. doi:10.1007/BF01044563

LaFree, G. (1989). Rape and criminal justice: The social construction of sexual assault, Belmont, California: Wadsworth.

Lindholm, T. (2008). Who can judge the accuracy of eyewitness statements? A comparison of professionals and lay-persons. *Applied Cognitive Psychology, 22*(9)*,* 1301-1314. doi:10.1002/acp.1439

Lindsay, R.C., & Wells, G.L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, *70*(3), 556-564. doi: 10.1037/0021-9010.70.3.556

Loftus, E.F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7*(4)*,* 560-72. doi: 10.1016/0010-0285(75)90023-7

Loftus, E.F., Miller, D.G., & Burns, H.J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*(1)*,* 19-31. doi: 10.1037/0278-7393.4.1.19

Loftus, E.F., & Palmer, J.C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585-589. doi:10.1016/S0022-5371(74)80011-3

Magnussen, S., Melinder, A., Stridbeck, U., & Raja, A.Q. (2010). Beliefs about factors affecting the reliability of eyewitness testimony: A comparison of judges, jurors and the general public. *Applied Cognitive Psychology, 24*(1)*,* 122-133. doi: 10.1002/acp.1550

Maguire, K., & Pastore, A.L. (2003). *Sourcebook of Criminal Justice Statistics: 2002.* (Report No. NCJ 203301). Washington DC: Bureau of Justice Assistance. Retrieved from: https://www.hsdl.org/?view&did=711164

Malpass, R.S. (2006). A policy evaluation of simultaneous and sequential lineups. *Psychology, Public Policy, and Law, 12*(4), 394-418. doi: 10.1037/1076-8971.12.4.394

Malpass, R.S., & Devine, P.G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66*(4), 482-489. doi: 10.1037/0021-9010.66.4.482

McCloskey, M., & Egeth, H.E. (1983). Eyewitness identification: What can a psychologist tell a jury? *American Psychologist, 38*(5), 550-563. doi: 10.1037/0003-066X.38.5.550

McKenna, J.D., Treadway, M., & McCloskey, M.E. (1992). Expert psychological testimony on eyewitness reliability: Selling psychology before its time. In P. Suedfeld & P.E. Tetlock (Eds.), *Psychology and social policy* (pp. 283-293). New York, New York: Hemisphere Publishing Cooperation.

Mecklenberg, S.H. (2006). *The Illinois pilot program on sequential double-blind identification procedures*. Report submitted to the state legislature of the State of Illinois on behalf of the Illinois State Police, Springfield, Illinois.

Mecklenburg, S.H., Bailey, P.J., & Larson, M.R. (2008). The Illinois field study: A significant contribution to understanding real world eyewitness identification issues. *Law and Human Behavior, 32*(1)*,* 22-27. doi: 10.1007/s10979-007-9108-6

Memon, A., & Gabbert, F. (2003). Unravelling the effects of sequential presentation in culprit-present lineups. *Applied Cognitive Psychology, 17*(6), 703-714. doi: 10.1002/acp.909

Merola, M. (1982). Federal, state and local governments: Partners in the fight against violent crime. *Journal of Criminal Law and Criminology, 73*(3), 965-984.

Mickes, L., Flowe, H.D., & Wixted, J.T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*(4), 361-376. doi:10.1037/a0030609

Münsterberg, H. (1908). *On the witness stand.* Garden City, New York: Doubleday, Page & Company.

National Center for Prosecution Management (1974). *Report to the Bronx County District Attorney on the case evaluation system.* (Report No. NCJ 029088). Retrieved from: https://www.ncjrs.gov/pdffiles1/Digitization/29088NCJRS.pdf

National Institute of Justice (1999). *Eyewitness evidence: A guide for law enforcement.* (Report No. NCJ 178240). Retrieved from: https://www.ncjrs.gov/pdffiles1/nij/178240.pdf

Oppel Jr., R. (2011, September 25). Sentencing shift gives new leverage to prosecutors. *The New York Times*. Retrieved from http://www.nytimes.com/2011/09/26/us/tough-sentences-help-prosecutors-push-for-plea-bargains.html?pagewanted=all&_r=1&.

Peterson, J.L, Hickman, M.J., Strom, K.J., & Johnson, D.J. (2013). Effect of forensic evidence on criminal justice case processing. J*ournal of Forensic Sciences, 58*(S1), S78-S90. doi: 10.1111/1556-4029.12020

Peterson, J.L., Ryan, J.P., Houlden, P.J., & Mihajlovic, S. (1987). The uses and effects of forensic science in the adjudication of felony cases. *Journal of Forensic Sciences, 32*(6)*,* 1730-1753.

Pozzulo, J.D., Lemieux, J.M.T., Wilson, A., Crescini, C., & Girardi, A. (2009). The influence of identification decision and DNA evidence on juror decision-making. *Journal of Applied Social Psychology, 39*(9), 2069-2088. doi:10.1111/j.1559-1816.2009.00516.x

Pratt, T.C., Gaffney, M.J., Lovrich, N.P., & Johnson, C.L. (2006). This isn't CSI: Estimating the National backlog of forensic DNA cases and the barriers associated with case processing. *Criminal Justice Policy Review*, *17*(1), 32-47. doi:10.1177/0887403405278815

Raeder, Myrna (2012). Overturning wrongful convictions and compensating exonerees. In Springer, *Encyclopedia of Criminology and Criminal Justice,* (Bruinsma & Weisburd, eds.).

Samuels, J.E., Davies, E.H., & Pope, D.B. (2013). *Collecting DNA at arrest: Policies, practices, and implications.* (Report for Award No. 2009-DN-BX-0004). Washington, DC: The Urban Institute Justice Policy Center. Retrieved at: https://www.ncjrs.gov/pdffiles1/nij/grants/242813.pdf

Schacter, D.L., Dawes, R., Jacoby, L.L., Kahneman, D., Lempert, R., Roediger, H.L., & Resenthal, R. (2008). Police fourm: Studying eyewitness investigations in the field. *Law and Human Behavior, 32*(1)*,* 3-5. doi:10.1007/s10979-007-9093-9

Smith, L.L., & Bull, R. (2012). Identifying and measuring juror pre-trial bias for forensic evidence: Development and validation of the Forensic Evidence Evaluation Bias Scale. *Psychology, Crime & Law, 18*(9), 797-815. doi: 10.1080/1068316X.2011.561800

Smith, L.L., & Bull, R. (2013). Exploring the disclosure of forensic evidence in police interviews with suspects. *Journal of Police and Criminal Psychology, July,* 1-6. doi: 10.1007/s11896-013-9131-0

Smith, L.L., Bull, R., & Holliday, R. (2011). Understanding juror perceptions of forensic evidence: Investigating the impact of case context on perceptions of forensic evidence strength. *Journal of Forensic Sciences*, *56*(2), 409-414. doi: 10.1111/j.1556-4029.2010.01671.x

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*(3), 315-327. doi: 10.1037/0033-2909.118.3.315

Steblay, N.K. (1997). Social Influence in eyewitness recall: A meta-analytic review of lineup instruction efforts. *Law and Human Behavior, 21*(3)*,* 283-298. doi: 10.1023/A:1024890732059

Steblay, N.K. (2011). What we know now: The Evanston Illinois field lineups. *Law and Human Behavior, 35*(1), 1-12. doi: 10.1007/s10979-009-9207-7

Steblay, N.K., Dysart, J.E., Fulero, S., & Lindsay, R.C.L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*(5), 459-473. doi: 10.1023/A:1012888715007

Steblay, N.K., Dysart, J.E., & Wells, G.L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*(1), 99-139. doi: 10.1037/a0021650

Tollestrup, P.A., Turtle, J. W., & Yuille, J. C. (1994). Actual victims and witnesses to robbery and fraud: An archival analysis. In D. F. Ross, J. D. Read, & M. P. Toglia, (Eds.), Adult eyewitness testimony: Current trends and developments (pp. 144-160). New York City, New York: Cambridge University Press.

Wagstaff, G.F., MacVeigh, J., Boston, R., Scott, L., Brunas-Wagstaff, J., & Cole, J. (2003). Can laboratory findings on eyewitness testimony be generalized to the real world? An archival analysis of the influence of violence, weapon presence, and age on eyewitness accuracy. *Journal of Psychology, 137*(1), 17-28.

Wells, G.L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*(12)*,* 1546-1557. doi: 10.1037/0022-3514.36.12.1546

Wells, G.L., Malpass, R., Lindsay, R.C.L., Fisher, R.P., Turtle, J.W., & Fulero, S.M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*(6), 581-598. doi: 10.1037/0003-066X.55.6.581

Wells, G.L., Memon, A., & Penrod, S.D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45-75. doi: 10.1111/j.1529-1006.2006.00027.x

Wells, G.L, Small, M., Penrod, S.D., Malpass, R.S., Fulero, S.M., & Brimacombe, C.A.E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*(6), 603-607. doi: 10.1023/A:1025750605807

Wells, G.L., Steblay, N.K., & Dysart, J.E. (2011). A test of the simultaneous vs. sequential lineup methods: An initial report of the AJS national eyewitness identification field studies. Des Moines, Iowa: American Judicature Society. Retrieved from: http://www.popcenter.org/library/reading/PDFs/lineupmethods.pdf

Wise, R.A., & Safer, M.A. (2010). A comparison of what U.S. judges and students know and believe about eyewitness testimony. *Journal of Applied Social Psychology, 40*(6), 1400-1422. doi:10.1111/j.1559-1816.2010.00623.x

Wixted, J.T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*(3), 275-278. doi: 10.1177/1745691612442906

Wright, D.B., & McDaid, A.T. (1996). Comparing system and estimator variables using data from real lineups. *Applied Cognitive Psychology, 10*(1), 75-84. doi: 10.1002/(SICI)1099-0720(199602)10:1<75::AID-ACP364>3.0.CO;2-E

Yuille, J.C., & Wells, G.L. (1991). Concerns about the application of research findings: The issue of ecological validity. In J. Doris (Ed.). The suggestibility of children's recollections: Implications for eyewitness testimony (pp. 118-128). Washington, DC: American Psychological Association.

Appendix A

Police Foundation Strength of Evidence Scale

Rater ID: ________

Case #: A ________ - ____

I. PHYSICAL EVIDENCE: Tangible items that directly link the suspect to the crime.

1 Evidence and/or information that is weak. 3 Evidence and/or information that is strong. 5

FACTOR

FACTOR RATING

1. DNA - biological material recovered from crime scene

DNA sample is contaminated (1.40)

Fluids from the crime scene prove joint presence of DNA for both victim AND suspect (4.46)

1________2________3________4________5 □

Fluid sample lacks sufficient quantity of DNA for testing (1.55)

Multiple DNA tests with consistent findings (4.69)

2. Surveillance Tapes/ Photos from Crime Scene - images or photos that capture the crime

Parts of surveillance tape are clear and parts are not (2.81)

Tape provides a profile view of the suspect (3.64)

Video captures the entire crime in real time (4.83)

1________2________3________4________5 □

Videotape footage is within close proximity of the suspect committing the crime (3.69)

3. Fingerprints - fingerprint evidence recovered from crime scene

Full fingerprint is smudged (1.70)

Suspect's fingerprint found at the crime scene (4.17)

1________2________3________4________5 □

Portion of fingerprint is smudged (2.51)

Suspect print found on weapon used in commission of crime (4.39)

4. Wire Taps/ Audio Tapes - audio material that gives information about the crime

No identifying information is provided on the tape (1.47)

Volume of audio tape is loud (3.42)

1________2________3________4________5 □

Audio tape has intermittent background noise (3.04)

Name of suspect is identified in audio taped conversation (3.59)

5. Trace Evidence - including bite marks, tire marks, tools, etc

Indistinguishable bite marks found on victim (1.73)

Tool owned by suspect matches marks on victim's door (3.43)

1________2________3________4________5 □

Tire marks at crime scene are consistent with suspect's car (3.24)

Casing from crime scene matches the casings from suspect's gun (4.24)

6. Recovery of Items - such as drugs, guns, knives, and stolen items

Perp's high school class ring is recovered at crime scene and is from suspect's class, school, & year (3.48)

Recovered bullet matches suspect's gun (4.40)

1________2________3________4________5 □

No shell casings found (1.90)

Stolen item found in suspect's girlfriend's car (3.60)

Victim's driver's license is on the suspect (4.35)

7. Other Miscellaneous Evidence

Suspect's home and crime scene fall in the same cell tower radius. (1.95)

Suspect writing sample matches writing sample on a robbery note (3.77)

1________2________3________4________5 □

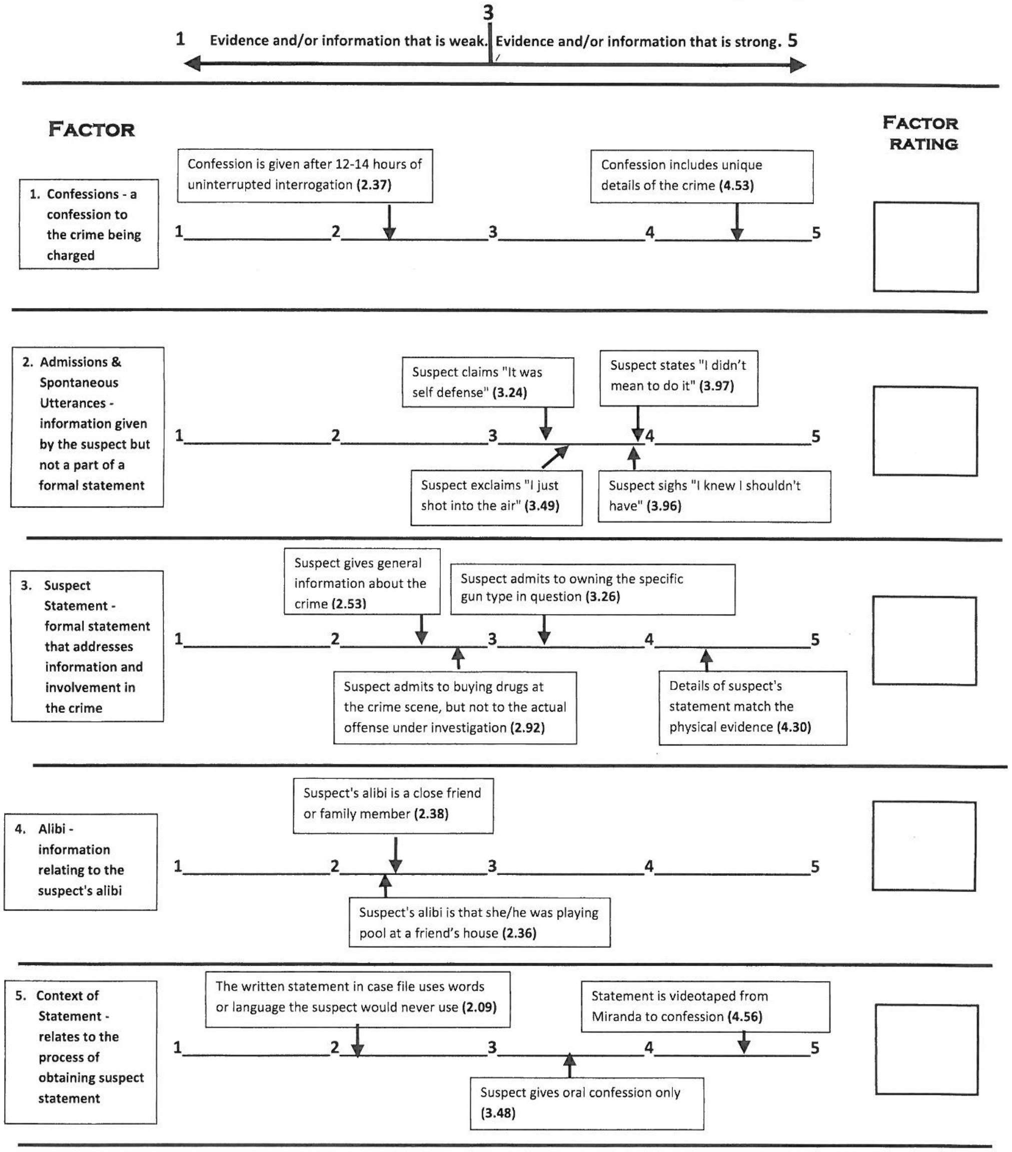
Charges made on stolen card are made near the crime scene (3.55)

Items are purchased online using the victim's card and are shipped to suspect's home (4.40)

RATER #____________________

PHYSICAL EVIDENCE OVERALL RATING □

II. SUSPECT STATEMENT INFORMATION: Details provided by the suspect to the police that link/fail to link the suspect to the particular crime being charged.



1 Evidence and/or information that is weak. 3 Evidence and/or information that is strong. 5

FACTOR

FACTOR RATING

1. Confessions - a confession to the crime being charged

1____2____3____4____5 □

Confession is given after 12-14 hours of uninterrupted interrogation **(2.37)**

Confession includes unique details of the crime **(4.53)**

2. Admissions & Spontaneous Utterances - information given by the suspect but not a part of a formal statement

1____2____3____4____5 □

Suspect claims "It was self defense" **(3.24)**

Suspect states "I didn't mean to do it" **(3.97)**

Suspect exclaims "I just shot into the air" **(3.49)**

Suspect sighs "I knew I shouldn't have" **(3.96)**

3. Suspect Statement - formal statement that addresses information and involvement in the crime

1____2____3____4____5 □

Suspect gives general information about the crime **(2.53)**

Suspect admits to owning the specific gun type in question **(3.26)**

Suspect admits to buying drugs at the crime scene, but not to the actual offense under investigation **(2.92)**

Details of suspect's statement match the physical evidence **(4.30)**

4. Alibi - information relating to the suspect's alibi

1____2____3____4____5 □

Suspect's alibi is a close friend or family member **(2.38)**

Suspect's alibi is that she/he was playing pool at a friend's house **(2.36)**

5. Context of Statement - relates to the process of obtaining suspect statement

1____2____3____4____5 □

The written statement in case file uses words or language the suspect would never use **(2.09)**

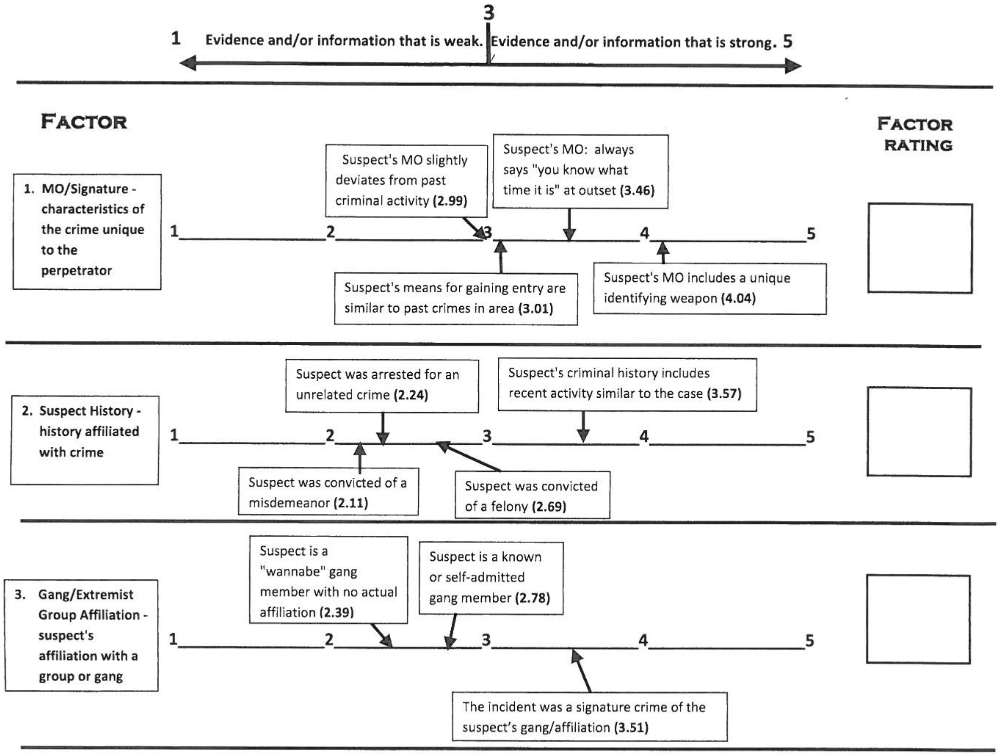
Statement is videotaped from Miranda to confession **(4.56)**

Suspect gives oral confession only **(3.48)**

RATER #________________

SUSPECT STATEMENT OVERALL RATING □

III. SUSPECT HISTORY: Suspect's law enforcement history and/or group/gang affiliation that speaks to the likelihood that the suspect committed the crime.



3

1 Evidence and/or information that is weak. Evidence and/or information that is strong. 5

FACTOR

FACTOR RATING

1. MO/Signature - characteristics of the crime unique to the perpetrator

1_____2_____3_____4_____5

Suspect's MO slightly deviates from past criminal activity **(2.99)**

Suspect's MO: always says "you know what time it is" at outset **(3.46)**

Suspect's means for gaining entry are similar to past crimes in area **(3.01)**

Suspect's MO includes a unique identifying weapon **(4.04)**

□

2. Suspect History - history affiliated with crime

1_____2_____3_____4_____5

Suspect was arrested for an unrelated crime **(2.24)**

Suspect's criminal history includes recent activity similar to the case **(3.57)**

Suspect was convicted of a misdemeanor **(2.11)**

Suspect was convicted of a felony **(2.69)**

□

3. Gang/Extremist Group Affiliation - suspect's affiliation with a group or gang

1_____2_____3_____4_____5

Suspect is a "wannabe" gang member with no actual affiliation **(2.39)**

Suspect is a known or self-admitted gang member **(2.78)**

The incident was a signature crime of the suspect's gang/affiliation **(3.51)**

□

RATER #____________________

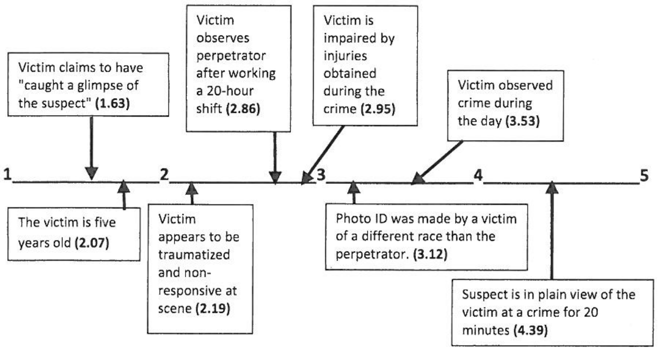
SUSPECT CHARACTERISTICS OVERALL RATING □

IV. VICTIM CHARACTERISTICS: Information that speaks to the veracity of the victim(s) involved in the crime.

1 Evidence and/or information that is weak. 3 Evidence and/or information that is strong. 5

FACTOR | **FACTOR RATING**

1. Reliability - factors that influence the accuracy of information provided by victim



Victim claims to have "caught a glimpse of the suspect" **(1.63)**

The victim is five years old **(2.07)**

Victim appears to be traumatized and non-responsive at scene **(2.19)**

Victim observes perpetrator after working a 20-hour shift **(2.86)**

Victim is impaired by injuries obtained during the crime **(2.95)**

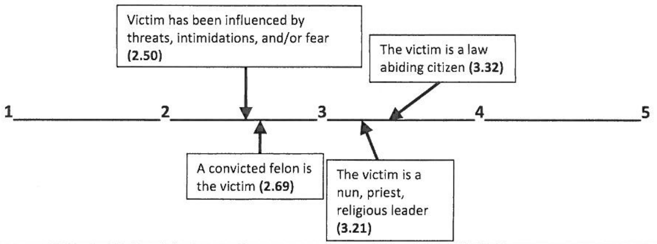
Photo ID was made by a victim of a different race than the perpetrator. **(3.12)**

Victim observed crime during the day **(3.53)**

Suspect is in plain view of the victim at a crime for 20 minutes **(4.39)**

□

2. Credibility - factors that influence the likelihood that the victim is giving truthful information



Victim has been influenced by threats, intimidations, and/or fear **(2.50)**

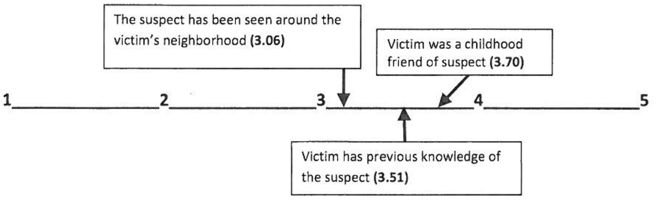
A convicted felon is the victim **(2.69)**

The victim is a nun, priest, religious leader **(3.21)**

The victim is a law abiding citizen **(3.32)**

□

3. Knowledge of/or Familiarity with Suspect



The suspect has been seen around the victim's neighborhood **(3.06)**

Victim has previous knowledge of the suspect **(3.51)**

Victim was a childhood friend of suspect **(3.70)**

□

RATER #____________________

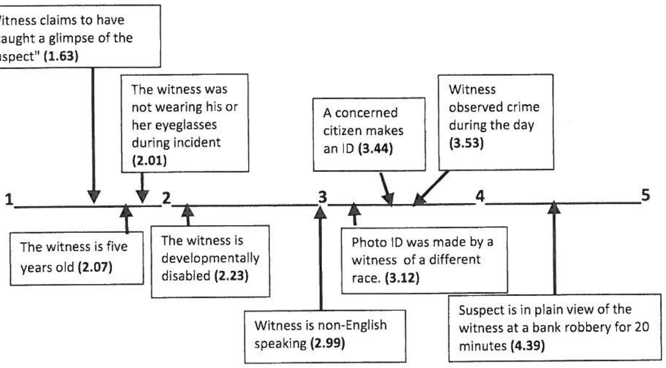
VICTIM CHARACTERISTICS OVERALL RATING □

V. WITNESS CHARACTERISTICS: Information that speaks to the veracity of any witness whether or not they observed the incident.

1 Evidence and/or information that is weak. 3 Evidence and/or information that is strong. 5

FACTOR | **FACTOR RATING**

1. Reliability - factors that influence the accuracy of information provided by the witness



Witness claims to have "caught a glimpse of the suspect" **(1.63)**

The witness was not wearing his or her eyeglasses during incident **(2.01)**

The witness is five years old **(2.07)**

The witness is developmentally disabled **(2.23)**

Witness is non-English speaking **(2.99)**

Photo ID was made by a witness of a different race. **(3.12)**

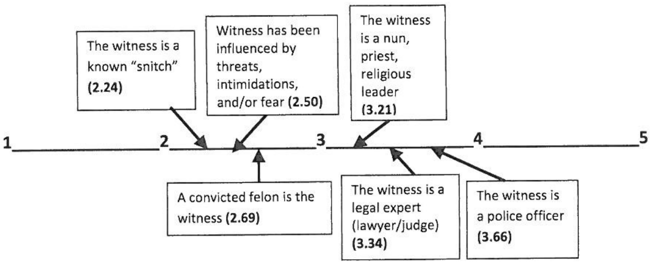
A concerned citizen makes an ID **(3.44)**

Witness observed crime during the day **(3.53)**

Suspect is in plain view of the witness at a bank robbery for 20 minutes **(4.39)**

□

2. Credibility - factors that influence the likelihood that the witness is giving truthful information



The witness is a known "snitch" **(2.24)**

Witness has been influenced by threats, intimidations, and/or fear **(2.50)**

A convicted felon is the witness **(2.69)**

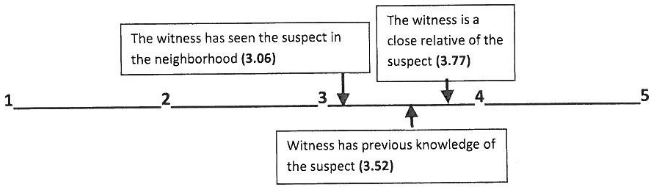
The witness is a nun, priest, religious leader **(3.21)**

The witness is a legal expert (lawyer/judge) **(3.34)**

The witness is a police officer **(3.66)**

□

3. Knowledge of/or Familiarity with Suspect



The witness has seen the suspect in the neighborhood **(3.06)**

Witness has previous knowledge of the suspect **(3.52)**

The witness is a close relative of the suspect **(3.77)**

□

RATER #______________________

WITNESS CHARACTERISTICS OVERALL RATING □

VI. IDENTIFICATION INFORMATION: Independent corroboration of information linking the suspect to the particular incident, regardless of source.

1 Evidence and/or information that is weak. 3 Evidence and/or information that is strong. 5

FACTOR	1 — 2 — 3 — 4 — 5	FACTOR RATING
1. Unique ID Information - clothing, tattoo, hair, line-ups, etc. identifying the suspect	Perpetrator was wearing a non-descript T-shirt and jeans **(1.56)** Perpetrator is described as having brown hair and brown eyes **(1.92)** Witness fails to describe unique characteristics of suspect **(1.92)** Perpetrator wore thick glasses **(2.91)** Perpetrator is described as a white male, 6'0 tall, 200 lbs. with mustache/ braids **(3.26)** Tattoos/scars described by the victim/witness match the suspect **(4.23)** Victim names suspect by name **(4.27)**	
2. Details of Crime - details obtained through the investigation	Vehicle was stolen and suspect was observed in the vehicle close to the crime scene **(3.87)** Suspect leaves personal property at crime scene **(4.28)**	
3. Witness ID Information - account of incident is given	Witness describes the weapon as a black gun **(1.85)** Variations and/or inconsistencies in witness account **(2.04)** Witness statement is consistent with details of the crime **(4.29)** Multiple corroborating witness statements **(4.41)**	
4. 3rd Party/ Complainant Information	Co-conspirator implicates a person unrelated to the crime **(2.20)** Third party verbal statement implicates the suspect **(2.86)** Pawn shop owner states suspect came in wanting to sell jewelry **(3.62)**	
5. Circumstances Surrounding Arrest - factors regarding the arrest	Multiple parties near stolen object or weapon **(2.11)** Suspect is cool, collected & cooperative during arrest **(2.22)** Suspect is arrested 2 or 3 blocks from crime scene and within minutes of crime **(3.72)** Suspect is hiding near crime scene **(3.89)**	
6. Flipping the Suspect - information from a co-conspirator	Co-conspirator reveals information about crime involving other person/suspect **(3.08)**	
7. Anonymously Provided Information	Receive vague tip from an anonymous person **(1.44)** Receive detailed tip from an anonymous person **(2.81)** Anonymous tip corroborated by other evidence **(3.40)**	

RATER #________________

IDENTIFICATION INFORMATION OVERALL RATING

Appendix B

Individual Final Rating Form

Final Rater Evaluation Form

Rater # : _________

Case #: A_______- ___

Instructions:

(1) Transfer your ratings from the instrument to the "initial rating" column of this form for all six categories.

(2) Determine your overall rating of the evidence against the specific suspect (do not add or average your category ratings).

(3) Indicate your post-consensus rating in the "final rating" column (even if unchanged). If you changed the score, note the reason for the change by checking the appropriate box(es) BELOW the Category. IF the reason is not listed, check other and DESCRIBE the reason. '

(4) If you changed your final OVERALL RATING, PLEASE DESCRIBE YOUR REASONS FOR CHANGE IN YOUR OWN WORDS (E.G. Missed physical evidence caused me to give a higher overall evidentiary strength rating, lowered scores in two categories which led me to believe the overall evidence was not as strong, etc.)

CATEGORY	INITIAL RATING	FINAL RATING
Physical Evidence		
Please indicate reason for change. Check (√) all that apply. € Missed Evidence € Evidence Reconsidered € Misinterpreted Evidence € Evidence Weighed in Wrong Category € Other: ____________________		
Suspect Statement Information		
Please indicate reason for change. Check (√) all that apply. € Missed Evidence € Evidence Reconsidered € Misinterpreted Evidence € Evidence Weighed in Wrong Category € Other: ____________________		
Suspect History		
Please indicate reason for change. Check (√) all that apply. € Missed Evidence € Evidence Reconsidered € Misinterpreted Evidence € Evidence Weighed in Wrong Category € Other: ____________________		

Please complete SIDE 2 ⇨

Victim Characteristics		

Please indicate reason for change. Check (√) all that apply.

€ Missed Evidence € Evidence Reconsidered € Misinterpreted Evidence

€ Evidence Weighed in Wrong Category € Other: ______________________________

Witness Characteristics		

Please indicate reason for change. Check (√) all that apply.

€ Missed Evidence € Evidence Reconsidered € Misinterpreted Evidence

€ Evidence Weighed in Wrong Category € Other: ______________________________

Identification Information		

Please indicate reason for change. Check (√) all that apply.

€ Missed Evidence € Evidence Reconsidered € Misinterpreted Evidence

€ Evidence Weighed in Wrong Category € Other: ______________________________

OVERALL CASE RATING		***If changed, please answer question below***

Please indicate reason for change. **In your own words, please describe the reason you changed your overall case rating. It could consist of more than one reason.**

PLEASE CONTINUE TO NEXT PAGE!! ⇨