UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IMPROVING JUROR SENSITIVITY TO EYEWITNESS CONFIDENCE INFLATION USING A VISUAL AID

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

KYLIE N. KEY

Norman, Oklahoma

2019

IMPROVING JUROR SENSITIVITY TO EYEWITNESS CONFIDENCE INFLATION USING A VISUAL AID

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Scott Gronlund

Dr. Robert Hamm

Dr. Robert Terry

Dr. Hairong Song

Dr. Edward Cokely

Table of Contents

Abstract

Contrary to the widespread belief that eyewitnesses are always fallible and that an eyewitness' confidence is not indicative of identification accuracy, a new body of literature tells a different story: A highly confident eyewitness, measured properly (first fair test of memory, immediate confidence recorded), is likely to be correct; and conversely, an eyewitness that reports low confidence in the same situation is likely making an error. Although there is evidence that jurors intuitively understand this strong confidence-accuracy relationship, prior research shows that they do not understand the measurement nuances, and that interventions are needed. This dissertation reviews some traditional approaches to improving how jurors think about eyewitness evidence, like pattern jury instructions and expert testimony, concluding that these approaches are too complex and, if anything, simply make jurors skeptical. The present studies designed and tested a novel intervention, presenting a simplified message focused on initial eyewitness confidence from fair memory test circumstances, using a visual aid with supporting text instruction. Overall, the results showed some support for this intervention as an approach to sensitizing jurors to eyewitness evidence. Specifically, the intervention sensitized jurors to eyewitness confidence inflation, compared to modified Henderson instructions and a control condition with no instruction. These results suggest that the visual aid intervention could be a viable alternative to traditional approaches, pending further research to validate this novel aid.

Keywords: jury decision-making; skepticism; sensitivity; decision aid; comprehension; eyewitness confidence

Chapter 1: Introduction

Both laboratory research and real criminal cases document that eyewitness memory is fallible; eyewitnesses sometimes misidentify the perpetrator of the crime they witnessed. This has been a contributing factor in at least 34% of the more than 1700 wrongful convictions that have been overturned by DNA or other evidence (National Registry of Exonerations, 2018), and researchers estimate that this may just be the tip of the iceberg. Laboratory research shows that memory can be impacted by a variety of system variables, factors under the control of the legal system (e.g. identification procedure, witness instructions, double blind administration), and estimator variables, the impact of which can only be estimated (e.g., weapon focus, cross-race, exposure duration, lighting) (for a review, see Wells, 1993). This is problematic in and of itself, but is magnified when considering the role of eyewitness confidence.

Confidence can be artificially inflated by factors like post-identification feedback that confirm an eyewitness' choice or by an unfair identification procedure (whereby a suspect stands out unduly) (see Key et al., 2017; Neuschatz et al., 2016; Steblay et al., 2014; Wells & Bradfield, 1998). Moreover, the confidence-accuracy correlation is around $r = .4$, which has been interpreted as a relatively weak relationship (Sporer et al., 1995). All these factors led researchers to conclude that eyewitnesses are untrustworthy, even when highly confident (Deffenbacher, 1980; Sporer et al., 1995). As I will discuss below, this interpretation influenced legal cases by way of pattern jury instructions and expert testimony. However, Juslin, Olsson, and Winman (1996) showed that the correlation coefficient does not provide the best depiction of the confidence-accuracy relationship because it aggregates across all levels of confidence, providing a single index of the relationship between confidence and accuracy. In fact, the correlation coefficient can vary from 0 to 1 despite the underlying data exhibiting a very strong

relationship between confidence and accuracy. Juslin et al. argued that the confidence-accuracy relationship is better revealed by confidence calibration curves, which depict the proportion of accurate judgments made at each level of confidence. This information also answers the question of interest to jurors: what is the likelihood that the witness in this case is accurate, given his or her level of confidence?

A recent body of literature, relying on calibration curves, concludes that eyewitness memory can be highly reliable when measured properly, in laboratory research, in field studies, and in real criminal cases (Behrman & Davey, 2001; Behrman & Richards, 2005; Brewer & Wells, 2006; Garrett, 2011; Gronlund & Benjamin, 2018; Sauerland & Sporer, 2007; Sauerland & Sporer, 2009; Wixted, Mickes, Clark, Gronlund, & Roediger, 2015a; Wixted, Mickes, Dunn, Clark, & W. Wells, 2015b; Wixted, Mickes, & Fisher, 2018; Wixted & Wells, 2017). For example, Wixted et al.'s (2015b) field study revealed that about 95% of eyewitnesses reporting the highest level of confidence identified the police suspect, compared with only about 20% of low confidence witnesses. These high confidence identifications also were more likely to be accompanied by corroborating evidence of guilt. A meta-analysis conducted by Wixted and Wells (2017) collapsed across data from 15 laboratory studies varying system and estimator variables and found a strong confidence-accuracy relationship. Importantly, this strong relationship holds even in the presence of conditions that impair memory accuracy (Semmler Dunn, Mickes, & Wixted, 2018). For example, even if there is a weapon present or a cross-race identification, both of which impair accuracy, the confidence-accuracy relationship remains strong (although there are fewer eyewitnesses who are able to express high confidence under the adverse weapon-present or cross-race conditions).

However, two measurement nuances are necessary to expound here: (1) the eyewitness must choose someone from the identification procedure, rather than reject the lineup, and (2) the confidence judgment must be made immediately after the first, fair test of memory (Wixted et al., 2015a). These two nuances are critical for ensuring the reliability of confidence judgments. Confidence judgments that are delayed (e.g., those reported in the courtroom, hereafter referred to as courtroom confidence), or from an unfair test (e.g., a biased lineup, a single-person showup, more than one exposure to the suspect), are less meaningfully related to accuracy due to factors like confidence inflation. Confidence inflation typically is defined as an initial expression of low confidence but a subsequent expression of higher confidence. Garrett's (2011) analysis of the first 161 DNA exoneration cases revealed that confidence inflation occurred in at least 57% of the cases (with missing data for the remainder of cases). In other words, the eyewitnesses making initial misidentifications in these cases actually signaled that there was a good chance they were making an incorrect identification by reporting low confidence (or choosing a filler or rejecting the lineup). Given the high frequency with which confidence inflation occurs, paired with the severe consequences that it can have for the justice system, I am interested in two questions:

1. Do jurors understand that a strong confidence-accuracy relationship exists in eyewitness memory?
2. Do jurors understand its measurement nuances? If not, what interventions can improve this understanding?

To explore these questions, I reviewed the literature on jurors' behavior, which indicates that jurors do understand the strong confidence-accuracy relationship but ignore the measurement nuances. I also reviewed the literature on traditional approaches to juror intervention. The ineffectiveness of traditional approaches led me to propose research to test an

alternative method of intervention, which seeks to enhance the ability of jurors to weigh eyewitness evidence (i.e., sensitivity). Specifically, I expect to sensitize jurors to the important role of initial eyewitness confidence judgments, the problem with confidence inflation, and the relatively small impact of system and estimator variables on the confidence-accuracy relationship.

1.1 Jurors' Behavior Regarding Eyewitness Evidence

There is no doubt that jurors attend to eyewitness confidence and use this information in their decision-making. In fact, research shows that this often may be the only variable that they attend to, as it was the only significant predictor of jurors' perceptions of the evidence and guilt, compared to nine other common system and estimator variables (Cutler, Penrod & Stuve, 1988; Cutler, Dexter & Penrod, 1990). Jurors also perceive highly confident witnesses as more credible, and tend to ignore testimonial inconsistencies of highly confident witnesses (Brewer & Burke, 2002). The confidence heuristic, or reliance on a third party's confidence as an indicator of accuracy, likely explains these results (Thomas & McFayden, 1995). Ironically, novices who are solving difficult problems or engaging in single-trial decision-making (like jurors) are the most likely to rely on the confidence heuristic. Of course, to the extent that the eyewitness is actually correct, this is not problematic. However, failure to distinguish between reliable and unreliable expressions of eyewitness confidence is problematic and can contribute to wrongful convictions. To reiterate, high confidence expressed immediately using a fair test is diagnostic of accuracy, but delayed, courtroom confidence judgments, or those arising from an unfair test, are often not diagnostic.

Few studies have directly examined the impact of courtroom confidence judgments on jurors' perceptions and verdict decisions. Bradfield and McQuiston (2004) manipulated whether confidence inflation occurred at trial and whether the defense attorney challenged the

eyewitness. The control condition had the highest rate of guilty verdicts because both the initial and courtroom confidence judgments were high. This suggests that jurors do intuitively understand that high confidence signals that a suspect is likely guilty. The mere inflation condition exhibited confidence inflation from initial to courtroom confidence judgments, but the defense attorney did not challenge this. This condition did have lower perceptions of defendant's guilt, eyewitness accuracy, and strength of the prosecution's case, but did NOT decrease guilty verdicts. The only condition with decreased guilty verdicts was the condition in which there was confidence inflation, but this was explicitly challenged by the defense attorney. Jones, Williams and Brewer (2008) reported that, although 88% of jurors considered confidence inflation to be a form of testimonial inconsistency, it did not impact their verdict decisions. Finally, Key et al. (accepted pending minor revisions) replicated and extended Bradfield and McQuiston's (2004) study, finding that jurors were seven times as likely to vote guilty when confidence was high than when it was low. But, importantly for the present research, Key et al. obtained this pattern even for courtroom confidence judgments and when confidence inflation occurred. All that mattered was if an eyewitness was *ever* confident, not that an eyewitness was initially confident.

These results indicate that, although jurors are attuned to the strong confidence-accuracy relationship (higher rates of guilty verdicts when confidence is high), they fail to appreciate the measurement nuances under which this relationship holds (first, fair test of memory). The outcome of this could be wrongful convictions based on an initially hesitant, but later highly confident, eyewitness testimony (i.e., confidence inflation), like Jennifer Thompson's low confidence initial ID of Ronald Cotton (Innocence Project, 2019; Garrett, 2011). This state of affairs necessitates intervention. In the next section, I review the traditional interventions that have tried to improve juror comprehension of eyewitness evidence.

1.2 Traditional Approaches to Intervention

The two most common interventions in the literature are testimony from an expert witness and pattern jury instructions delivered by judges. These instructions, or best practice guidelines, are usually written with the help of experts. Keep in mind that the new view of eyewitness memory, regarding the strong confidence-accuracy relationship under proper conditions, is only recently being endorsed. In fact, the latest surveys of experts revealed that most experts endorsed a weak confidence-accuracy relationship (Kassin, Tubb, Hosch & Memon, 2001; Simons & Chabris, 2011). As a result, interventions focused on the idea that eyewitness memory is fallible and further harmed by a number of system and estimator variables. This view of eyewitness memory, when shared with jurors, may not lead to enhanced decision-making. To foreshadow, the interventions that arose from this old-science view (Gronlund & Benjamin, 2018) of eyewitness memory tended to induce skepticism, or a lower proportion of guilty verdicts, compared to a control condition with no instructions. Of course, this conservative shift is only good if the suspect is innocent. The ideal intervention will increase sensitivity; that is, increase the ability to discriminate between accurate and inaccurate eyewitnesses. A brief history of these interventions follows.

Expert Witnesses. Experts can be hired by the prosecution, defense, or both. The expert is allowed to review the case materials and look for variables that may have impacted eyewitness memory (helpful and harmful), write reports, assist the attorneys with preparation for questioning and cross-examination, and/or testify to judges/jurors about the eyewitness variables relevant to the specific case at hand. The judge must pre-approve the testimony in a Daubert hearing, and will usually only do so if the expert demonstrates expertise in the relevant field (e.g., with peer-

reviewed publications), offers testimony "beyond the ken" of jurors' lay-knowledge, and has previously testified in similar cases.

Research shows that adversarial methods of expert testimony, where there are experts working for opposing counsel, can improve sensitivity. However, non-adversarial methods, where there is a single expert, induce skepticism without improving sensitivity (Cutler, Dexter, & Penrod, 1989). Cutler et al. suggested an explanation based on the elaboration likelihood model (Petty & Cacioppo, 2012), wherein information is processed at face value (uncritically) if the source is credible. A critical evaluation of the evidence is more likely to occur when the source is not viewed as credible. Viewed this way, having a single expert, a credible source, should not result in a critical evaluation of the evidence; instead, hearing about the variables that can harm eyewitness accuracy should result in lower guilty verdicts across the board. A similar effect might occur for pattern jury instructions.

Pattern Jury Instructions. The first set of pattern jury instructions were the Telfaire instructions (*U.S. v Telfaire*, 1972), which contained information about whether the witness had capacity and opportunity to view the culprit, whether the identification was a product of the witness' own recollection of the event, and whether the witness was credible. These instructions were the same for all cases, and judges exercised no discretion regarding whether to present the instructions or not. Research showed that these instructions either had no effect (Cutler, Penrod & Dexter, 1990) or induced skepticism across the board compared to a control condition (Greene, 1988).

Revised instructions, called the Henderson instructions, were created in New Jersey and contained information about approximately 20 variables that can affect eyewitness memory (New Jersey v. Henderson, 2011). Unlike the Telfaire instructions, the Henderson instructions

require judges to exercise discretion about which variables are relevant to the case at hand. These instructions have come under scrutiny in several news reports and scientific articles for inducing skepticism. For example, Papailiou, Yokum, and Robertson (2015) examined the effects of instruction type (Telfaire, Henderson) and eyewitness evidence strength (strong, weak) on verdict decisions. They predicted an interaction, such that the proportion of guilty verdicts should be highest in the Henderson + strong evidence condition and lowest in the Telfaire + weak evidence condition. What they found was that Henderson instructions induced even more skepticism than the Telfaire instructions in both the strong and weak evidence conditions, with only 12% guilty verdicts even in the strong evidence condition (see their Figure 1). In fact, the odds of a guilty verdict were 2.55 times higher in the Telfaire than the Henderson instructions. Dillon et al. (2017) reached similar conclusions.

Papailiou et al. (2015) proposed a few reasons for the induced skepticism, including: reduced comprehension due to the increased length of the Henderson instructions, poor operationalization of what "high" level of the variables means (e.g., what constitutes "high" stress?), and implicit trust of the credible expert (judge), whose mere reading of the instructions could lead jurors to believe that an eyewitness must be inaccurate. In my opinion, these results are not particularly surprising. If giving jurors information about five variables (Telfaire) induces skepticism, it is no wonder that giving information about 20 variables (Henderson) induces *even more* skepticism. However, a content analysis of these instructions, covered in the next section, points to a different explanation.

1.3 Content Analysis of Traditional Approaches

One could argue that the sheer amount of information presented to jurors is the reason for skepticism. Laypeople with no prior knowledge about eyewitness issues likely have difficulty

weighing and integrating the multitude of eyewitness variables in a given case to determine whether an eyewitness is accurate or not (Strawn & Buchanan, 1975; Strawn & Munsterman, 1981). Even trained legal professionals could struggle to integrate all of this information. To explore this idea further, I compared the information from five sets of best practice guidelines and instructions written for police officers, courts (judges), and jurors (see Table 1). The table contains the 20 most common eyewitness variables mentioned, organized into system and estimator variables. The five sets of guidelines and instructions are organized by the audience: police on the far left, courts for the middle three, and jurors on the far right. An "x" indicates that the variable is included in that set of guidelines or instructions.

Table 1

Information Contained in Police, Court, and Jury Instructions

	IACP (2010)	Yates (2017)	ABA (2004)	Utah (in prep)	NJ v. Henderson (2012)
System Variables					
Double-blind Admin	x	x	x	x	x
Lineup vs Showup	x	x	x	x	x
Filler Similarity	x	x	x	x	x
Lineup Presentation	x	x	x	x	x
Lineup Instructions	x	x	x	x	x
Single Suspect	x	x	x	x	x
Post-ID Feedback	x	x	x	x	x
Co-Witness Contamination	x	x		x	
# Fillers	x	x	x	x	x
Filler Selection Method	x	x	x	x	
Mugshot Exposure			x		x
Audio/Video Recording	x	x	x	x	
Estimator Variables					
Cross-Race ID			x	x	x
Exposure Time			x	x	x
Retention Interval			x	x	x
Stress			x	x	x
Weapon Focus			x	x	x
Intoxication			x		x
Distance					x
Lighting					x
Disguise					x
Confidence	x	x	x	x	x

Note: The "x" denotes that the variable is included in the instructions.

As the table shows, police are instructed about system variables (that are under their control) and eyewitness confidence. This makes sense, as the goal of police instructions is to ensure the proper collection of the eyewitness evidence. On the other hand, courts and jurors are instructed about the system variables, eyewitness confidence, plus estimator variables (around 20 variables in total). The goal of court and jury instructions is to inform them about factors that may influence the eyewitness' accuracy. Note that these groups play no part in the collection of

the eyewitness evidence, but they nevertheless are required to weigh the evidence and ultimately make critical decisions about it (suggestible eyewitness procedure, sentencing decisions).

Furthermore, consider the instructions about eyewitness confidence. Remember, jurors intuitively believe that confidence and accuracy are strongly related, and that is the one factor that influences their verdict decisions (not the nine other system and estimator variables; see Cutler et al., 1988; Cutler et al., 1990). That means that jurors' natural inclination already matches the new view of eyewitness memory—the importance of initial confidence, compared to other system and estimator variables. But a closer look at the instructions reveals that only the most recent instructions (Yates, 2017; Utah, in prep) portray this view. The instructions currently implemented in practice for the courts and jurors (ABA, 2004; *NJ v. Henderson*, 2012) state that there is a weak confidence-accuracy relationship. Given that this is contrary to what jurors believe, this might explain why the instructions induce skepticism.

To summarize, prior research by Bradfield and McQuiston (2004) and others show that jurors intuitively understand the strong relationship between confidence and accuracy, but are not sensitive to the measurement nuances (first, fair test). Traditional approaches like expert testimony and pattern jury instructions tend to induce skepticism rather than improving sensitivity. A content analysis revealed that the instructions contain information about approximately 20 variables in total (even though these variables hold much less predictive power than initial confidence), and portray the confidence-accuracy relationship contrary to what recent research shows (and what jurors intuitively believe). Given that in real cases the variables involved typically point to conflicting influences (i.e., some variables harm memory, but others help it), jurors likely struggle to integrate all of the variables together and make a coherent

decision about whether an eyewitness is accurate. It is no surprise that the current approaches are not working, and a different intervention is needed.

1.4 The Current Research

A simpler message, with fewer variables to consider, might be a better way to improve jurors' sensitivity to eyewitness evidence. Indeed, research by Pennington and Hastie (1991; 1992) showed that a simple, cohesive story about the evidence is the most compelling to jurors. The judgment and decision-making literature has also shown that decision aids that simplify the information (e.g., icon arrays, bar or line graphs, pie charts, checklists) improved decision-making in most studies (Cokely et al., in press; Garcia-Retamero & Cokely, 2013; Garcia-Retamero & Cokely, 2017; Hamm, Beasley & Johnson, 2014). Other research using a teaching aid for eyewitness evidence (i.e., a PowerPoint presentation) also improved sensitivity (Pawlenko et al., 2013).

The simple message about eyewitness memory is that there is a strong confidence-accuracy relationship under proper measurement conditions. Even variables believed to harm memory (system and estimator variables) do not appreciably distort this relationship (Semmler et al., 2018; for a review see Gronlund & Benjamin, 2018). Therefore, initial confidence from a first, fair test, appears to be the most important factor to consider when determining whether an eyewitness is accurate. The goal of the current research is to design an intervention that sensitizes jurors to this simplified message.

Specifically, I designed a novel intervention (hereafter, the Key Intervention) that presents the simplified eyewitness message as a visual aid, with supporting text instructions. This intervention was inspired by prior work from other domains showing the effectiveness of simplifying the message that a decision-maker receives either in story format (Pennington &

Hastie, 1991; 1992) or by decision aid (Garcia-Retamero & Cokely, 2013; 2017; Hamm et al., 2014). This novel Key Intervention will be compared to a control group—to determine whether this new intervention improves sensitivity compared to giving no eyewitness evidence instruction—and to a modified Henderson instructions, to determine whether the new intervention outperforms those that are currently used in real cases (which induce skepticism, Papailiou et al., 2015). My research excluded other traditional interventions: Telfaire instructions because they are longer used in most courts, and expert testimony because financial constraints often preclude its use. The Key Intervention, however, could be implemented at little cost. To reiterate, the goal of this research was to design an intervention that DOES work; the goal was NOT to figure out why traditional interventions do not work.

To that end, I conducted two studies. The first study tested whether the Key Intervention improved sensitivity to confidence inflation in a relatively simple eyewitness context. Although this study is a good first step in the determining whether a simplified message focused on confidence can be effective at sensitizing jurors, it also was important to determine whether the new instructions can improve sensitivity to the measurement nuances (first, fair test). Specifically, in situations where confidence is no longer predictive of accuracy (not a fair test), but is high, can the instructions sensitize mock jurors to ignore confidence? Therefore, Experiment 2 included a third factor, memory test fairness. The efficacy of the interventions was evaluated using a post-trial questionnaire, comprehension items, and usability and workload metrics. Demographic variables and cognitive ability measures were included as covariates.

Chapter 2: Experiment 1

2.1 Method

2.1.1 Participants

Participants were (N = 842) Amazon's Mechanical Turk jury-eligible workers who were compensated a small amount (<$5.00) for participation[1]. Only those who self-reported being 18+ years old with the ability to speak and read English at the high school level or better were allowed to complete the study. A total of 88 people did not progress past these criteria for participation, and 27 participants dropped out before completing any dependent measures (after which, n = 727 usable data points remained). Usable data points were excluded from analysis for malingering, operationalized as: completing the study in less than half the median total duration for the control condition and failure of two or more (of four) comprehension-check questions. This resulted in exclusion of n = 40 additional participants (~5% of the sample), similar to previous similar studies (Papailiou et al., 2015; Safer et al., 2016). After exclusion, a final sample of N = 687 participants remained for data analysis. This exceeded the ideal sample size of N = 640, calculated based on prior research showing that n = 80 per condition is sufficient to detect significant effects of interventions (Cutler et al., 1989). The Intuitional Review Board at the University of Oklahoma approved this research.

Not all participants reported demographic information, but of those who did, 42.6% were female (of n = 676), and the average age was 20.67 years (SD = 11.9; n = 675). Participants (n = 674) reported ethnicity as: Caucasian (65.0%), African American (15.7%), Hispanic/Latino (4.9%), Native American (1.5%), Asian/Pacific Islander (8.0%), Other (0.1%), No Response (1.0%), and Multiple (3.7%). Around one fourth of the sample (26.2%) reported being currently

[1] The research was funded by a University of Oklahoma Robberson Research Grant, which was partially matched by Psychology Department funds.

in college. Highest level of education was reported as (n = 676): High School/GED (7.4%), Some College (27.5%), Bachelor's Degree (47.3%), Master's Degree (14.3%), Doctoral/Professional Degree (2.1%), Other (0.3%), and No Response (1.0%). Reported political affiliations (n = 676) were: Very Conservative (8.7%), Conservative (17.2%), Moderate (21.9%), Liberal (30.6%), Very Liberal (18.5%), Other (1.0%), Don't Know (0.9%), and No Response (1.2%). Finally, yearly household income (n = 672) was reported as: $0-$30,000 (18.6%), $31,000-$60,000 (37.9%), $61,000-$90,000 (24.9%), $91,000 or more (16.4%), and No Response (2.2%).

2.1.2 Design

The study conformed to a 2 (Transcript Strength: Strong, Weak) x 3 (Intervention: Key Intervention, Modified Henderson, Control) between-participants design. In the Strong version of the transcript, the eyewitness was 90% confident initially and in court; in the Weak transcript, confidence inflated from 20% initially to 90% in court. The Key Intervention was a visual aid depicting a balance beam, on which confidence measured properly had more weight than confidence arising from any other circumstances. Text instructions supported this aid, stating in words that confidence is predictive of accuracy if measured properly, as well as how to interpret the balance beam[2] (see Appendix A). The original Henderson instructions were modified by simplifying the language, removing redundant sentences, and correcting the confidence-accuracy relationship instruction (the modified instruction matched the content regarding confidence in the Key Intervention, see Appendix C). These modifications reduced the word count from around

[2] Note: The balance beam was chosen as the visual aid because pilot testing revealed that it sensitized jurors more than icon arrays depicting the same information (see Appendix B).

2300 words in original Henderson to around 600 words in the modified Henderson. The Control condition contained no instructions regarding eyewitness evidence.

The primary dependent variables were verdict decisions, post-trial questionnaire items, comprehension check questions, usability and workload items, and total duration. Demographics (gender, age, ethnicity, education, political orientation, and income) and cognitive measures (numeracy and graph literacy) were also included as covariates, as these variables may relate to decision making and comprehension of visual aids (Garcia-Retamero & Cokely, 2017; Ybarra, 2018).

2.1.3 Materials

The mock trial, adapted from prior research (Bradfield & McQuiston, 2004; Key et al., accepted pending minor revisions), describes the fictitious case of *People v. Roger Sanchez*. Mr. Sanchez is charged with threatening and robbing Ms. Cameron of her purse, its contents, and $730 cash. Mr. Sanchez was arrested a short distance from the scene of the crime because he matched the general description of the culprit given to police. He had $700 in cash on his person. He was later identified by Mrs. Cameron from a police lineup as the man who stole her purse, and her confidence was expressed initially and in the courtroom. Mr. Sanchez claims that he was misidentified and that the cash he was carrying was from his paycheck.

The transcript begins with direct examination of the witness by the prosecution, and she is asked to describe the crime, the perpetrator, and give a courtroom confidence judgment (90%). This is followed by cross-examination by the defense attorney, during which the witness is asked to provide the confidence judgment she gave initially after ID (20% or 90%), and she is questioned about system and estimator variables that may have impacted her accuracy. A total of 10 system and estimator variables were embedded in the trial transcript, favoring the defense (6

variables harm eyewitness accuracy, 4 favor eyewitness accuracy and the prosecution). These were held constant in the strong and weak versions of the transcript, so that the only manipulated variable in the transcript was the eyewitness confidence inflation. See Appendix D for the full trial transcript.

The pre-deliberation instructions were adapted from the Henderson instructions (see Appendix E), reduced from 675 words to about 175.

The post-trial questionnaire items were adapted from prior research (Bradfield & McQuiston, 2004; Key et al., accepted pending minor revisions). The post-trial questionnaire included 17 questions regarding: verdict decisions, strength of the evidence and sentencing recommendation, the witnessing conditions, and beliefs about eyewitness memory in general. Verdict decision was dichotomous (1=guilty, 2=not guilty). All other questions were reported on Likert scales ranging from 1 to 10, where higher values on the scale favor the prosecution. See Appendix F for the post-trial questionnaire.

The comprehension check questions were adapted from prior research (Bradfield & McQuiston, 2004; Key et al., accepted pending minor revisions). There were four multiple-choice questions about basic case facts, including: type of crime, the defendant's name, whether the witness identified a suspect, and what the witness was threatened with. To ensure participants paid attention to the eyewitness' confidence judgments, they were asked to choose initial and courtroom confidence percentages from a list (10% increments), and rate whether the courtroom confidence was greater than, equal to, or less than the initial confidence. See Appendix G for comprehension check items.

Usability was measured with six items based on the System Usability Scale (SUS; Brooke, 1996; validated by Bangor, Kortum, Miller, 2008), where 0 = strongly disagree and 10 =

strongly agree. Workload was measured with the six items from the NASA TLX (Hart & Staveland 1988), where 0 = lowest, 10 = highest. Participants were asked to rank how they felt while weighing all the information presented to them about the case *People v. Sanchez*. See Appendix H for the usability and workload items.

Numeracy was measured with the Berlin Numeracy Test-Schwartz (BNT-S) (Cokely et al., 2012), which is a combination of the four BNT items and the three easy items from the Schwartz et al. (1997) test. Graph literacy was measured with the difficult subset of multiple-choice items from Woller-Carter (2015). Participants were allowed to use a calculator and/or scratch paper.

2.1.4 Procedure

Participants completed the study individually rather than together as real jurors would, and the entire study was self-paced. Participants gave informed consent and self-reported whether they were age 18 or older and spoke and read English at the high school level or better. Anyone who did not meet these inclusion criteria were not allowed to continue. Participants were instructed to pay close attention to the trial transcript, because their responses could influence the outcome of real cases. On the same page, they read a summary of the case facts before proceeding to the transcript. The transcript was divided into several sections, each of which appeared on a different page of the survey. Participants read the trial transcript, pre-deliberation instructions and received the intervention. Then they completed the post-trial questionnaire items, comprehension check items, usability, workload, cognitive ability measures and demographic items. Finally, they were thanked for their participation and received payment. Due to experimenter error, the cognitive ability measures and some of the post-trial questionnaire items were only completed by a subset of participants (n = 200).

2.2 Results

I expected that the Key Intervention would improve juror sensitivity compared to the Modified Henderson instructions, which should induce skepticism, and the Control condition, which should show no difference across Strong and Weak evidence conditions. Other metrics, including comprehension, usability, and workload, were also explored.

The data were analyzed using a series of regression models (logistic regression for verdict decision, linear regression for all other dependent variables). To maintain a family-wise error rate of $\alpha = .05$, a Holm-Bonferroni correction was used for each model; the *p* values reported in this manuscript are the adjusted *p* values. For all models, the Key Intervention condition served as the reference class, to test the hypothesis that this condition outperforms the other two. The results are reported first for the dependent measures completed by all participants in the study (Full Sample, $n = 687$), and then for the measures that only a subset of participants completed (Subset, $n = 200$). Only the findings of interest are reported in text, but full reporting of each model and graphs displaying results are contained in Appendix I (Full Sample) and Appendix J (Subset). A correlation matrix for the covariates is reported in Table 2.

Table 2.
Correlation Matrix of Covariates

	Numer	Grap	Gend	Age	Ethnici	Educat	Political	Inco
Numeracy	1	0.54	-0.12	0.17	-0.14	0.02	0.08	0.01
GraphLiteracy	0.54	1	0.05	0.13	-0.2	-0.02	0.12	0.03
Gender	-0.12	0.05	1	0.12	0.12	0.11	0.17	0.14
Age	0.17	0.13	0.12	1	-0.16	-0.02	-0.01	-0.07
Ethnicity	-0.14	-0.2	0.12	-0.16	1	0.39	0.25	0.16
Education	0.02	-0	0.11	-0.02	0.39	1	0.13	0.42
Political Orientation	0.08	0.12	0.17	-0.01	0.25	0.13	1	0.1
Income	0.01	0.03	0.14	-0.07	0.16	0.42	0.1	1

2.2.1 Verdict Decisions

Full Sample. A logistic regression was used to determine whether verdict was predicted by the manipulations and demographic variables. A log-likelihood test evaluated the overall fit of the model. The overall model was significant, indicating differences in verdict across conditions, χ^2 (df = 12) = -410.04, $p < .0001$, AIC = 844.09, $r^2 = .12$. Verdict varied significantly by Transcript Strength, $B = -1.64$ ($SE = .29$), $z = -5.49$, $p = < .0001$: Jurors voted guilty more in the Strong (65%) than the Weak (36%) transcripts, indicating sensitivity to the differences in eyewitness evidence quality across the two conditions. The Key Intervention did have a higher rate of guilty verdicts (54%), but this did not differ significantly from the Modified Henderson (46%) or the Control condition (50%), $ps > .68$, see Figure 1. Verdict also varied as a function of Ethnicity, $B = .21$ ($SE = .07$), $z = 3.01$, $p = .03$ and Political Orientation, $B = -.25$ ($SE = .06$), $z = -3.89$, $p = .001$. Those of minority ethnicity voted guilty less often, as did those with liberal political orientation. No other significant differences emerged, including interactions between Transcript Strength and Intervention.

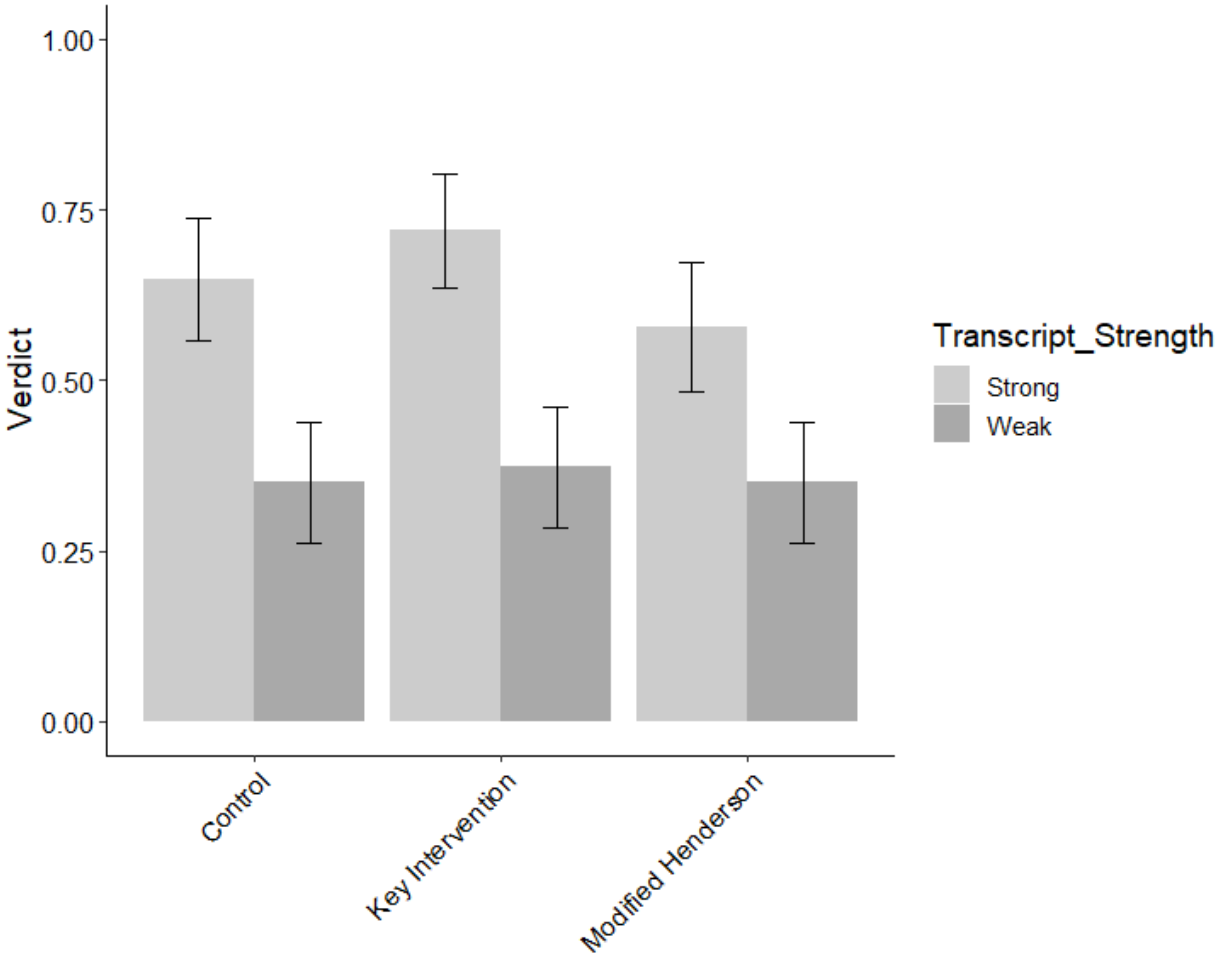


Figure I.1. Average VERDICT as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals. Note: this figure reappears in Appendix I for convenience.

Subset of Sample. Adding numeracy and graph literacy to the model did not change its overall fit, and these variables did not emerge as significant predictors of verdict decisions, *p*s > .20. In fact, no predictors were significant, *ps* > .11.

2.2.2 Post-Trial Questionnaire Items

Full Sample. The post-trial questionnaire item results tell a similar story. Intervention did not appreciably predict averages for any post-trial questionnaire item. However, Evidence

Strength significantly predicted average rating for Likelihood of Guilt and Confidence in Witness: There were higher averages (favoring the prosecution) in the Strong than Weak evidence condition. Interestingly, Evidence Strength also influenced Courtroom Confidence Predicts Verdict (strong $M = 7.00$, $SD = 2.63$; weak $M = 5.59$, $SD = 2.91$), but not Initial Confidence Predicts Verdict (strong $M = 7.30$, $SD = 2.55$; weak $M = 7.60$, $SD = 2.41$). The two takeaways from these results are: 1) participants are more persuaded by courtroom confidence in the strong evidence condition (perhaps because it confirms the initial confidence judgment, rather than conflicting with it), and 2) participants have higher averages for (are more influenced by) initial than courtroom confidence judgments made by the witness. In other words, they appear to be already sensitive to confidence inflation, and this could explain why the Key Intervention had little impact.

Ethnicity and Political Orientation were also significant predictors for the post-trial questionnaire items, with minority ethnicity and liberal political orientation giving ratings that favored the defense, on average. For Confidence Inflation Equals Accuracy, Education emerged as a significant predictor ($p = .04$), suggesting that those higher in education are more knowledgeable about how confidence inflation over time signals a witness less likely to be accurate. The only variable that had no significant predictors was Confidence in Verdict, indicating that jurors' confidence in their own decision-making did not vary as a function of the manipulations or any demographic variables.

Subset of Sample. Remember, some post-trial questionnaire items were not included for all participants, so a richer story is revealed for the subset of participants who completed all items. Transcript Strength was a significant predictor of the post-trial questionnaire items regarding witness confidence and memory, including: Courtroom Confidence Influenced

Verdict, Confidence in Witness, and Witness Memory. This is evidence of improved sensitivity for the variables that were the focus of the Key Intervention, in support of my hypothesis. But there were no significant predictors of witnessing conditions (e.g., Attention, Good Basis for ID), Case Strength, ratings of Witness Accuracy, and beliefs about eyewitnesses in general (Eyewitness Accuracy General, Confidence Predicts Accuracy, or Confidence Inflation Occurs). This makes sense, as these were not part of the Key Intervention, so improved sensitivity should not have occurred.

Including cognitive abilities in the models negated the predictive power of all demographic variables, except that Political Orientation predicted Sentence Recommendation (with heavier sentences given by those who identified as conservative). Graph literacy emerged as a significant predictor of Sentence Recommendation, Witness Accuracy, Attention, Witness Memory and Confidence Inflation Equals Accuracy. These findings may signal that those higher in graph literacy evaluated the witness in light of diagnostic information (i.e., her confidence), whereas those lower in graph literacy may not have. Numeracy did not predict any post-trial questionnaire items.

2.2.3 Comprehension Check Questions

Full Sample. For comprehension check questions, the overall model was significant, indicating significant differences in comprehension, F (11, 656) = 6.38, $p < .0001$, adjusted r^2 = .08. Comprehension did not vary as a function of Transcript Strength or Intervention, or their interactions. This finding makes sense, as the comprehension questions pertained to information presented in the trial, which participants read before the intervention (thus, intervention should

not have any bearing on comprehension scores). Comprehension scores were higher with Age, lower with Education[3], and higher with more liberal Political Orientation.

Participants were also asked to indicate the witness' initial and courtroom confidence judgment as a percentage. Transcript Strength significantly predicted Initial Confidence but not Courtroom Confidence. Remember, initial confidence percentage was the key manipulation of transcript strength, so these findings indicate that participants paid attention to the witness' confidence judgments.

Subset of Sample. Transcript Strength significantly predicted Initial Confidence Percentage and a judgment of whether the courtroom confidence judgment was higher, lower, or equal to initial confidence. Transcript Strength did not predict Courtroom Confidence Percentage itself, although Graph Literacy did, $B = .11$ ($SE = .03$), $t = 3.27$, $p = .02$). Graph Literacy also predicted comprehension score, $B = .04$ ($SE = .01$), $t = 3.89$, $p = .002$; Numeracy did not.

2.2.4 Usability and Workload

Full Sample. The overall model for usability was significant, F (11, 657) = 12.29, $p < .0001$, adjusted $r^2 = .16$; the overall average usability score (*B0*) was 7.44 ($SD = .40$) on a 11-point scale. Usability did not differ as a function of Transcript Strength, Intervention, or their interactions. However, usability scores were higher with Age, lower with Education, and higher with more liberal Political Orientation. The same pattern occurred for workload, overall model fit F (11, 657) = 11.97, $p < .0001$, adjusted $r^2 = .15$. The overall average workload score (*B0*)

[3] The reader may be wondering why the coefficient for education was negative, indicating a potential suppressor effect. However, a single regression model using education as the sole predictor of comprehension also yielded a negative coefficient, $B = -0.12$ ($SE = .03$), $t = -4.98$, $p < .0001$. In fact, this negative coefficient of education occurred for comprehension, usability, and workload (Experiment 1 and 2) even when it was the sole predictor in the model. The explanation for this negative coefficient could be explored in future research.

was 5.72 (*SE* = .41) on a 11-point scale. Workload did not differ as a function of Evidence Strength, Intervention, or their interactions. However, higher workload scores were associated with lower Age, higher Education, and more conservative Political Orientation.

Subset of Sample. Adding Numeracy and Graph Literacy as predictors did not change the null findings for the manipulated variables. However, the effects of the demographic variables were no longer significant, and Graph Literacy emerged as a significant predictor of Usability, *B* = .15 (*SE* = .05), *t* = 3.14, *p* = .03, and of Workload, *B* = -.19 (*SE* = .05), *t* = -4.00, *p* = .001. In sum, the averages indicate reasonable usability and workload for all of the interventions, and only demographic variables and graph literacy (not manipulations) predicted the scores.

2.3 Discussion

The purpose of Experiment 1 was to compare a novel intervention (i.e., the Key Intervention) presenting a simplified message about eyewitness evidence using a visual aid, to a traditional jury intervention, which induces skepticism, and a Control condition with no eyewitness information. This was done in a relatively simple trial context focused on eyewitness evidence that manipulated evidence strength through confidence (consistently high—90%—in the Strong condition, inflated in the Weak condition—20% to 90%).

I expected the Key Intervention to improve juror sensitivity compared to the Modified Henderson instructions, which should induce skepticism, and the Control condition, which should show no difference across Strong and Weak evidence conditions. Although there were trends in these directions, there were no significant differences across Interventions for any of the dependent variables. Instead, participants appeared to be already sensitive to the evidence, favoring the prosecution much more given the Strong than the Weak transcript. This might be due to the magnitude of the inflation, which was exaggerated compared to what had been used in previous research (50% to 100% used by Bradfield & McQuiston, 2004; Key et al., accepted

pending minor revisions, rather than 20% to 90% here). In the current study, perhaps the exaggerated inflation used here increased jurors' sensitivity to the confidence evidence, leaving little room for the Intervention to have an impact.

To provide a stronger test of the hypothesis that the Key Intervention should outperform traditional interventions, Experiment 2 extended this work by mimicking the confidence inflation used in previous research (50% to 90%). My expectation was that this would reduce the sensitivity that participants demonstrated in Experiment 1, thus allowing more room for the interventions to have an impact. Moreover, Experiment 2 used a more ambiguous and ecologically valid trial context, including the manipulation of memory test fairness (expressed via the testimony of the investigating law enforcement officer). I hypothesized that if the Key Intervention really does work, mock jurors should be sensitive to the confidence inflation, such that they favor the prosecution more given the Strong than Weak transcript. But importantly, this sensitivity should only occur when the memory test is fair. Given an unfair memory test, the Strong and Weak transcripts should yield similar favor to the prosecution, because the initial confidence did not arise from a fair memory test, and therefore the reliance on confidence should be downplayed even if there is no confidence inflation.

Chapter 3: Experiment 2

3.1 Method

3.1.1 Participants

Participants were (N = 945) jury-eligible adults of Qualtrics panel members and Amazon's Mechanical Turk workers who were compensated a small amount (< $5.00) for

participation[4]. Only those who self-reported being age 18+ years old with the ability to speak and read English at the high school level or better were allowed to complete the study. A total of 87 people did not progress past these criteria for participation, and 26 participants dropped out before completing any dependent measures (after which, n = 831 usable data points remained). Usable data points were excluded from analysis for malingering, operationalized as: completing the study in less than half the median total duration for the control condition and failure of two or more (of four) comprehension-check questions. This resulted in exclusion of n = 41 participants (~5% of the usable sample, similar to previous similar studies, Papailiou et al., 2015; Safer et al., 2016). After exclusion, a final sample of N = 790 participants remained for data analysis. This fell short of the ideal sample size of N = 960, calculated based on prior research showing that n = 80 per condition is sufficient to detect significant effects of interventions (Cutler et al., 1989)[5]. The Intuitional Review Boards at University of Oklahoma and University of Alabama in Huntsville approved this research.

Not all participants reported demographic information, but of those who did, 52.2% were female (n = 740), and the average age was 22.18 years (SD = 13.02; n = 736). Participants (n = 738) reported ethnicity as: Caucasian (65.7%), African American (10.6%), Hispanic/Latino (6.4%), Native American (1.2%), Asian/Pacific Islander (10.2%), Other (0.7%), No Response (1.2%), and Multiple (4.1%). Around one-fourth of the sample (27.6%) reported being currently in college. Highest level of education was reported as (n = 740): High School/GED (8.2%), Some College (28.2%), Bachelor's Degree (41.4%), Master's Degree (17.6%),

[4] The research was funded by University of Oklahoma Robberson Research Grant, partially matched by psychology department funds, and an internal research grant from University of Alabama in Huntsville.
[5] Data collection was terminated at the exhaustion of grant funding.

Doctoral/Professional Degree (2.8%), Other (0.7%), and No Response (1.1%). Reported political affiliations (*n* = 739) were: Very Conservative (8.8%), Conservative (21.0%), Moderate (25.0%), Liberal (27.9%), Very Liberal (12.9%), Other (.8%), Don't Know (1.1%), and No Response (2.6%). Finally, yearly household income (*n* = 733) was reported as: $0-$30,000 (21.1%), $31,000-$60,000 (33.7%), $61,000-$90,000 (23.9%), $91,000 or more (17.9%), and No Response (3.4%).

3.1.2 Design

The study employed a 2 (Transcript Strength: Strong, Weak) x 3 (Intervention: Key Intervention, Modified Henderson, Control) x 2 (Memory Test: Fair, Unfair) between-participants design. Again, Transcript Strength was manipulated via eyewitness confidence ratings, but in this experiment confidence inflated from 50% to 90% in the Weak condition; confidence was consistently high at 90% in the Strong condition. The Fair memory test includes a fair lineup and double-blind lineup administration; the Unfair memory test includes a biased lineup (in which the suspect is described as standing out) and single-blind lineup administration.

3.1.3 Materials

The materials were the same as those used in Experiment 1, with two exceptions. First, the confidence inflation was less exaggerated (50% to 90%), mimicking previous research (Bradfield & McQuiston, 2004; Key et al., accepted pending minor revisions). Second, the transcript included testimony of the investigating law enforcement officer. This testimony focused on how the lineup was created and administered; this is how Memory Test fairness was manipulated.

3.1.4 Procedure

The procedures were the same as those used in Experiment 1.

3.2 Results

I expected that the Key Intervention would improve juror sensitivity compared to the Control condition and Modified Henderson instructions (which should induce skepticism). However, if the instructions truly sensitize jurors, this pattern should be observed for the fair memory test only—not for the unfair memory test. In the unfair memory test condition, the "strong" confidence evidence did not arise from a proper test of memory, so there should be no difference between the Strong and Weak evidence conditions.

The data were analyzed using a series of regression models (logistic regression for verdict decision, linear regression for all other dependent variables). To maintain a family-wise error rate of $\alpha = .05$, a Holm-Bonferroni correction was used for each model; the *p* values reported in this manuscript are the adjusted *p* values. For all models, the Key Intervention served as the reference class to test the hypothesis that this condition improved sensitivity compared to the others. Only the findings of interest are reported in text, but see Appendix K for the full reporting of each model with graphs displaying the results. See Table 3 for a correlation matrix of covariates.

Table 3.
Correlation Matrix of Covariates

	Nume	Graph	Gend	Age	Ethnic	Educat	Political	Inco
Numeracy	1	0.54	-0.01	0.07	-0.02	-0.01	0.16	0.11
Graph Literacy	0.54	1	0.04	0.11	-0.01	-0.12	0.21	0.08
Gender	-0.01	0.04	1	0.14	-0.03	-0.02	0.09	-0.09
Age	0.07	0.11	0.14	1	-0.15	0.02	-0.05	0.01
Ethnicity	-0.02	-0	-0.03	-0.15	1	0.15	0.17	0.09
Education	-0.01	-0.1	-0.02	0.02	0.15	1	-0.02	0.31
Political Orientation	0.16	0.21	0.09	-0.05	0.17	-0.02	1	0.04
Income	0.11	0.08	-0.09	0.01	0.09	0.31	0.04	1

3.2.1 Verdict Decisions

Full Sample. A logistic regression was used to determine whether verdict was predicted by the manipulations and covariates. A log-likelihood test evaluated the overall fit of the model. The overall model was significant, indicating differences in verdict across conditions, χ^2 (df = 20) = -445.62, $p < .0001$, AIC = 931.23, $r^2 = .11$. There was an effect of Transcript Strength, as verdicts were higher in the Strong (54%) than Weak (39%) transcripts, $B = -1.49$ ($SE = .41$), $z = -3.68$, $p < .0004$. At first glance, this result could be taken to mean that the change to 50% confidence in the trial transcript did not have the intended effect of reducing sensitivity. But as can be seen in Figure 2, the difference between Strong and Weak is limited to the Key Intervention in the Fair memory test (confidence intervals do not overlap). This supports my primary hypothesis that the Key Intervention would sensitize jurors to confidence inflation only for the Fair memory test, whereas the other two interventions would not induce increased sensitivity.

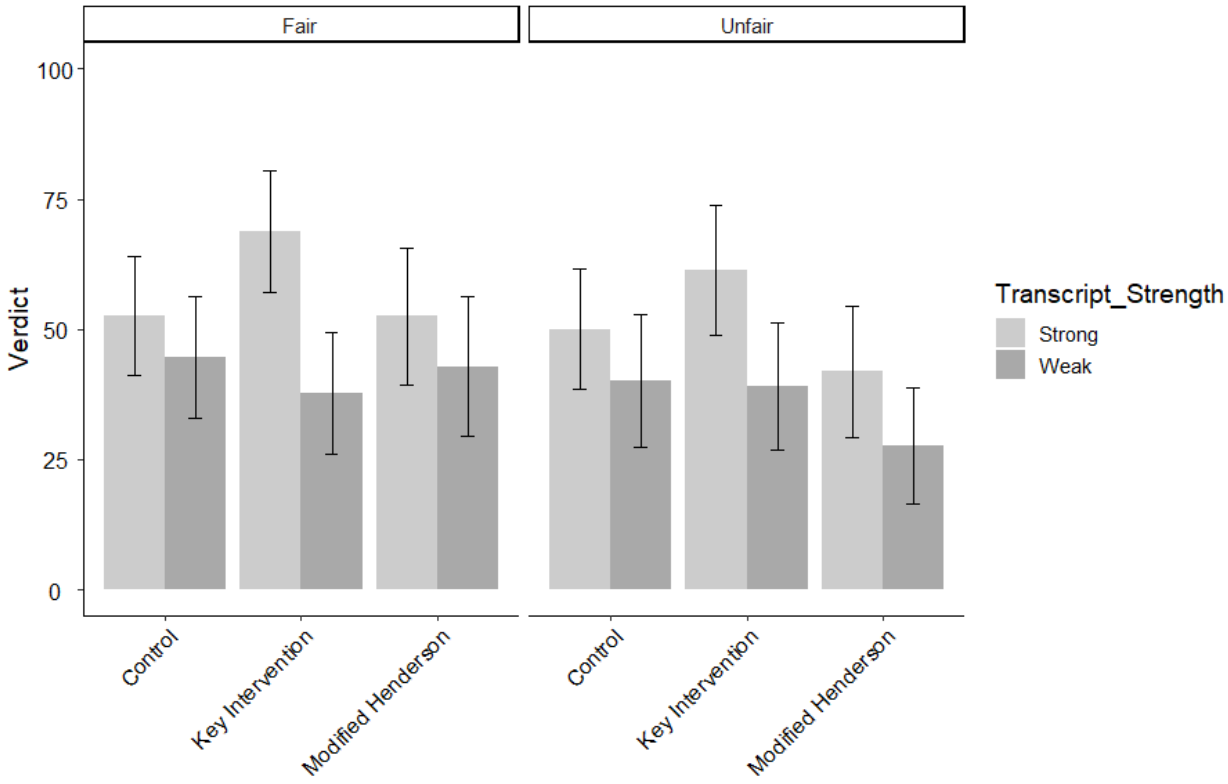


Figure K.1. Average VERDICT percentage as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals. Note: Figure reappears in Appendix K for convenience.

As further evidence of this conclusion, when collapsing across Transcript Strength and Intervention type the other experimental manipulations, verdict did not differ between Fair (50%) and Unfair (43%) memory test conditions, B = -.37 (SE = .42), z = -.10, p > .05. Moreover, collapsing across Memory Test Fairness and Transcript Strength, the Key Intervention did have a higher rate of guilty verdicts (51%), although it did not differ significantly from the Modified Henderson (41%) or the Control condition (47%), Bs < -.79, zs < -2.00, ps > .58. Only when the Key Intervention is paired with the Fair memory test does improved sensitivity occur as predicted. No other significant differences emerged, including 2-

way and 3-way interactions, except for one covariate, Graph Literacy, B = -.13 (SE = .03), z = -4.78, p < .0001.

3.2.2 Post-Trial Questionnaire Items

The post-trial questionnaire item results tell a similar story. The experimental manipulations did not appreciably predict averages for most post-trial questionnaire items, although the averages for many items are higher (non-overlapping CIs) in Strong than Weak for the Key Intervention. The Key Intervention improved understanding for Initial Confidence Influenced Verdict compared to the control condition, B = 1.40 (SE = .45), t = 3.07, p = .04. This provides some evidence that the Key Intervention sensitized jurors to the pertinent information (i.e., initial confidence).

Political Orientation was a significant predictor of most post-trial questionnaire items. Conservative political orientation was associated with higher agreement with questions pertaining to the eyewitness in this case and to general eyewitness beliefs, signaling a bias favoring the prosecution. Graph Literacy also emerged as a significant predictor of most post-trial questionnaire items, with ratings favoring the prosecution being associated with a lower Graph Literacy score. Numeracy significantly predicted Confidence in Verdict, but did not predict any other post-trial questionnaire items.

3.2.3 Comprehension Check Questions

For comprehension check questions, the overall model was significant, indicating significant differences in comprehension, F (19, 707) = 15.34, p < .0001, adjusted r^2 = .27. Average comprehension ($B0$) was 2.89 out of 4 (SE = .12), t = 24.09, p < .0001. Comprehension did not vary as a function of Transcript Strength, Lineup Fairness, Intervention, or their interactions. This is not surprising, because the questions pertained to the basic case facts presented before the

experimental manipulations. Comprehension scores were lower with Education, B = -.07 (SE = .02), t = -3.75, p = .003, and higher with Graph Literacy, B = .06 (SE = .006), t = 9.70, p < .0001.

To ascertain whether participants paid attention to the witness' confidence judgment, they were asked to indicate initial and courtroom confidence of the witness, as well as make a rating of whether the courtroom confidence judgment was higher, lower, or equal to initial confidence. There were significant differences across Strong and Weak transcripts for initial confidence percentage and the ranking, but not courtroom confidence. As the transcript did contain differences in confidence initially but not in the courtroom, this indicates participants attended to the confidence judgments. This may be particularly the case for those higher in Graph Literacy, as it was also a significant predictor for Initial and Courtroom Confidence percentages.

3.2.4 Usability and Workload

The overall model for usability was significant, F (19, 707) = 18.63, p < .0001, adjusted r^2 = .32; the overall average usability score ($B0$) was 6.28 on a 11-point scale (SE = .38), t = 16.75, p < .0001. Usability did not differ as a function of Transcript Strength, Lineup Fairness, Intervention, or their interactions. However, usability scores were higher with Age, B = .02 (SE = .005), t = 3.97, p = .001, lower with minority Ethnicity, B = -.14 (SE = .04), t = -3.50, p = .008, and lower with Education, B = -.28 (SE = .06), t = -4.76, p < .0001. Higher usability was also associated with higher Graph Literacy, B = .20 (SE = .02), t = 10.48, p < .0001.

The same pattern occurred for workload, overall model fit, F (19, 707) = 23.63, p < .0001, adjusted r^2 = .37. The overall average workload score ($B0$) was 8.59 on a 11-point scale (SE = .37), t = 22.78, p < .0001). Workload did not differ as a function of Transcript Strength, Lineup Fairness, Intervention, or their interactions. However, higher workload scores were associated with lower Age, B = -.01 (SE = .005), t = -3.02, p = .039, minority Ethnicity, B = .14

(*SE* = .04), *t* = 3.51, *p* = .008, higher Education, *B* = .38 (*SE* = .06), *t* = 6.41, *p* < .0001, and more conservative Political Orientation, *B* = -.25 (*SE* = .04), *t* = -6.12, *p* < .0001. Higher workload was also associated with lower Graph Literacy, *B* = -.22 (*SE* = .02), *t* = -11.50, *p* < .0001.

3.3 Discussion

The purpose of this study was to determine whether the Key Intervention significantly improved sensitivity, but only when the memory test was fair. This should be the case because confidence arising from an unfair memory test is not a reliable indicator of accuracy, and should not be weighted more heavily in the Strong (90% confidence consistently) than Weak transcript (confidence inflation from 50% to 90%). As expected, the change in confidence inflation (from 20% to 90% in Experiment 1 to 50% to 90% in Experiment 2) greatly reduced the differences between Strong and Weak transcripts observed in Experiment 1. There also were no differences across memory test fairness. This reduced sensitivity across the board allowed room for the intervention to play a role. Specifically, in line with my primary hypothesis, the Key Intervention did show improved sensitivity in verdicts (difference between strong and weak) for the Fair memory test only, whereas the other two interventions did not. This pattern also held for some key post-trial questionnaire items, with ratings favoring the prosecution only when the evidence was actually strong (Fair test).

Most covariates were not statistically associated with the dependent variables of interest, with a few exceptions of note. Age, Ethnicity, and Education predicted usability and workload,. Political Orientation was a significant predictor of most post-trial questionnaire items, with conservative political affiliation being associated with higher agreement on questions pertaining to witness-specific and eyewitness-general beliefs—signaling a bias favoring the prosecution.

Graph Literacy also emerged as a significant predictor of most post-trial questionnaire items (as well as verdict), with prosecution-favoring ratings associated with lower Graph

Literacy. This may indicate that those higher in Graph Literacy evaluated the witness using diagnostic information about her (i.e., her confidence), whereas those lower in graph literacy may not have. Moreover, numeracy was shown to be a significant predictor of Confidence in Verdict, replicating prior work showing that numeracy predicts decision-making confidence (Cokely et al., in press).

Chapter 4: General Discussion

Contrary to the widely touted view that eyewitness memory is not reliable and that eyewitness confidence can be high even when inaccurate, a newer body of literature modifies this conclusion: an initially confident witness is likely to be accurate, compared to a low confidence witness, assuming proper measurement. Even some system and estimator variables that are thought to distort accuracy do not appreciably impact the confidence-accuracy relationship (Semmler et al., 2018, Wixted & Wells, 2017). Research from Cutler and colleagues (Cutler et al., 1988; Cutler et al., 1990) shows that jurors already focus on eyewitness confidence and disregard other variables when making verdict decisions. However, they are not sensitive to the measurement nuances, as their verdict decisions are not impacted by *when* the confidence judgment occurred (initially or not) nor by confidence inflation across time (Bradfield & McQuiston, 2004; Key et al., accepted pending minor revisions). Moreover, the traditional approaches to intervention, like expert testimony and pattern jury instructions, do not improve sensitivity.

A content analysis of these approaches revealed two possible explanations for jurors' skepticism, not sensitivity: (a) the instructions contain information about ~20 eyewitness system and estimator variables, which may be difficult to integrate into a coherent conclusion about eyewitness accuracy, particularly when the variables point to conflicting conclusions; (b) the

instructions contain the mistaken message that confidence and accuracy are not related, which conflicts with jurors' intuitive understanding. To overcome these issues, I tested the idea that a simpler message, focused on initial confidence and disregarding the other system and estimator variables, may improve sensitivity compared to the traditional interventions. Specifically, I designed a new intervention (the Key Intervention), which presented the simplified message using a visual aid with supporting text instructions. This novel intervention was pitted against a modified version of the Henderson instructions currently used in real court cases, and a control condition with no eyewitness instruction. This was done in a relatively simple criminal trial context, where confidence inflation was the only difference between strong and weak evidence conditions (Experiment 1) and in a more complex scenario where memory test fairness was also manipulated (Experiment 2).

Experiment 1 showed that, surprisingly, mock jurors were already sensitive to confidence inflation. Thus, there was no room for intervention type to appreciably impact verdicts or other measures. I hypothesized that this was due to the exaggerated inflation (20% to 90%) in the current study compared to previous research showing no sensitivity (Bradfield & McQuiston, 2004; Key et al., accepted pending minor revisions). Thus, Experiment 2 modified the research by reverting to confidence inflation similar to what was used by previous research (50% to 90%), and expanded the research by adding a manipulation of memory test fairness. If the Key Intervention really sensitizes jurors, sensitization should occur only when the confidence arose from a fair memory test. The methodology change was an improvement, as the sensitivity effect from Experiment 1 was reduced, making room for the interventions to have an effect. Specifically, in support of my primary hypothesis, verdict and some post-trial questionnaire

items were higher in Strong than Weak in the Key Intervention, and importantly, only for Fair memory test.

There were no significant differences between evidence conditions (i.e., a lack of sensitivity) for the modified Henderson and the control condition. This is in line with the hypothesis that these two interventions would not improve sensitivity. Somewhat surprisingly though, the modified Henderson instruction did not induce skepticism (lower verdicts across all conditions), as expected. This could be because the modifications I undertook to the Henderson instructions by reducing the length and simplifying the legal terminology included in the original Henderson, either of which could have been driving the skepticism effect normally found (Papailiou et al., 2015). But the modified Henderson still included several secondary details about eyewitness memory (i.e., various system and estimator variables), which is likely why this condition did not improve sensitivity. Comprehension may also have been positively affected by these modifications, as there were no differences across intervention type for these metrics.

The goal of these studies was to design an intervention that significantly outperformed those currently used in real cases (original Henderson and expert witnesses). These studies provide some evidence that the Key Intervention might satisfy that aim. As expected, verdicts were higher in the Strong than Weak eyewitness evidence condition, but only when the memory test was fair. This pattern also held for some post-trial questionnaire items. However, the Key Intervention did not improve the other metrics of interest, including usability, workload, or comprehension compared to the other interventions. This might be because the modifications to Henderson improved understanding of the information compared to the original Henderson, such that this condition did not perform less favorably than the others. Overall, these findings suggest that the Henderson modifications may have resulted in an intervention that, although not as good

as the Key Intervention, perhaps is better than what is currently used (although this would need to be tested explicitly by comparing how the modified Henderson fares to the original).

Of course, the present studies are not without limitations. These studies manipulated evidence strength via the eyewitness confidence judgment (consistently high versus confidence inflation). Future research could manipulate other eyewitness variables (i.e., those not included in the current studies), or could even include other types of evidence (confession, forensic evidence). These other variables may well trump eyewitness confidence, and this is important to know. Another major limitation is that some system variables (e.g., double-blind administration, fair lineup creation) may be necessary to ensure that the confidence judgment came from a first, fair test. Instructing jurors to disregard these factors (as the Key Intervention does) may actually harm their decision-making, an idea that will be important to test in future research. Despite these limitations, the current studies add to the literature by testing a novel approach to jury intervention, and provides some evidence for the Key Intervention being a successful approach to enhancing sensitivity to eyewitness evidence, at least under the tested circumstances.

Adoption of the Key intervention in the legal system is uncertain though, because each jurisdiction chooses its own methodology for court proceedings. Its utility may also be impacted by individual differences among jurors (i.e., demographics and cognitive abilities), and overcoming this barrier will be difficult, because the legal standard is to select a jury of peers. As an alternative to intervening with all jurors who need to deliberate about the evidence, it may be easier to intervene on behalf of judges. Specifically, judges are expected to be gatekeepers of eyewitness evidence, disallowing overly suggestive eyewitness procedures. Can judges reliably apply higher standards regarding suggestibility by admitting eyewitness confidence evidence only if it arose from fair testing circumstances? This would require clear criteria for what

constitutes a fair memory test, which could be presented to judges using a decision aid. A checklist of criteria, a decision tree, or the balance beam used in the Key Intervention, may be viable decision aids to use in this context. However, what constitutes fair is still under debate in the scientific literature (see Wixted & Wells, 2017, and a response by Mickes, Clark & Gronlund, 2018, for a discussion of some basic criteria). Moreover, recent evidence suggests that subject factors (e.g., self-efficacy; face recognition ability) may account for at least as much variance in confidence as does accuracy (Grabman, Dobolyi, Berelovich, & Dodson, 2019; Kantner & Dobbins, 2019). In other words, the robustness of the confidence-accuracy relationship is under scrutiny. Perhaps in the future, judges can disallow evidence from eyewitnesses whose subject factors predict they have poor face memory or are not well calibrated; but exactly *which* subject factors matter is still being explored. Thus, it is unclear when the judges-as-gatekeepers solution could be implemented, or whether it would even be effective. However, the Key Intervention could be implemented immediately, and sensitizes mock jurors to the most important eyewitness evidence researchers agree on now (confidence inflation).

In conclusion, these studies designed and tested a novel approach to juror intervention, utilizing a visual aid and simplified message regarding eyewitness evidence. This new intervention, the Key Intervention, effectively sensitized mock jurors to confidence inflation of the eyewitness. Importantly, it also sensitized them to the measurement nuances necessary to ensure the reliable recording of eyewitness confidence judgments. This intervention seems to be a significant improvement over current approaches used by the legal system, and if implemented, would be expected to improve jury decision making in criminal cases involving eyewitness evidence.

References

ABA. (2004). Statement of best practices for promoting the accuracy of eyewitness identification procedures.

Bradfield, A., & McQuiston, D. E. (2004). When does evidence of eyewitness confidence inflation affect judgments in a criminal trial?. *Law and Human Behavior*, *28*, 369-387.

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: an archival analysis. *Law and Human Behavior, 25*, 475-491.

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: a reality monitoring analysis. *Law and Human Behavior, 29*, 279-301.

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364.

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

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry, 189* (194), 4-7.

Cokely, E.T., Feltz, A., Ghazal, S., Allan, J.N., Petrova, D., & Garcia-Retamero, R., (2016). Decision making skill: From intelligence to numeracy and expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (2nd Eds.), *Cambridge Handbook of Expertise and Expert Performance.* New York, NY: Cambridge University Press.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7*(1), 25-47.

Cutler, B. L., Dexter, H. R., & Penrod. S. D. (1989). Expert testimony and jury decision making: An empirical analysis. *Behavioral Sciences and the Law, 7,* 215-225.

Cutler, B. L., Dexter, H. R., & Penrod, S. D. (1990). Nonadversarial methods for sensitizing jurors to eyewitness evidence. *Journal of Applied Social Psychology, 20,* 1197-1207.

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12,* 41-55.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship?. *Law and Human Behavior, 4,* 243-260.

Dillon, M. K., Jones, A. M., Bergold, A. N., Hui, C. Y., & Penrod, S. D. (2017). Henderson instructions: Do they enhance evidence evaluation?. *Journal of Forensic Psychology Research and Practice, 17,* 1-24.

Garcia-Retamero, R. & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science, 22,* 392-399. doi: 10.1177/0963721412491570

Garcia-Retamero, R. & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors, 59,* 582-627. doi: 10.1177/0018720817690634

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutors go wrong*. Cambridge, MA: Harvard University Press.

Greene, E. (1988). Judge's instruction on eyewitness testimony: Evaluation and revision. *Journal of Applied Social Psychology, 18,* 252-276.

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-

time, and justifications. *Journal of Applied Research in Memory and Cognition, 8,* 233–243.

Gronlund, S. D. & Benjamin, A. S. (2018). The new science of eyewitness memory. *Psychology of Learning and Motivation, 69,* 241-284.

Hamm, R. M., Beasley, W. H., & Johnson, W. J. (2014). A balance beam aid for instruction in clinical diagnostic reasoning. *Medical Decision Making, 34,* 854-862.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology, 52,* pp. 139-183). North-Holland.

IACP National Law Enforcement Policy Center (2010). Eyewitness identification: Model policy.

Innocence Project (2019). Ronald Cotton. Retrieved from https://www.innocenceproject.org/cases/ronald-cotton/

Jones, E. E., Williams, K. D., & Brewer, N. (2008). "I had a confidence epiphany!": Obstacles to combating post-identification confidence inflation. *Law and Human Behavior, 32,* 164-176.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1304–1316. doi:10.1037/0278-7393.22.5.1304.

Kantner, J. & Dobbins, I. G. (2019). Partitioning the sources of recognition confidence: The role of individual differences. *Psychonomic Bulletin & Review,* 1-8.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The" general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist*, *44*(8), 1089.

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the" general acceptance" of eyewitness testimony research: A new survey of the experts. *American Psychologist*, *56*(5), 405.

Key, K. N., Neuschatz, J. S., Gronlund, S. D., Deloach, D., Wetmore, S. A., & McAdoo, R. M. (accepted pending minor revisions). High eyewitness confidence is always compelling: That's a problem.

Key, K. N., Wetmore, S. A., Cash, D. K., Neuschatz, J. S., & Gronlund, S. D. (2017). The effect of post-ID feedback on retrospective self-reports in showups. *Journal of Police and Criminological Psychology*, 369-377. doi: 10.1007/s11896-017-9228-y

Mickes, L., Clark, S. E. & Gronlund, S. D. (2017). Distilling the confidence-accuracy message: A comment on Wixted and Wells (2017). *Psychological Science in the Public Interest, 18*, 6-9. Doi: 10.1177/1529100617699240

National Registry of Exonerations (2018). Contributing factors and type of crime. Retrieved from http://www.law.umich.edu/special/exoneration/Pages/ExonerationsContribFactorsByCrime.aspx

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In M. Miller, & B. Bornstein (Eds.), Advances in psychology and law (edn, Vol. 1, pp. 43 – 69). Switzerland: Springer International Publishing.

New Jersey Eyewitness Instruction. (2012). New Jersey model criminal jury charges. Retrieved from: http://www.judiciary.state.nj.us/pressrel/2012/jury_instruction.pdf

New Jersey v. Henderson, 208 N. J. 208, 287 (2011).

Pawlenko N.B., Safer M.A., Wise R.A., & Holfeld B. (2013). A teaching aid for improving jurors' assessments of eyewitness accuracy. *Applied Cognitive Psychology*, *27*(2), 190–197. doi: 10.1002/acp.2895

Papailiou, A. P., Yokum, D. V., & Robertson, C. T. (2015). The novel New Jersey eyewitness instruction induces skepticism but not sensitivity. *PloS ONE*, *10*(12), e0142695.

Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review*, *13*, 519-557.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, *62*, 189-206.

Petty, R. E. & Cacioppo, J. T. (2012). *Communication and persuasion: Central and peripheral routes to attitude change.* Springer Science & Business Media.

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law*, *13*, 611-625.

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46-62.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127* (11), 966.

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied.* doi.org/10.1037/xap0000157

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

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

Steblay, N. K., Wells, G. L, & Douglass, A. B. (2014). The eyewitness post identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, & Law*, *20*, 1-18. doi: 10.1037/law0000001

Strawn, D. U., & Buchanan, R. W. (1975). Jury confusion: A threat to justice. *Judicature, 59*, 478-483.

Strawn, D. U., & Munsterman, G. T. (1981). Helping juries handle complex cases. *Judicature*, *65*, 444-447.

Thomas, J. P. & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, *16*, 87-113.

U.S. v Telfaire, 469 F.2d 552 (D. C. Cir. 1972)

Utah instructions (in prep). Rule 617: Eyewitness identification. Memorandum.

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*, 553–571. doi:10.1037/0003-066X.48.5.553

Wells, G. L., & Bradfield, A. L. (1998). “Good, you identified the suspect:” Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83,* 360 –376. doi: 10.1037/0021-9010.83.3.360

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger III, H. L. (2015a). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70,* 515-526.

Wixted, J. T., Mickes, L., Clark, S. E., Dunn, J. C., & Wells, W. (2015b). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, USA, 113,* 304–309. doi:10.1073/pnas.1516814112

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science, 13,* 324-335.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18,* 10-65.

Woller-Carter, M. (2015). Development of the intelligent graphs for everyday risky decisions tutor. Open Access Dissertation, Michigan Technological University.

Yates, S. Q. (2017). Eyewitness identification: Procedures for conducting photo arrays. U.S. Department of Justice Memorandum. Retrieved from https://www.justice.gov/archives/opa/press-release/file/923201/download

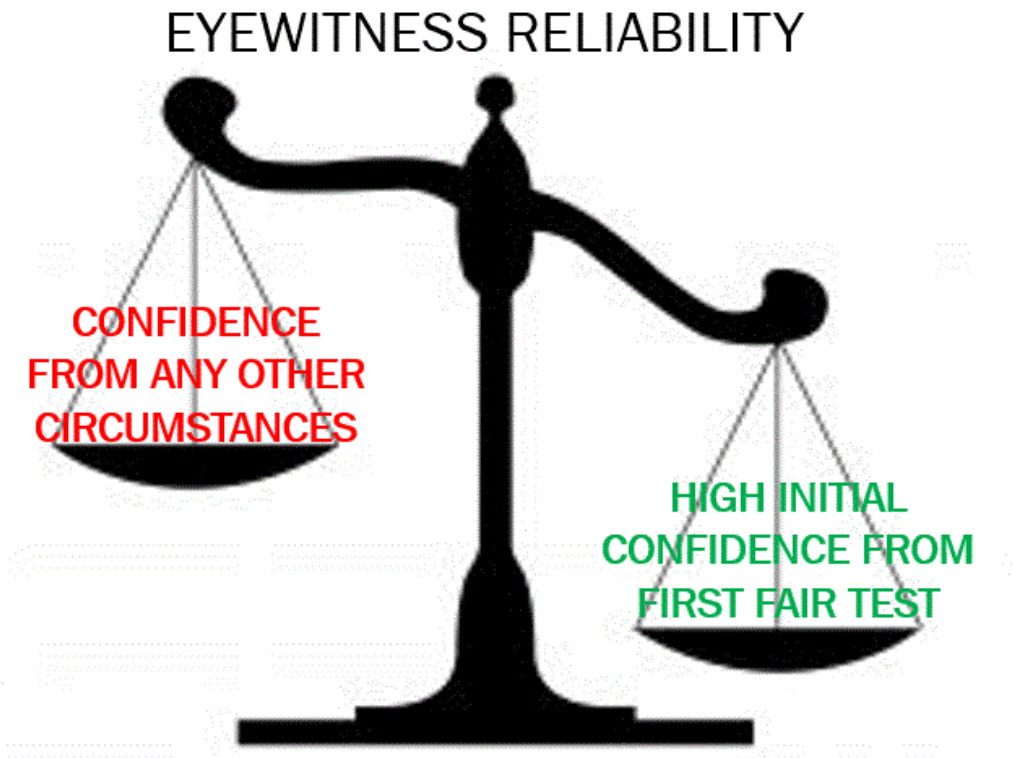
Ybarra, V. (2018). Self-evaluation of skills and overconfidence vulnerability: are most people blind to their own decision making biases?. Open Access Dissertation, University of Oklahoma. Retrieved from: https://shareok.org/handle/11244/299932

Appendix A

Key Intervention

Although it is commonly believed that there are many variables that can make eyewitnesses less accurate, like stress or poor viewing conditions, new research shows that these variables do not matter as much as eyewitness confidence. If an eyewitness is highly confident initially, they are very likely to be accurate. If an eyewitness is not confident initially, they are much less likely to be accurate. However, this is only true for the first identification attempt (e.g., the victim didn't see the perpetrator in the news before viewing him in a lineup), and only if the suspect does not stand out unfairly (the lineup must be fair). Later expressions of confidence, like in the courtroom, are not related to accuracy and should be downplayed.

Please look at the following image(s), which summarizes these ideas. The image shows how much weight, or importance, to give to the eyewitness confidence, depending on the confidence level itself and whether it came from a first, fair identification attempt. More weight, or importance, makes that side of the balance beam lower.



Appendix B

A pilot study was used to determine the appropriate visual aid for these experiments. The pilot study compared the performance of two visual aids, balance beam and icon arrays, both of which have been used successfully in other fields (Garcia-Retamero & Cokely, 2017; Hamm et al., 2014). These visual aids were compared to a control condition with no visual aid.

Method

Participants

Participants were *n* = 387 workers from Amazon's Mechanical Turk and students from introductory psychology courses at University of Oklahoma, who were compensated a small amount for participation (<$5.00 for Turk, class credit for students).

Design

The design was a 2 (Transcript Strength: Strong, Weak) x 3 (Visual Aid: Balance Beam, Icon Array, Control) between-participants factorial.

Materials & Procedure

The Materials and Procedure were the same as those in Experiment 1, Full Sample.

Results

Regression models with the control as the reference class were used to determine significant differences across conditions. There were no significant differences in verdict decisions, χ^2 (df = 7) = -270.26, p = .57, AIC = 554.52, r^2 = .04. However, the Balance Beam produced a much higher difference (.26) between Strong and Weak transcripts (evidence of sensitivity), compared to the Control condition (.06). See Table 1. Likelihood of Guilt was also trending higher in the Balance Beam (M = 59.52, SD = 28.48) than the Control (M = 55.6, SD = 27.37), t = 1.89, p = .059. No other differences across experimental manipulations occurred for any post-trial questionnaire item, usability, or workload metrics. Based on these results, indicating some sensitivity for the Balance Beam, it was chosen as the visual aid for these experiments.

Table B.1

Proportion of guilty verdicts across experimental manipulations

	Strong	Weak	Strong-Weak
Control	0.43	0.37	0.06
Balance Beam	0.56	0.30	0.26
Icon Array	0.58	0.39	0.20

Appendix C

Modified Henderson Instructions

You should consider the following factors that are related to the witness, the alleged perpetrator, and the criminal incident itself.

1. The Witness's Opportunity to View and Degree of Attention: In making this assessment you should consider the following:

a. Stress: You should consider a witness's level of stress and whether that stress, if any, distracted the witness or made it harder for him or her to identify the perpetrator.

b. Duration: A brief or fleeting contact is less likely to produce an accurate identification than a more prolonged exposure to the perpetrator.

c. Weapon Focus: The presence of a weapon can distract the witness and take the witness's attention away from the perpetrator's face.

d. Distance: The greater the distance between an eyewitness and a perpetrator, the higher the risk of a mistaken identification.

e. Lighting: Inadequate lighting can reduce the reliability of an identification.

f. Disguises/Changed Appearance: Disguises like hats, sunglasses, or masks can reduce the accuracy of an identification. Similarly, changes in appearance (like to facial features, hair, or body weight) can reduce accuracy.

2. Prior Description of Perpetrator: Another factor for your consideration is the accuracy of any description the witness gave after observing the incident and before identifying the perpetrator. Did the prior description match the photo or person picked out later, did the prior description provide details or was it just general in nature, and was the witness's testimony at trial consistent with, or different from, the prior description of the perpetrator.

3. Confidence and Accuracy: Eyewitness confidence is a very good predictor of accuracy. If an eyewitness is highly confident initially, they are very likely to be accurate. If an eyewitness is not confident initially, they are much less likely to be accurate. However, this is only true for the first identification attempt (e.g., the victim didn't see the perpetrator in the news before viewing him in a lineup), and only if the suspect does not stand out unfairly (the lineup must be fair). Later expressions of confidence, like in the courtroom, are not related to accuracy and should be downplayed.

4. Time Elapsed: Memories fade with time. In other words, the more time that passes, the greater the possibility that a witness's memory of a perpetrator will weaken. In evaluating the reliability of a witness's identification, you should also consider the circumstances under which any out-of-

court identification was made, and whether it was the result of a suggestive procedure. You should consider the following factors:

a. Line-up Composition: A suspect should not stand out from other members of the lineup.

b. Fillers: Lineups should include a number of possible choices for the witness, commonly referred to as “fillers.”

c. Multiple Viewings: When a witness views the same person in more than one identification procedure, it can be difficult to know whether a later identification comes from the witness’s memory of the original event or from an earlier exposure (e.g., saw in the new). You may consider whether the witness viewed the suspect multiple times during the identification process and, if so, whether that affected the reliability of the identification.

In determining the reliability of the identification, you should also consider whether the identification procedure was properly conducted.

1. Double-blind: A lineup administrator who knows which person or photo in the lineup is the suspect may intentionally or unintentionally convey that knowledge to the witness. That increases the chance that the witness will identify the suspect, even if the suspect is innocent. For that reason, whenever feasible, live lineups and photo arrays should be conducted by an officer who does not know the identity of the suspect.

2. Instructions: Identification procedures should begin with instructions to the witness that the perpetrator may or may not be in the lineup and that the witness should not feel compelled to make an identification.

Appendix D

Trial Transcript (adapted from Bradfield & McQuiston, 2004)

Experimental Manipulations in Red Font, with the Weak Evidence condition in [brackets]

People v. Roger Sanchez

Mr. Walker: Your Honor, the State would like to call Mrs. Edna Cameron to the witness stand.

Judge Cooper: Mrs. Cameron, if you could come up to the witness stand. If you'll remain standing and raise your right hand. Do you solemnly swear that you will tell the truth, the whole truth, and nothing but the truth?

Mrs. Cameron: I do.

Judge Cooper: Please state and spell your name for the Court and give your address.

Mrs. Cameron: Edna D. Cameron, C-A-M-E-R-O-N. 1337 Ashton Blvd. Chicago.

Judge Cooper: Thank you. Please be seated. The prosecution may now begin examining the witness.

DIRECT EXAMINATION OF MRS. CAMERON BY MR. WALKER:

Mr. Walker *(prosecutor)*: Mrs. Cameron, could you please explain, in your own words, what happened on the morning of August 17th of last year?

Mrs. Cameron: I was on my way to the bank when I passed the convenience store on Grand Avenue.

Mr. Walker: Around what time were you near the convenience store?

Mrs. Cameron: Well, the bank opens at 8 and I wanted to get there right when it opened so it was probably around 7:45.

Mr. Walker: Was the lighting pretty good at that time?

Mrs. Cameron: Yes.

Mr. Walker: Ok, what happened next?

Mrs. Cameron: Roger Sanchez came up from behind me and….

Mr. Reeder *(defense attorney)*: Objection, Your Honor. My client is innocent until proven guilty. I object to the witness naming him as the man who robbed her.

Judge Cooper: Objection sustained. Mrs. Cameron, please do not refer to the man who robbed you as Mr. Sanchez. Continue.

Mr. Walker: So, Mrs. Cameron, a man came up behind you?

Mrs. Cameron: Yes, he said "Give me your purse. I have a gun." Then I felt something poking into my back like a gun would. I turned around slowly and handed my purse to him.

Mr. Walker: I bet that was very stressful.

Mrs. Cameron: Yes, it was scary.

Mr. Walker: What did you do then?

Mrs. Cameron: I walked into the convenience store and asked the clerk to call the police.

Mr. Walker: Describe the man you saw when you turned around to give him your purse.

Mrs. Cameron: He was about 5'10", 200 pounds, with a beard, and he was wearing a red sweatshirt.

Mr. Walker: And you saw him pretty close up, right? He was close enough for you to see?

Mrs. Cameron: Yes, he was.

Mr. Walker: Did he have anything covering his face?

Mrs. Cameron: No, he didn't.

Mr. Walker: How long did you look at him?

Mrs. Cameron: A couple of seconds, I guess.

Mr. Walker: Do you think that was enough time to get a good look at him?

Mrs. Cameron: Yes, I do.

Mr. Walker: Mrs. Cameron, have you seen the man who robbed you since the morning of August 17th?

Mrs. Cameron: Yes.

Mr. Walker: When was that?

Mrs. Cameron: The police asked me to look at some pictures and try to pick out the person who robbed me.

Mr. Walker: And what happened?

Mrs. Cameron: I picked out the picture of the person who robbed me.

Mr. Walker: Is that person in this courtroom today?

Mrs. Cameron: Yes, he is.

Mr. Walker: Could you point him out, please?

Mrs. Cameron: He's sitting right over there.

Mr. Walker: For the record, Mrs. Cameron pointed at the defendant, Roger Sanchez. Mrs. Cameron, how confident are you here today that Roger Sanchez is the man who robbed you?

Mrs. Cameron: I'm 90% confident.

Mr. Walker: Thank you, Mrs. Cameron. I have no further questions, Your Honor.

Judge Cooper: Mr. Reeder, your witness.

CROSS-EXAMINATION OF MRS. CAMERON BY MR. REEDER:

Mr. Reeder *(defense attorney)*: Mrs. Cameron, am I correct when I say that you identified my client from a police lineup some time after the crime?

Mrs. Cameron: Yes, I did.

Mr. Reeder: And how long after the crime occurred did you make that identification?

Mrs. Cameron: It was a few days later.

Mr. Reeder: Did you indicate how confident you were at the time you initially made your identification?

Mrs. Cameron: Yes, I did.

Mr. Reeder: Judge Cooper, I would like to enter into evidence the statement that Mrs. Cameron made on the day of her identification.

Judge Cooper: Accepted as Exhibit A.

Mr. Reeder: Mrs. Cameron, is this the statement you completed on the date of the identification?

Mrs. Cameron: Yes, it is.

Mr. Reeder: Would you read what you wrote on the day of your identification for the court, please?

Mrs. Cameron: I wrote "I picked out photograph #3. I am 90% [20%, Experiment 1; 50%, Experiment 2] confident that this is the guy who robbed me."

Mr. Reeder: And you wrote this a few days after the robbery occurred, correct?

Mrs. Cameron: Yes, I did.

Mr. Reeder: Ok. You testified that the man who robbed you came up from behind you, correct?

Mrs. Cameron: Yes.

Mr. Reeder: So you didn't see him approach you?

Mrs. Cameron: No, I did not.

Mr. Reeder: So, overall, you didn't look at the man for very long, did you?

Mrs. Cameron: Well…not very long, but he was right in front of me for a couple seconds.

Mr. Reeder: And yet you are testifying that my client was the man who robbed you?

Mrs. Cameron: Yes, I am.

Mr. Reeder: I have no further questions, Your Honor.

Experiment 2 Only

Mr. Walker: Your Honor, the State would like to call Detective John Hale to the witness stand.

Judge Cooper: Detective Hale, if you could come up to the witness stand. If you'll remain standing and raise your right hand. Do you solemnly swear that you will tell the truth, the whole truth, and nothing but the truth?

Detective Hale: I do.

Judge Cooper: Please state and spell your name for the Court and give your address.

Mrs. Cameron John Hale. **:** H A L E 2039 West Oak Avenue. Chicago.

Judge Cooper: Thank you. Please be seated. The prosecution may now begin examining the witness.

DIRECT EXAMINATION OF DETECTIVE HALE BY MR. WALKER:

Mr. Walker: Sir, are you a detective with the Chicago Metropolitan Police Department?

Detective Hale: Yes sir.

Mr. Walker: And how long have you been a police officer with this department?

Detective Hale: 21 years.

Mr. Walker: Have you become the lead investigator for the robbery of Mrs. Edna Cameron?

Detective Hale: Yes.

Mr. Walker: Okay. During the course of the investigation, did you have occasion to show the victim a lineup of individuals who may have committed the crime?

Detective Hale: Yes, I made the lineup.

Mr. Walker: Can you describe how you made the lineup?

Detective Hale: Sure. We have a large [small] database of mugshot photos that we use. As a result, we find a very good [very limited] set of photos that [sort of] match the suspect.

Mr. Walker: Okay. Did she identify anyone from the lineup?

Detective Hale: Yes, she chose our suspect, Roger Sanchez.

Mr. Walker: And based on her identification from the lineup, you detained him and pressed charges?

Detective Hale: That's correct.

Mr. Walker: I have no further questions, Your Honor.

Judge Cooper: Mr. Reeder, your witness.

CROSS-EXAMINATION OF DETECTIVE HALE BY MR. REEDER:

Mr. Reeder *(defense attorney)*: Detective Hale, I want to talk more about the lineup. Do you think that the suspect stood out from other people in the lineup?

Detective Hale: No, I don't think he stands out. [Maybe. But we have a pretty small database of mugshot photos, so I wasn't able to find many people who look like him.]

Mr. Reeder: When we look at lineup fairness, we normally think about things like gender, race, skin tone, hair color and style, and eye color. How many of the people in the lineup matched these features of the suspect?

Detective Hale: I'd say all 5 of the other people in the lineup match the suspect's features pretty well. [Well, only one or two. Again, it was hard to find other people who look like him].

Mr. Reeder: Now I want to ask you about the exchange you had with the witness during the lineup. What did you ask the witness to do with the lineup?

Detective Hale: I told her to choose who, if anyone, she thought was the perpetrator of her crime. She could have chosen nobody. [I told her that we had the suspect and she should "pick him out".]

Mr. Reeder: Is it standard in your police department to say this?

Detective Hale: Yes, this is what we typically do.

Mr. Reeder: One last thing. Did you stay in the room with the witness while she was looking at the lineup, or did you have another police officer administer it?

Detective Hale: Another police officer administered it. This is our standard procedure. [I did it myself. I like to administer the lineups I make.]

Mr. Reeder: And considering all this, you believe the witness correctly identified the perpetrator of her crime?

Detective Hale: Yes, I do.

Mr. Reeder: Your Honor, I have no further questions.

Appendix E

Pre-Deliberation Instructions

For you to find this defendant guilty, the State must prove beyond a reasonable doubt that this defendant is the person who committed the crime. You must determine, therefore, not only whether the State has proven the offenses charged, but also whether the State has proven that this defendant is the person who committed it. The State has presented the testimony of an eyewitness. You will recall that this witness identified the defendant in court as the person who committed the crime of armed robbery. The State also presented testimony that on a prior occasion before this trial, this witness identified the defendant as the person who committed these offenses.

In evaluating this identification, you should consider the observations on which the identification was based, the witness's ability to make those observations, and the circumstances under which the identification was made.

Although you may wish for more evidence, the eyewitness is the primary source of evidence in this case. You must make a verdict decision about the defendant based on the information you have read.

Appendix F

Post-Trial Questionnaire

Abbreviated Variable Name	Description	Response Options
Verdict	As a member of the jury, please provide your verdict decision below	1=guilty; 2=innocent
Confidence In Verdict	How confident are you in your verdict decision?	1=not at all confident; 10=very confident
Likelihood of Guilt	How likely do you think it is that Roger A. Sanchez, the defendant, actually committed the crime? (Please type a percent from 0-100 in the line provided)	free response
Witness Accuracy	Do you think Mrs. Cameron, the eyewitness, made an accurate identification of the perpetrator?	1=definitely not; 10= definitely yes
Case Strength	Do you think the prosecution or defense's case is stronger?	1=defense's case is very strong; 5=both are equally strong; 10=prosecution's case is very strong
Sentence Length Recommendation	If Mr. Sanchez is convicted, what do you think his sentence should be?	1=lightest possible sentence; 10=heaviest possible sentence
Confidence In Witness	How confident are you that Mrs. Cameron identified the correct man in the police line-up?	1=not at all confident; 10=very confident
Initial Confidence Influenced Verdict	How much was your verdict influenced by the confidence rating Mrs. Cameron gave at the time of her identification?	1=not at all influenced; 10=strongly influenced
Courtroom Confidence Influenced Verdict	How much was your verdict influenced by the confidence rating Mrs. Cameron gave at the trial?	1=not at all influenced; 10=strongly influenced

Initial Confidence Percentage	At the time of her identification decision from the police line-up, how confident was Mrs. Cameron that she identified the correct person?	10%-100% in 10% increments
Courtroom Confidence Percentage	At the time of the trial, how confident was Mrs. Cameron that she identified the correct person?	10%-100% in 10% increments
Confidence Ranking	Compared to how confident she was at the trial, do you remember whether Mrs. Cameron was more confident, less confident, or equally as confident when she first made her identification from the line-up?	1=initial<courtroom; 2=initial equal to courtroom; 3=initial > courtroom
Good Look	Do you think Mrs. Cameron got a good look at the culprit, based on the information you received about the robbery?	1=not a good look; 10=very good look
Attention	How much attention do you think Mrs. Cameron was paying to the culprit's face when the crime occurred?	1=no attention; 10=total attention
Good Basis	To what extent do you feel that Mrs. Cameron had a good basis (enough information) to make a good identification?	1=no basis at all; 10=a very good basis
Witness Memory	How good of a memory for strangers' faces do you believe Mrs. Cameron has?	1=very poor; 10=very good
Eyewitness Accuracy General	How often would you estimate that eyewitness identifications are correct, in general?	1=almost never; 10=almost always
Confidence Predicts Accuracy	How good of an indicator of eyewitness accuracy do you think eyewitness confidence is?	1=not at all; 10=completely

Confidence Inflation Occurs	Indicate the extent to which you agree with this statement: An eyewitness can be unsure about an identification made at the police line-up, but can become more confident in that identification over time.	1=totally disagree; 10=totally agree
Confidence Inflation Equals Accuracy	Indicate the extent to which you agree with this statement: An eyewitness who is unsure about their identification at the police line-up but then becomes more confident over time, can still be an accurate witness.	1=totally disagree; 10=totally agree

Appendix G

Comprehension Check Items, correct answers in bold.

Q1 What type of crime was the defendant on trial for?

- **Robbery (1)**
- Murder (2)
- Sexual Assault (3)

Q2 The defendant was:

- John Walker (1)
- **Roger Sanchez (2)**
- Edna Cameron (3)

Q3 Did the witness of the crime identify a suspect from the police lineup?

- **Yes (1)**
- No (2)

Q4 What was the victim threatened with?

- A knife (1)
- A baseball bat (2)
- **A gun (3)**

Appendix H

USABILITY MEASURES

Please rank how you felt while you were weighing all the information presented to you about the case of *People v. Sanchez*. 0=strongly disagree, 10=strongly agree

I found the information unnecessarily complex

I thought the information was easy to understand

I think I would need help from an expert to be able to understand the information

I thought there was too much inconsistency in the information

I would imagine that most people would be able to understand this information very quickly

I felt very confident using this information to make a verdict decision

WORKLOAD MEASURES

Please rank how you felt while you were weighing all the information presented to you about the case of *People v. Sanchez*.

How mentally demanding was the task? 0=very low, 10=very high

How physically demanding was the task? 0=very low, 10=very high

How hurried or rushed was the pace of the task? 0=very low, 10=very high

How successful were you in accomplishing what you were asked to do? 0=perfect, 10=failure

How hard did you have to work to accomplish your level of performance? 0=very low, 10=very high

How insecure, discouraged, irritated, stressed, and annoyed were you? 0=very low, 10=very high

Appendix I

Regression models and graphs for Full Sample results, Experiment 1, *n* = 687. Remember, not all items were completed by full sample of participants; Appendix J presents Subset of Sample results (including graphs).

Table I.1.

VERDICT

	B	*SE*	*z*	*p*
(Intercept)	1.81	0.51	3.57	0.00
Transcript Strength (Weak)	-1.64	0.30	-5.49	0.00
Key Intervention v. Control	-0.36	0.30	-1.19	0.94
Key Intervention v. modified Henderson	-0.47	0.30	-1.57	0.69
Gender	-0.14	0.17	-0.84	1.00
Age	-0.01	0.01	-1.58	0.69
Ethnicity	0.21	0.07	3.01	0.02
Education	0.24	0.09	2.55	0.09
Political Orientation	-0.25	0.06	-3.89	0.00
Income	-0.20	0.09	-2.39	0.12
Transcript Strength (Weak)*Key Intervention v. Control	0.36	0.42	0.87	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.38	0.41	0.92	1.00

Note: Overall model fit, χ^2 (df = 12) = -410.04, $p < .0001$, $r^2 = .12$, AIC = 844.09

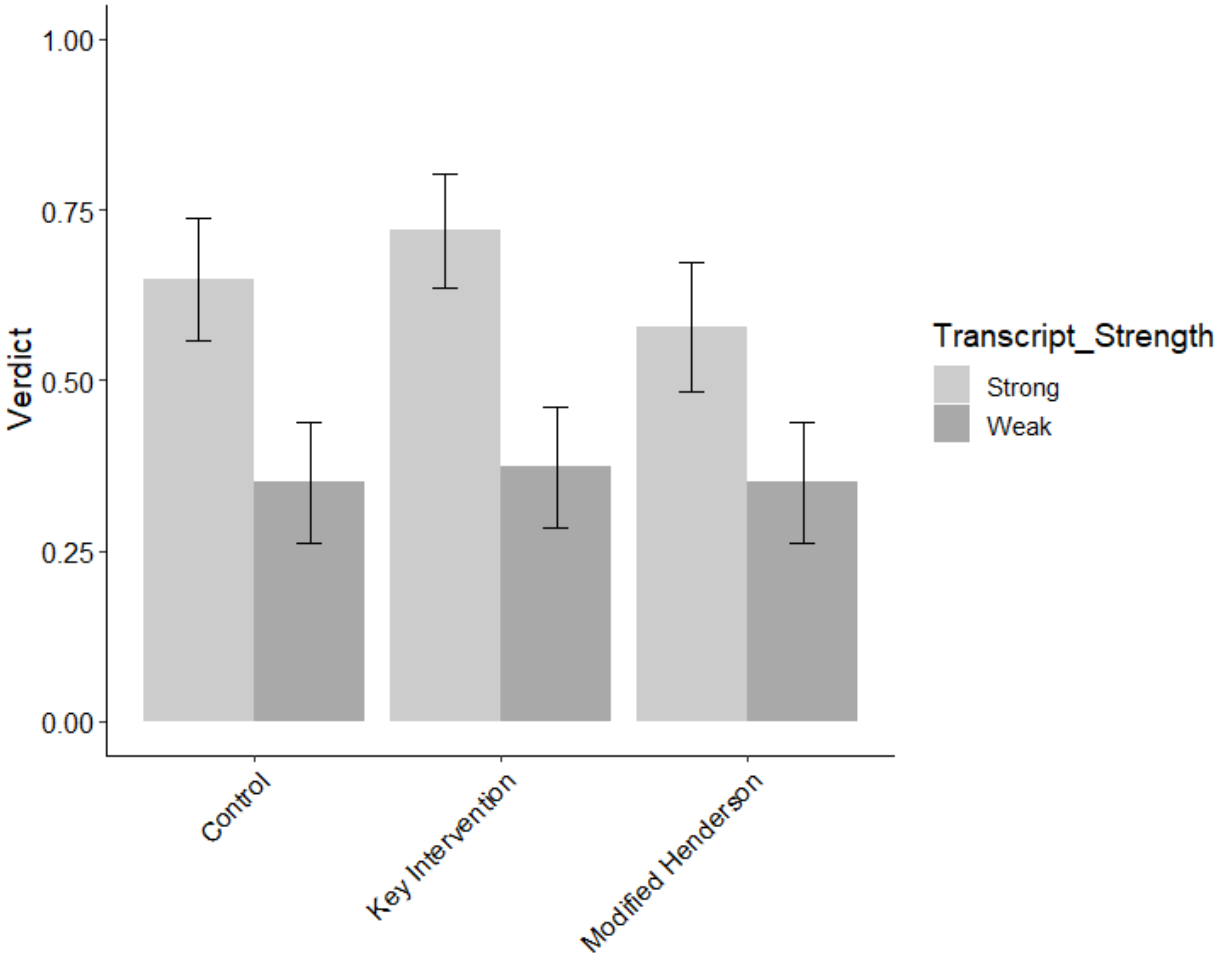


Figure I.1. Average VERDICT as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.2.
CONFIDENCE IN VERDICT

	B	*SE*	*t*	*p*
(Intercept)	7.41	0.42	17.83	0.00
Transcript Strength (Weak)	-0.30	0.24	-1.23	1.00
Key Intervention v. Control	0.10	0.25	0.42	1.00
Key Intervention v. modified Henderson	-0.10	0.25	-0.38	1.00
Gender	-0.06	0.14	-0.43	1.00
Age	0.00	0.01	0.24	1.00
Ethnicity	0.08	0.06	1.43	1.00
Education	0.21	0.08	2.66	0.09
Political Orientation	-0.11	0.05	-2.15	0.32
Income	0.00	0.07	-0.04	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.33	0.35	-0.93	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.09	0.35	0.27	1.00

Note: Overall model fit, F (11, 654) = 2.23, p = .01, adjusted r2 = .02

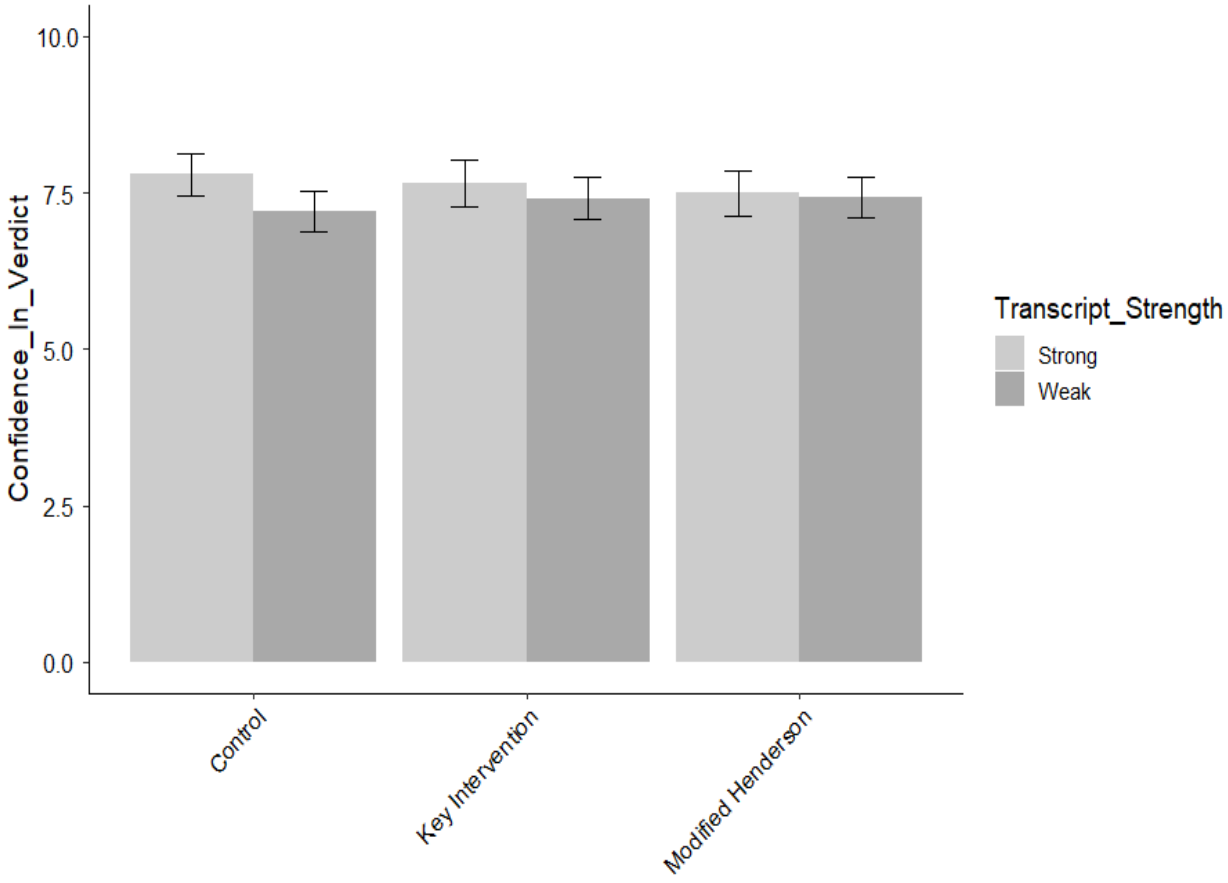


Figure I.2. Average CONFIDENCE IN VERDICT as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.3.
LIKELIHOOD OF GUILT

	B	*SE*	*t*	*p*
(Intercept)	81.55	5.80	14.05	0.00
Transcript Strength (Weak)	-21.48	3.41	-6.30	0.00
Key Intervention v. Control	-1.10	3.48	-0.32	1.00
Key Intervention v. modified Henderson	-2.21	3.50	-0.63	1.00
Gender	-2.16	1.97	-1.10	1.00
Age	0.05	0.09	0.49	1.00
Ethnicity	2.10	0.79	2.66	0.07
Education	0.71	1.10	0.65	1.00
Political Orientation	-3.71	0.74	-5.03	0.00
Income	-0.38	1.00	-0.38	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.28	4.87	0.26	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.61	4.88	0.33	1.00

Note: Overall model fit, F (11, 656) = 12.42, $p < .0001$, adjusted $r^2 = .16$

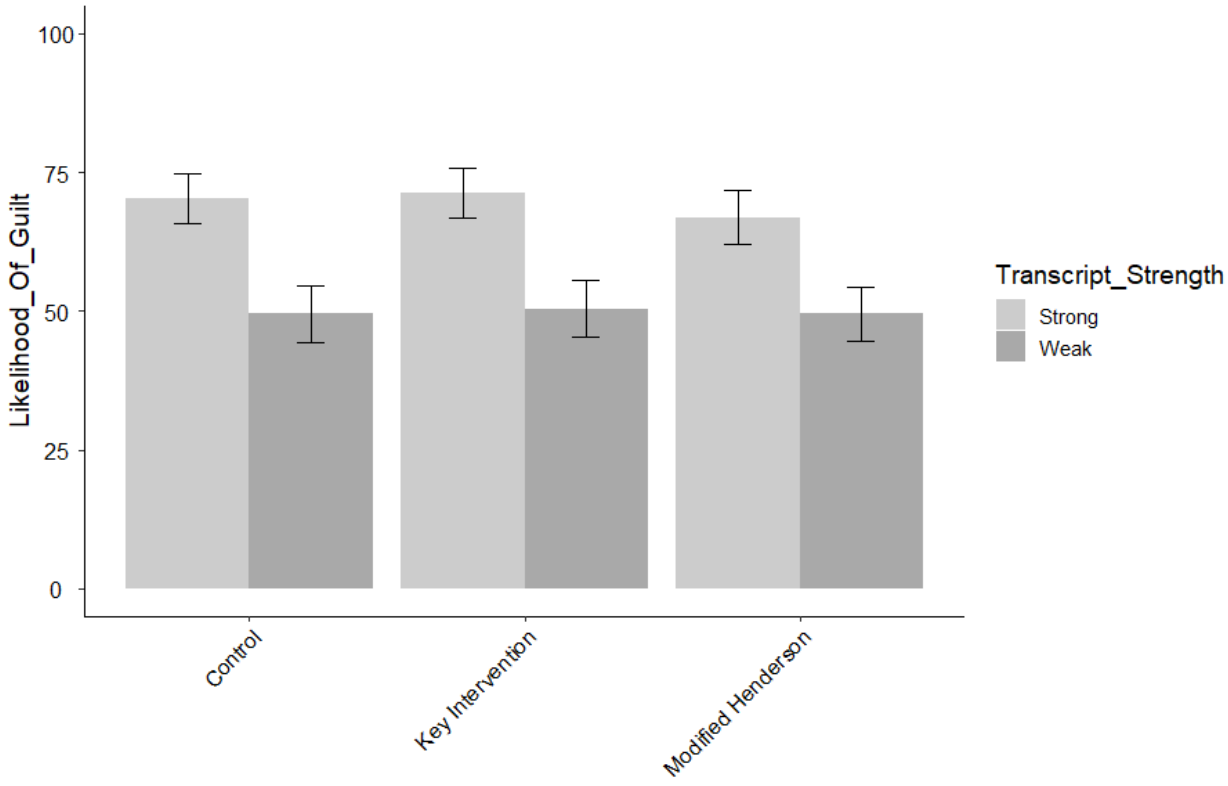


Figure I.3. Average LIKELIHOOD OF GUILT as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.4.
CONFIDENCE IN WITNESS

	B	*SE*	*t*	*p*
(Intercept)	9.83	0.63	15.49	0.00
Transcript Strength (Weak)	-2.96	0.37	-7.93	0.00
Key Intervention v. Control	-0.43	0.38	-1.14	1.00
Key Intervention v. modified Henderson	-0.63	0.38	-1.64	0.61
Gender	-0.42	0.22	-1.94	0.42
Age	0.00	0.01	-0.34	1.00
Ethnicity	0.28	0.09	3.21	0.01
Education	0.21	0.12	1.75	0.57
Political Orientation	-0.51	0.08	-6.31	0.00
Income	-0.07	0.11	-0.63	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.06	0.53	0.11	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.48	0.54	0.90	1.00

Note: Overall model fit, F (11, 656) = 20.04, $p < .0001$, adjusted $r^2 = .24$

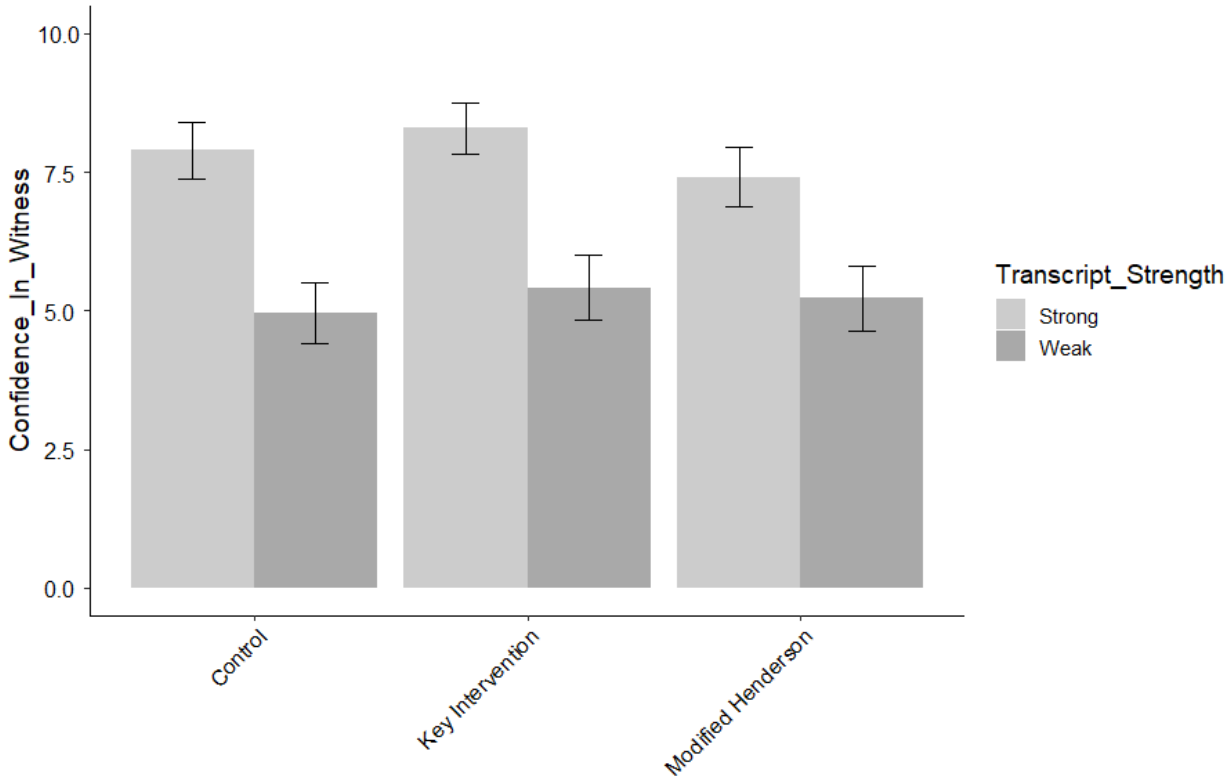


Figure I.4. Average CONFIDENCE IN WITNESS as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.5.
INITIAL CONFIDENCE INFLUENCED VERDICT

	B	*SE*	*t*	*p*
(Intercept)	6.20	0.58	10.68	0.00
Transcript Strength (Weak)	0.17	0.34	0.50	1.00
Key Intervention v. Control	0.07	0.35	0.21	1.00
Key Intervention v. modified Henderson	-0.61	0.35	-1.72	0.94
Gender	0.14	0.20	0.70	1.00
Age	0.01	0.01	0.77	1.00
Ethnicity	0.03	0.08	0.41	1.00
Education	0.15	0.11	1.34	1.00
Political Orientation	0.06	0.07	0.81	1.00
Income	0.10	0.10	0.99	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.42	0.49	-0.86	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.69	0.50	1.39	1.00

Note: Overall model fit, F (11, 597) = 1.27, p = .24, adjusted r^2 = .005

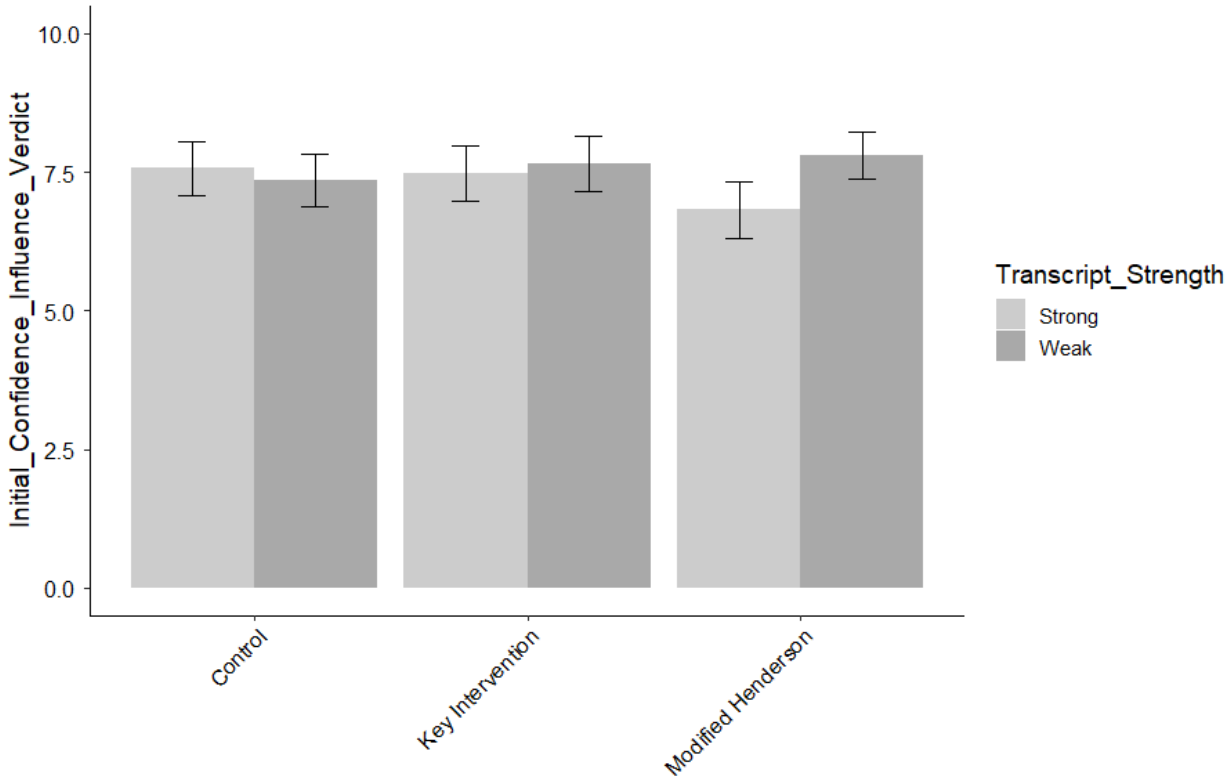


Figure I.5. Average INITIAL CONFIDENCE INFLUENCED VERDICT as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.6.

COURTROOM CONFIDENCE INFLUENCED VERDICT

	B	*SE*	*t*	*p*
(Intercept)	8.16	0.65	12.62	0.00
Transcript Strength (Weak)	-1.77	0.38	-4.63	0.00
Key Intervention v. Control	0.27	0.39	0.70	1.00
Key Intervention v. modified Henderson	-0.19	0.39	-0.49	1.00
Gender	-0.25	0.22	-1.14	1.00
Age	-0.01	0.01	-1.10	1.00
Ethnicity	0.27	0.09	2.98	0.03
Education	0.10	0.12	0.85	1.00
Political Orientation	-0.26	0.08	-3.15	0.02
Income	-0.16	0.11	-1.42	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.03	0.54	-0.05	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.90	0.55	1.64	0.81

Note: Overall model fit, F (11, 595) = 6.58, $p < .0001$, adjusted $r^2 = .09$

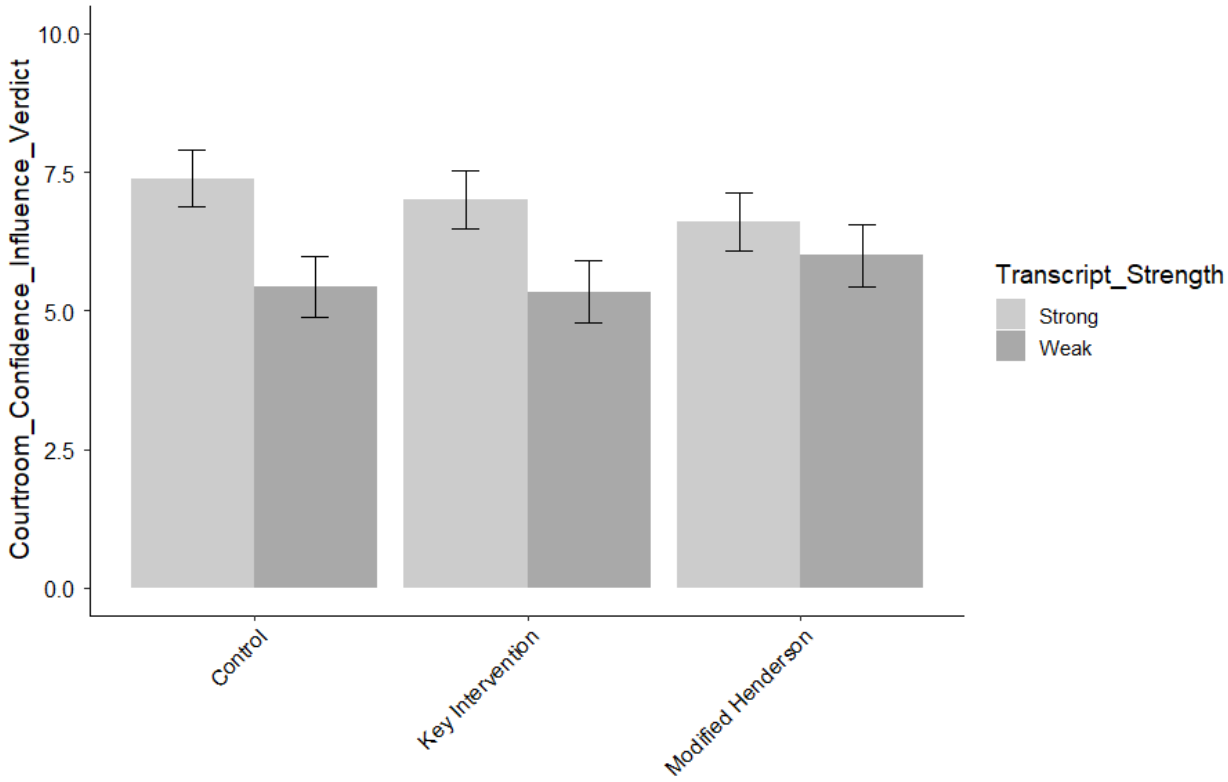


Figure I.6. Average COURTROOM CONFIDENCE INFLUENCED VERDICT as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.7.
CONFIDENCE INFLATION EQUALS ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	5.82	0.58	10.05	0.00
Transcript Strength (Weak)	-0.20	0.34	-0.57	1.00
Key Intervention v. Control	0.32	0.35	0.93	1.00
Key Intervention v. modified Henderson	-0.23	0.35	-0.65	1.00
Gender	-0.26	0.20	-1.32	1.00
Age	-0.02	0.01	-1.85	0.59
Ethnicity	0.14	0.08	1.76	0.63
Education	0.31	0.11	2.87	0.04
Political Orientation	-0.26	0.07	-3.56	0.00
Income	0.02	0.10	0.21	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.32	0.49	-0.65	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.19	0.49	0.38	1.00

Note: Overall model fit, F (11, 657) = 3.57, $p < .0001$, adjusted $r^2 = .04$

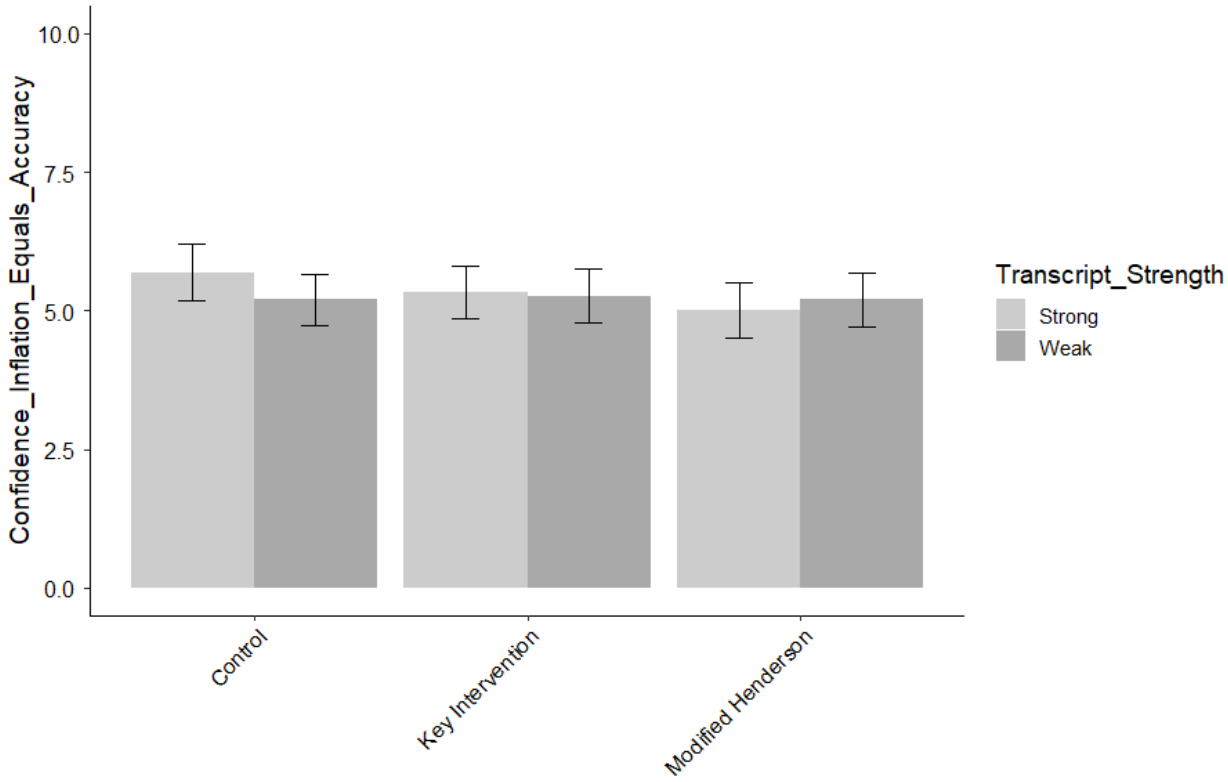


Figure I.7. Average CONFIDENCE INFLATION EQUALS ACCURACY as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.8.
COMPREHENSION CHECK QUESTIONS

	B	*SE*	*t*	*p*
(Intercept)	3.37	0.14	23.42	0.00
Transcript Strength (Weak)	0.12	0.08	1.46	0.73
Key Intervention v. Control	0.01	0.09	0.12	1.00
Key Intervention v. modified Henderson	0.13	0.09	1.44	0.73
Gender	0.11	0.05	2.31	0.17
Age	0.01	0.00	3.87	0.00
Ethnicity	-0.01	0.02	-0.36	1.00
Education	-0.14	0.03	-5.31	0.00
Political Orientation	0.06	0.02	3.05	0.02
Income	0.05	0.02	2.20	0.20
Transcript Strength (Weak)*Key Intervention v. Control	-0.08	0.12	-0.70	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.19	0.12	-1.58	0.69

Note: Overall model fit, F (11, 656) = 6.38, $p < .0001$, adjusted $r^2 = .08$

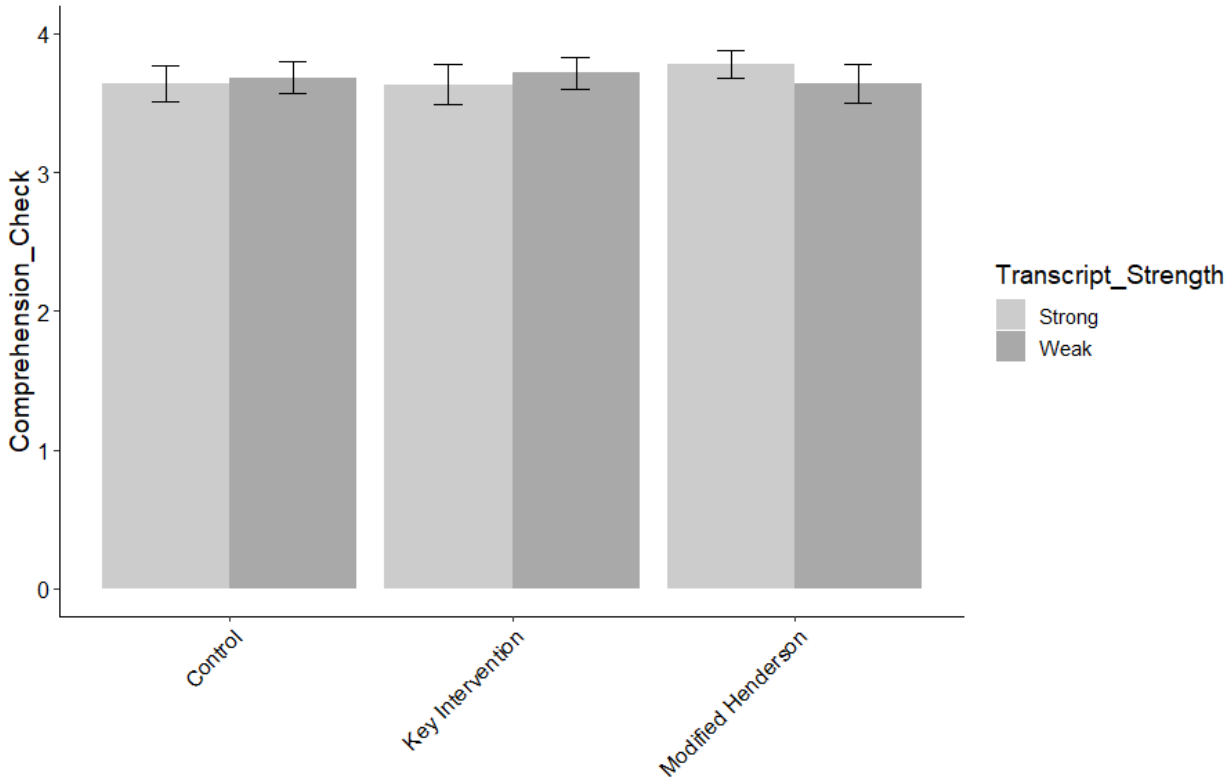


Figure I.8. Average COMPREHENSION CHECK QUESTIONS as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.9.
INITIAL CONFIDENCE PERCENTAGE

	B	*SE*	*t*	*p*
(Intercept)	9.53	0.44	21.88	0.00
Transcript Strength (Weak)	-5.45	0.26	-21.13	0.00
Key Intervention v. Control	0.00	0.26	0.01	1.00
Key Intervention v. modified Henderson	0.16	0.26	0.59	1.00
Gender	-0.24	0.15	-1.61	0.65
Age	-0.01	0.01	-1.87	0.44
Ethnicity	0.08	0.06	1.31	0.95
Education	0.18	0.08	2.12	0.31
Political Orientation	-0.14	0.06	-2.47	0.14
Income	-0.15	0.08	-1.92	0.44
Transcript Strength (Weak)*Key Intervention v. Control	-0.08	0.37	-0.22	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.08	0.37	0.21	1.00

Note: Overall model fit, F (11, 597) = 120.3, $p < .0001$, adjusted $r^2 = .68$

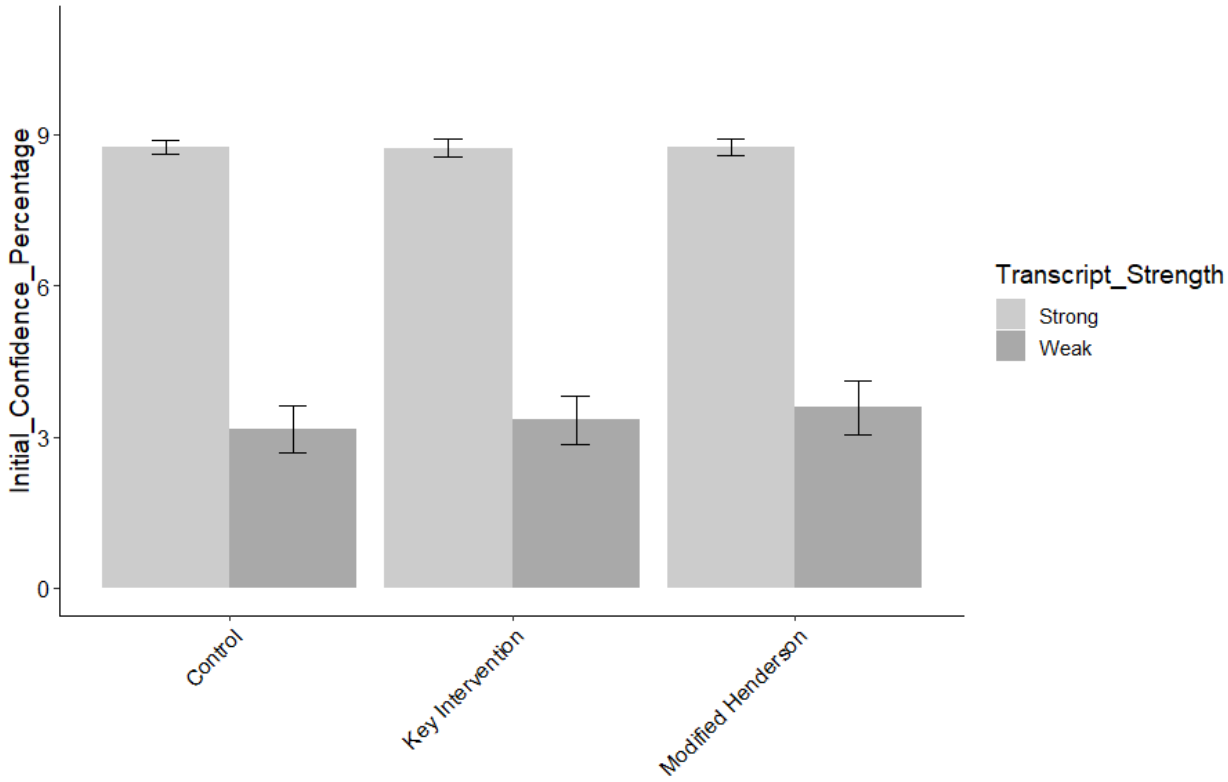


Figure I.9. Average INITIAL CONFIDENCE PERCENTAGE as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.10.
COURTROOM CONFIDENCE PERCENTAGE

	B	*SE*	*t*	*p*
(Intercept)	8.76	0.30	29.20	0.00
Transcript Strength (Weak)	-0.26	0.18	-1.44	1.00
Key Intervention v. Control	0.04	0.18	0.23	1.00
Key Intervention v. modified Henderson	0.04	0.18	0.21	1.00
Gender	0.09	0.10	0.84	1.00
Age	0.01	0.00	1.25	1.00
Ethnicity	-0.11	0.04	-2.64	0.09
Education	-0.14	0.06	-2.42	0.16
Political Orientation	0.04	0.04	0.97	1.00
Income	0.06	0.05	1.24	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.39	0.25	-1.55	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.03	0.25	0.10	1.00

Note: Overall model fit, F (11, 596) = 3.97, $p < .0001$, adjusted $r^2 = .05$

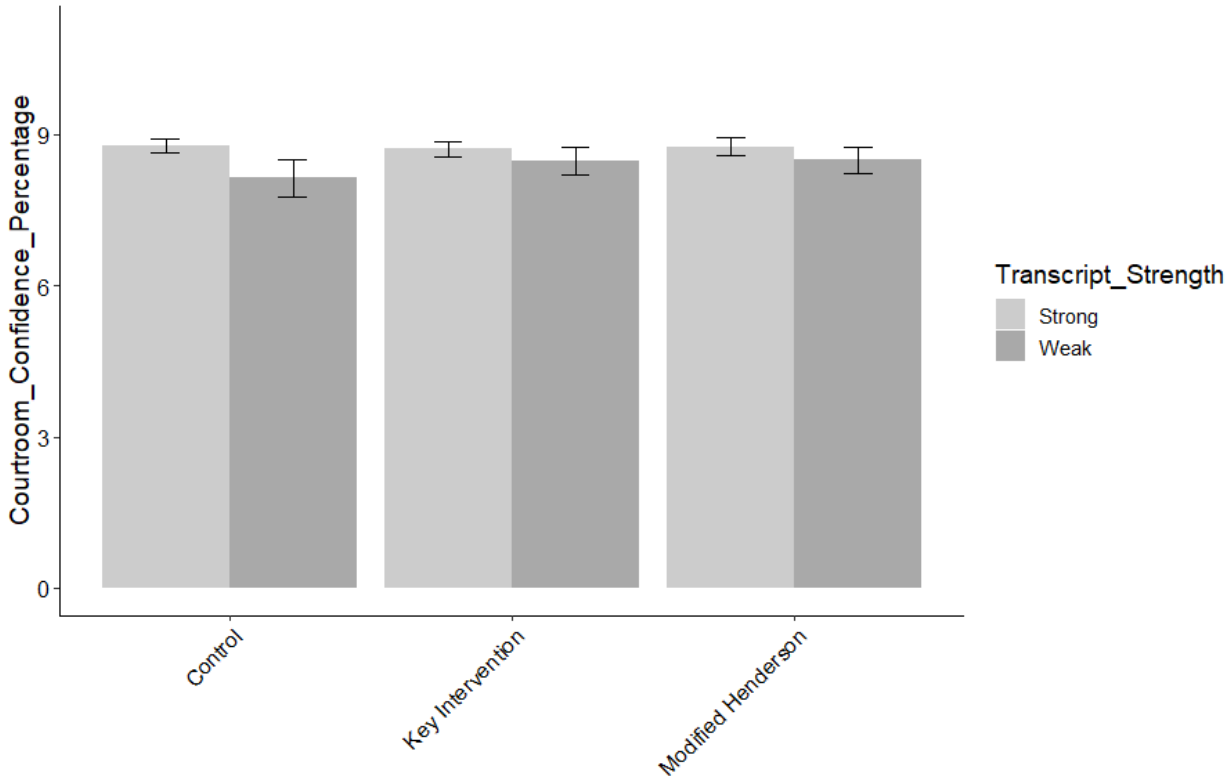


Figure I.10. Average COURTROOM CONFIDENCE PERCENTAGE as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.11.
USABILITY

	B	*SE*	*t*	*p*
(Intercept)	7.44	0.40	18.40	0.00
Transcript Strength (Weak)	0.31	0.24	1.31	0.38
Key Intervention v. Control	0.49	0.24	2.03	0.24
Key Intervention v. modified Henderson	0.28	0.24	1.15	0.38
Gender	0.28	0.14	2.07	0.24
Age	0.04	0.01	6.51	0.00
Ethnicity	-0.10	0.05	-1.82	0.28
Education	-0.45	0.08	-5.92	0.00
Political Orientation	0.22	0.05	4.26	0.00
Income	0.19	0.07	2.69	0.06
Transcript Strength (Weak)*Key Intervention v. Control	-0.55	0.34	-1.61	0.33
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.83	0.34	-2.42	0.11

Note: Overall model fit, F (11, 657) = 12.29, $p < .0001$, adjusted $r^2 = .16$

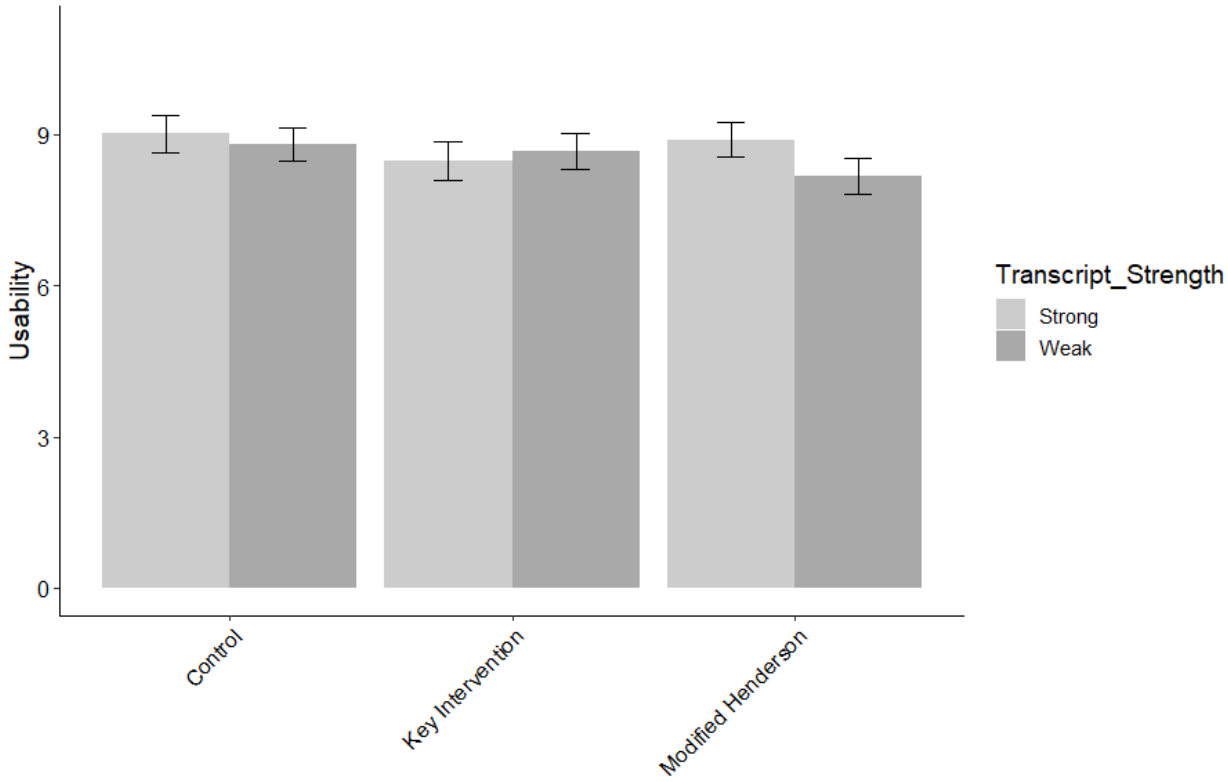


Figure I.11. Average USABILITY as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table I.12.
WORKLOAD

	B	*SE*	*t*	*p*
(Intercept)	5.72	0.41	14.11	0.00
Transcript Strength (Weak)	-0.10	0.24	-0.43	1.00
Key Intervention v. Control	-0.16	0.24	-0.68	1.00
Key Intervention v. modified Henderson	-0.13	0.24	-0.52	1.00
Gender	-0.15	0.14	-1.12	1.00
Age	-0.03	0.01	-4.61	0.00
Ethnicity	0.10	0.06	1.74	0.66
Education	0.53	0.08	6.93	0.00
Political Orientation	-0.27	0.05	-5.35	0.00
Income	-0.09	0.07	-1.24	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.23	0.34	-0.68	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.51	0.34	1.49	0.95

Note: Overall model fit, F (11, 657) = 11.97, $p < .0001$, adjusted $r^2 = .15$

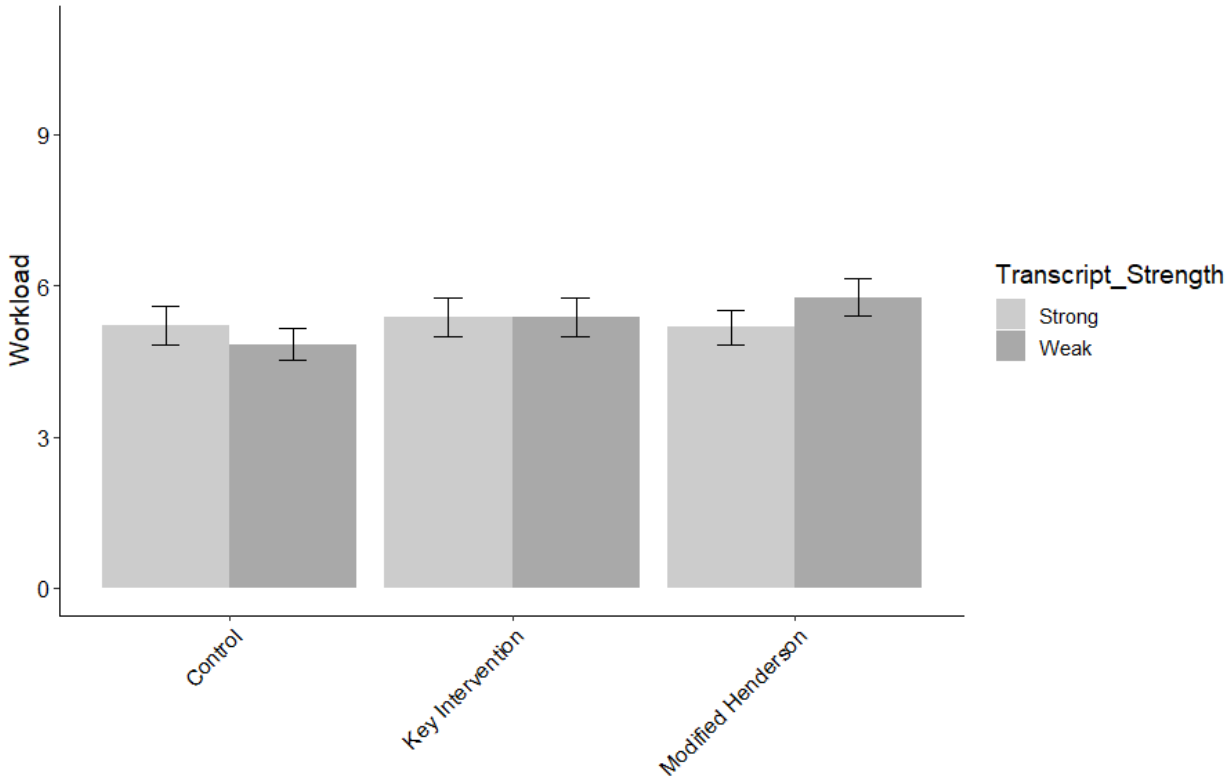


Figure I.12. Average WORKLOAD as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Appendix J

Regression models and graphs for Subset of Sample results, Experiment 1, *n* = 200. Graphs that appeared in Appendix I (Full Sample) are not repeated here.

Table J.1.
VERDICT

	B	*SE*	*z*	*p*
(Intercept)	2.22	1.05	2.12	0.37
Transcript Strength (Weak)	-1.55	0.60	-2.57	0.13
Key Intervention v. Control	0.18	0.59	0.30	1.00
Key Intervention v. modified Henderson	-0.69	0.58	-1.19	1.00
Gender	0.36	0.34	1.08	1.00
Age	-0.01	0.02	-0.70	1.00
Ethnicity	0.03	0.13	0.20	1.00
Education	0.31	0.19	1.64	1.00
Political Orientation	0.05	0.12	0.45	1.00
Income	-0.48	0.18	-2.66	0.11
Numeracy (BNT-S)	0.01	0.12	0.04	1.00
Graph Literacy	-0.16	0.07	-2.39	0.20
Transcript Strength (Weak)*Key Intervention v. Control	-0.47	0.90	-0.52	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.66	0.83	0.80	1.00

Note: Overall model fit, *x2* (*df* = 14) = -105.02, $p < .0001$, $r^2 = .17$, AIC = 238.04

Table J.2.
CONFIDENCE IN VERDICT

	B	*SE*	*t*	*p*
(Intercept)	6.64	0.92	7.23	0.00
Transcript Strength (Weak)	0.50	0.53	0.95	1.00
Key Intervention v. Control	1.10	0.55	2.02	0.54
Key Intervention v. modified Henderson	0.92	0.54	1.71	0.97
Gender	0.14	0.29	0.49	1.00
Age	0.01	0.01	0.64	1.00
Ethnicity	0.25	0.12	2.18	0.40
Education	0.20	0.16	1.24	1.00
Political Orientation	-0.17	0.11	-1.63	1.00
Income	0.00	0.15	0.00	1.00
Numeracy (BNT-S)	-0.10	0.10	-1.02	1.00
Graph Literacy	-0.02	0.06	-0.39	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-1.19	0.78	-1.53	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-1.01	0.73	-1.37	1.00

Note: Overall model fit, F (13, 173) = 1.44, p = .14, adjusted r^2 = .03

Table J.3.
LIKELIHOOD OF GUILT

	B	*SE*	*t*	*p*
(Intercept)	91.27	11.30	8.08	0.00
Transcript Strength (Weak)	-17.61	6.49	-2.71	0.10
Key Intervention v. Control	2.47	6.71	0.37	1.00
Key Intervention v. modified Henderson	-3.23	6.65	-0.49	1.00
Gender	2.63	3.57	0.74	1.00
Age	-0.01	0.18	-0.06	1.00
Ethnicity	0.46	1.43	0.32	1.00
Education	0.13	1.99	0.06	1.00
Political Orientation	-0.93	1.31	-0.71	1.00
Income	-1.52	1.83	-0.83	1.00
Numeracy (BNT-S)	-0.19	1.25	-0.15	1.00
Graph Literacy	-1.75	0.71	-2.45	0.18
Transcript Strength (Weak)*Key Intervention v. Control	-5.61	9.56	-0.59	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-1.61	9.04	-0.18	1.00

Note: Overall model fit, F (13, 173) = 6.08, $p < .00001$, adjusted $r^2 = .26$

Table J.4.
WITNESS ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	9.66	0.99	9.70	0.00
Transcript Strength (Weak)	-1.42	0.57	-2.48	0.17
Key Intervention v. Control	0.64	0.59	1.09	1.00
Key Intervention v. modified Henderson	-0.49	0.58	-0.84	1.00
Gender	0.22	0.31	0.71	1.00
Age	0.00	0.02	-0.13	1.00
Ethnicity	0.12	0.13	0.94	1.00
Education	0.14	0.18	0.80	1.00
Political Orientation	-0.09	0.12	-0.82	1.00
Income	-0.27	0.16	-1.70	1.00
Numeracy (BNT-S)	-0.05	0.11	-0.49	1.00
Graph Literacy	-0.20	0.06	-3.23	0.02
Transcript Strength (Weak)*Key Intervention v. Control	-1.28	0.84	-1.52	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.31	0.80	-0.38	1.00

Note: Overall model fit, F (13, 173) = 3.39, p = .0001, adjusted r^2 = .14

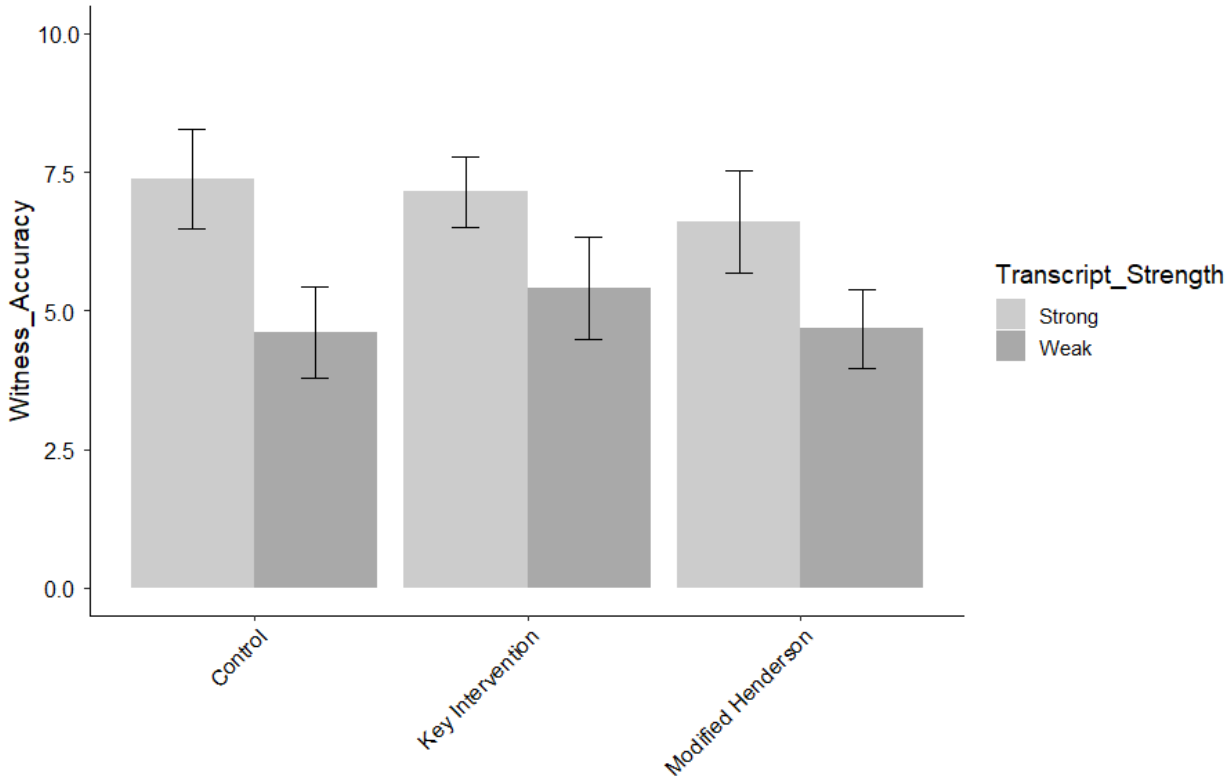


Figure J.1. Average WITNESS ACCURACY as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.5.
CASE STRENGTH

	B	*SE*	*t*	*p*
(Intercept)	8.63	1.15	7.54	0.00
Transcript Strength (Weak)	-1.03	0.66	-1.57	1.00
Key Intervention v. Control	0.93	0.68	1.37	1.00
Key Intervention v. modified Henderson	-0.05	0.67	-0.07	1.00
Gender	0.34	0.36	0.93	1.00
Age	-0.02	0.02	-1.14	1.00
Ethnicity	0.07	0.14	0.49	1.00
Education	0.32	0.20	1.60	1.00
Political Orientation	-0.27	0.13	-2.04	0.51
Income	-0.17	0.19	-0.89	1.00
Numeracy (BNT-S)	0.07	0.13	0.54	1.00
Graph Literacy	-0.20	0.07	-2.82	0.07
Transcript Strength (Weak)*Key Intervention v. Control	-1.24	0.97	-1.28	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.26	0.92	-0.28	1.00

Note: Overall model fit, F (13, 173) = 4.16, $p < .0001$, adjusted $r^2 = .18$

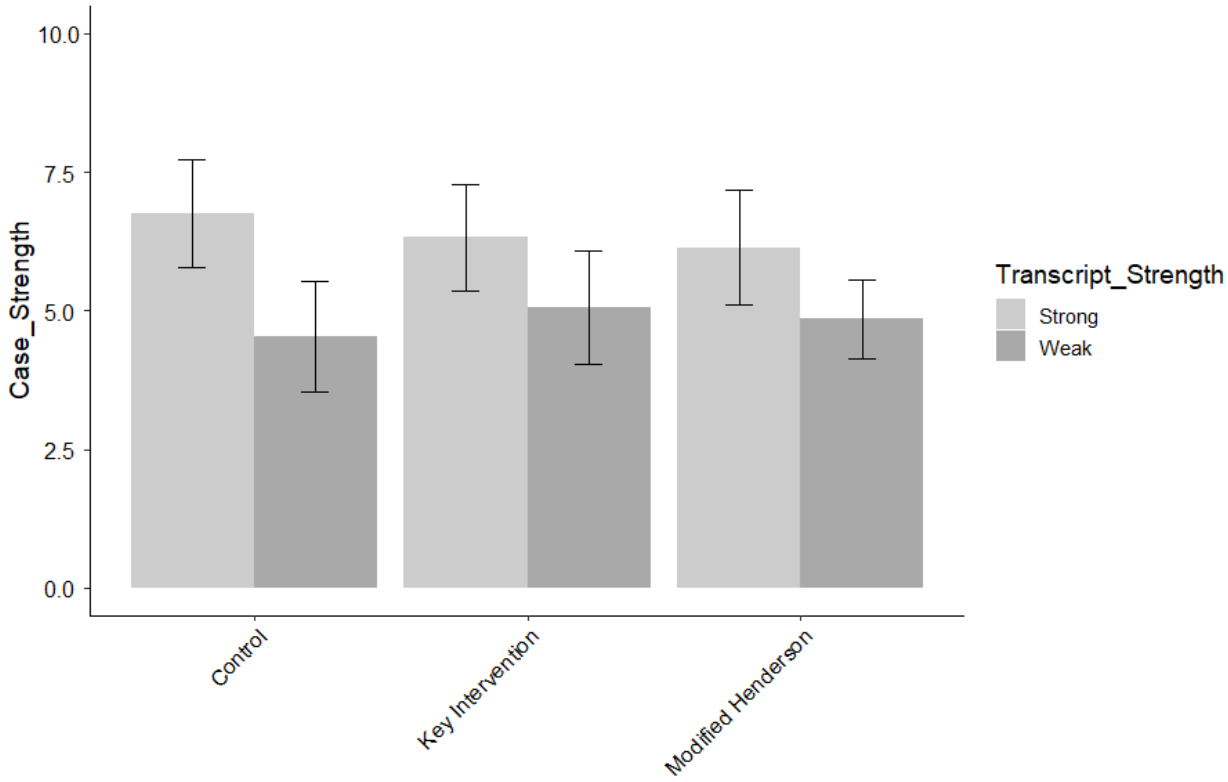


Figure J.2. Average CASE STRENGTH as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.6.
SENTENCE LENGTH RECOMMENDATION

	B	*SE*	*t*	*p*
(Intercept)	8.96	1.03	8.69	0.00
Transcript Strength (Weak)	-0.21	0.59	-0.36	1.00
Key Intervention v. Control	0.28	0.61	0.45	1.00
Key Intervention v. modified Henderson	0.30	0.61	0.50	1.00
Gender	-0.19	0.33	-0.58	1.00
Age	0.02	0.02	1.21	1.00
Ethnicity	0.29	0.13	2.21	0.32
Education	-0.03	0.18	-0.17	1.00
Political Orientation	-0.42	0.12	-3.50	0.01
Income	0.13	0.17	0.79	1.00
Numeracy (BNT-S)	-0.03	0.11	-0.24	1.00
Graph Literacy	-0.21	0.07	-3.23	0.02
Transcript Strength (Weak)*Key Intervention v. Control	-1.13	0.87	-1.29	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.76	0.83	-0.92	1.00

Note: Overall model fit, F (13, 173) = 8.19, $p < .0001$, adjusted $r^2 = .33$

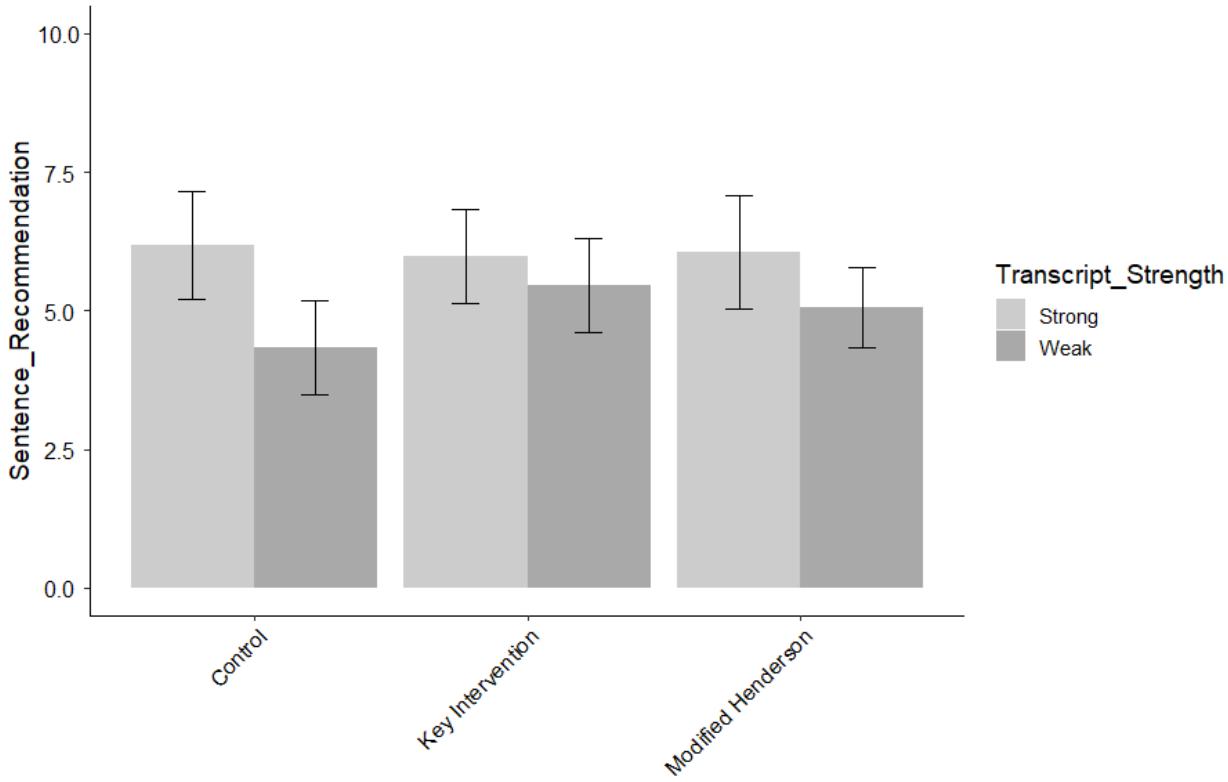


Figure J.3. Average SENTENCE LENGTH RECOMMENDATION as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.7.
CONFIDENCE IN WITNESS

Estimate	*B*	*SE*	*t*	*p*
(Intercept)	11.70	1.16	10.12	0.00
Transcript Strength (Weak)	-2.65	0.66	-3.99	0.00
Key Intervention v. Control	0.20	0.69	0.30	1.00
Key Intervention v. modified Henderson	-0.85	0.68	-1.25	1.00
Gender	-0.02	0.37	-0.04	1.00
Age	0.00	0.02	0.11	1.00
Ethnicity	0.37	0.15	2.54	0.13
Education	-0.02	0.20	-0.10	1.00
Political Orientation	-0.29	0.13	-2.19	0.30
Income	-0.24	0.19	-1.27	1.00
Numeracy (BNT-S)	-0.02	0.13	-0.17	1.00
Graph Literacy	-0.21	0.07	-2.86	0.06
Transcript Strength (Weak)*Key Intervention v. Control	-0.64	0.98	-0.65	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.27	0.93	0.29	1.00

Note: Overall model fit, F (13, 173) = 8.19, $p < .0001$, adjusted $r^2 = .33$

Table J.8.
INITIAL CONFIDENCE INFLUENCE VERDICT

	B	*SE*	*t*	*p*
(Intercept)	5.93	1.10	5.40	0.00
Transcript Strength (Weak)	0.26	0.65	0.40	1.00
Key Intervention v. Control	-0.36	0.65	-0.56	1.00
Key Intervention v. modified Henderson	-0.93	0.64	-1.46	1.00
Gender	0.07	0.35	0.20	1.00
Age	0.03	0.02	1.81	0.94
Ethnicity	0.15	0.14	1.07	1.00
Education	0.04	0.20	0.19	1.00
Political Orientation	0.11	0.13	0.82	1.00
Income	0.16	0.18	0.86	1.00
Numeracy (BNT-S)	0.02	0.12	0.20	1.00
Graph Literacy	-0.01	0.07	-0.18	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.44	0.95	-0.46	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.53	0.90	0.59	1.00

Note: Overall model fit, F (13, 160) = .91, p = .54, adjusted r^2 = -.007

Table J.9.
COURTROOM CONFIDENCE INFLUENCE VERDICT

	B	*SE*	*t*	*p*
(Intercept)	10.78	1.15	9.36	0.00
Transcript Strength (Weak)	-2.27	0.69	-3.31	0.02
Key Intervention v. Control	0.27	0.70	0.39	1.00
Key Intervention v. modified Henderson	-0.56	0.68	-0.82	1.00
Gender	0.02	0.37	0.04	1.00
Age	-0.01	0.02	-0.52	1.00
Ethnicity	0.20	0.15	1.33	1.00
Education	-0.08	0.20	-0.40	1.00
Political Orientation	0.06	0.14	0.41	1.00
Income	-0.30	0.19	-1.57	1.00
Numeracy (BNT-S)	-0.23	0.13	-1.74	0.92
Graph Literacy	-0.17	0.07	-2.24	0.32
Transcript Strength (Weak)*Key Intervention v. Control	-0.11	1.00	-0.11	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.98	0.93	1.05	1.00

Note: Overall model fit, F (13, 160) = 4.96, $p < .0001$, adjusted $r^2 = .23$

Table J.10.
GOOD LOOK

	B	*SE*	*t*	*p*
(Intercept)	8.28	1.11	7.45	0.00
Transcript Strength (Weak)	-0.73	0.64	-1.14	1.00
Key Intervention v. Control	0.36	0.66	0.54	1.00
Key Intervention v. modified Henderson	-0.64	0.65	-0.97	1.00
Gender	-0.04	0.35	-0.11	1.00
Age	0.02	0.02	1.29	1.00
Ethnicity	-0.01	0.14	-0.10	1.00
Education	0.20	0.20	1.02	1.00
Political Orientation	-0.01	0.13	-0.07	1.00
Income	-0.22	0.18	-1.20	1.00
Numeracy (BNT-S)	-0.12	0.12	-1.01	1.00
Graph Literacy	-0.18	0.07	-2.52	0.17
Transcript Strength (Weak)*Key Intervention v. Control	-1.42	0.94	-1.51	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.18	0.89	-0.20	1.00

Note: Overall model fit, F (13, 173) = 3.03, p = .0005, adjusted r^2 = .12

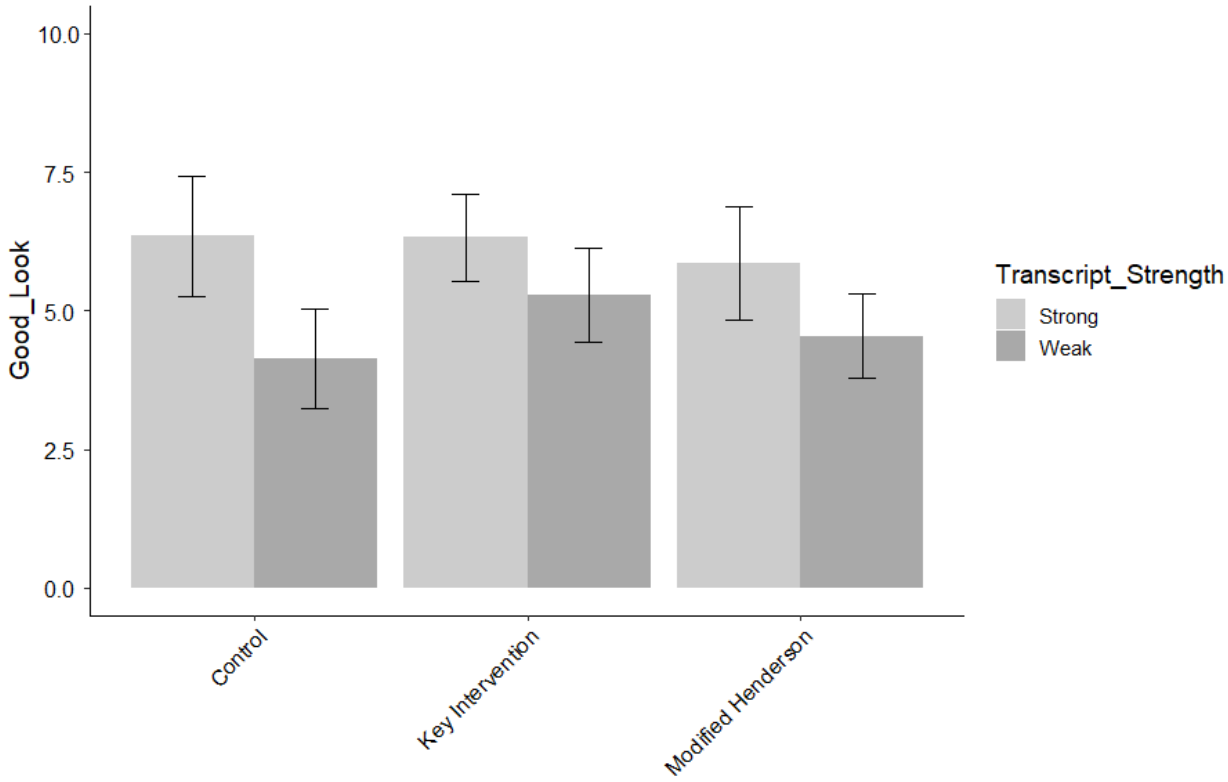


Figure J.4. Average GOOD LOOK as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.11.
ATTENTION

	B	SE	*t*	*p*
(Intercept)	9.37	1.09	8.63	0.00
Transcript Strength (Weak)	-1.27	0.62	-2.04	0.43
Key Intervention v. Control	0.15	0.64	0.23	1.00
Key Intervention v. modified Henderson	-1.64	0.64	-2.56	0.14
Gender	0.20	0.34	0.58	1.00
Age	0.02	0.02	0.95	1.00
Ethnicity	0.08	0.14	0.55	1.00
Education	0.20	0.19	1.03	1.00
Political Orientation	-0.01	0.13	-0.11	1.00
Income	-0.44	0.18	-2.51	0.14
Numeracy (BNT-S)	0.00	0.12	-0.04	1.00
Graph Literacy	-0.24	0.07	-3.46	0.01
Transcript Strength (Weak)*Key Intervention v. Control	-0.50	0.92	-0.55	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.88	0.87	1.01	1.00

Note: Overall model fit, F (13, 173) = 4.20, $p < .0001$, adjusted $r^2 = .18$

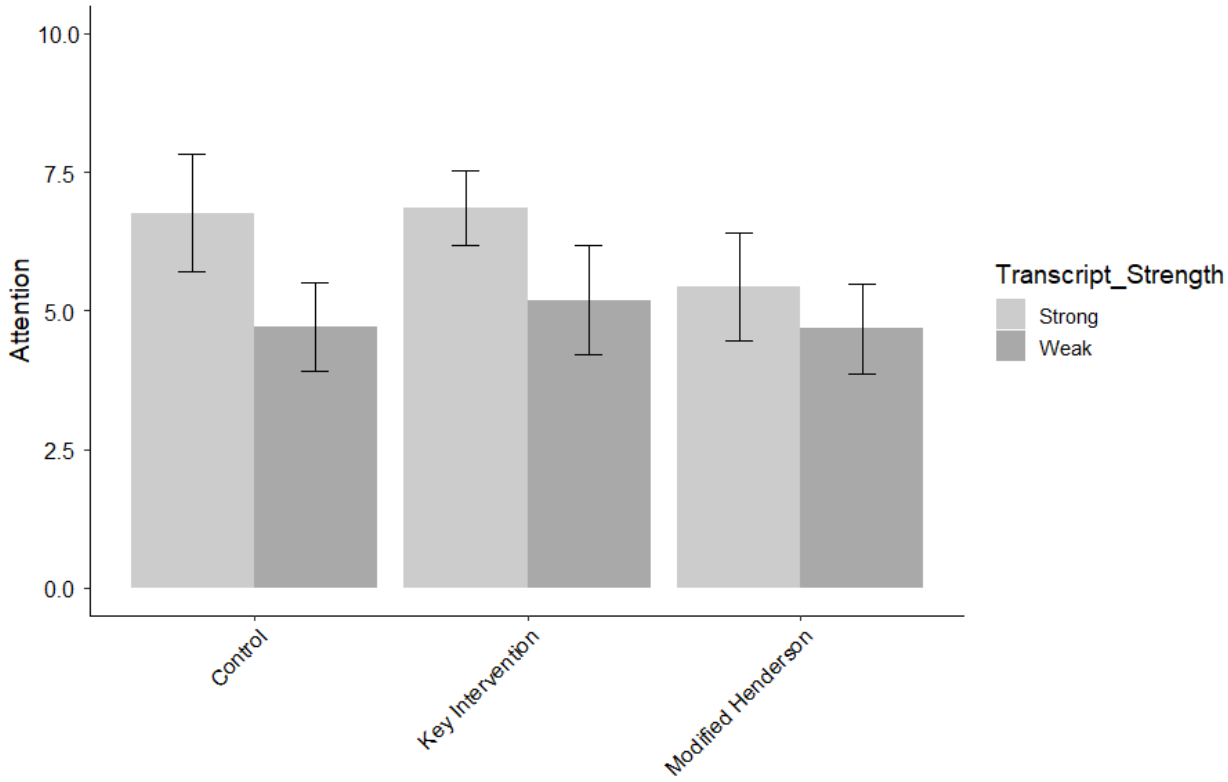


Figure J.5. Average ATTENTION as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.12.
GOOD BASIS

	B	*SE*	*t*	*p*
(Intercept)	8.78	1.12	7.81	0.00
Transcript Strength (Weak)	-1.26	0.65	-1.95	0.63
Key Intervention v. Control	0.20	0.67	0.29	1.00
Key Intervention v. modified Henderson	-1.09	0.66	-1.65	1.00
Gender	0.15	0.36	0.42	1.00
Age	0.01	0.02	0.52	1.00
Ethnicity	0.03	0.14	0.24	1.00
Education	0.20	0.20	0.99	1.00
Political Orientation	-0.08	0.13	-0.62	1.00
Income	-0.23	0.18	-1.25	1.00
Numeracy (BNT-S)	-0.05	0.12	-0.40	1.00
Graph Literacy	-0.18	0.07	-2.51	0.17
Transcript Strength (Weak)*Key Intervention v. Control	-1.17	0.95	-1.24	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.16	0.90	0.18	1.00

Note: Overall model fit, F (13, 173) = 3.66, $p < .0001$, adjusted $r^2 = .16$

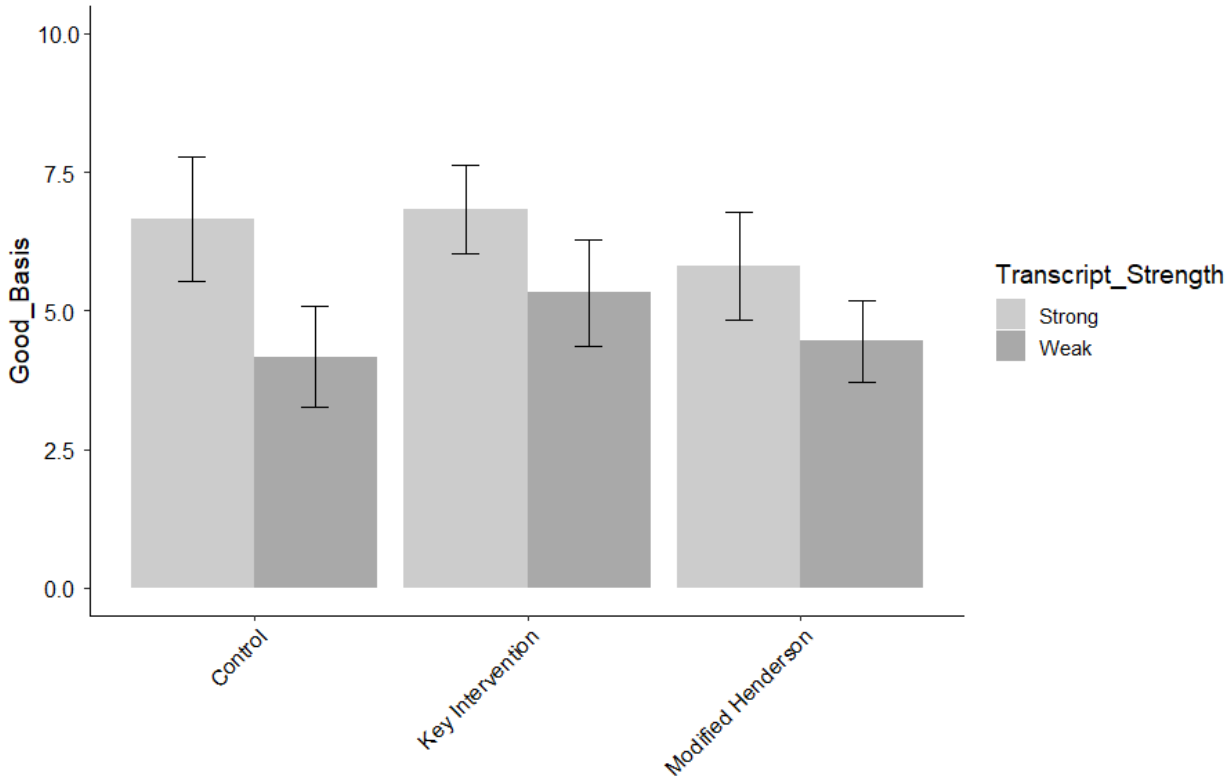


Figure J.6. Average GOOD BASIS as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.13.
WITNESS MEMORY

	B	*SE*	*t*	*p*
(Intercept)	7.65	0.94	8.16	0.00
Transcript Strength (Weak)	-1.55	0.54	-2.87	0.06
Key Intervention v. Control	0.44	0.56	0.78	1.00
Key Intervention v. modified Henderson	-0.61	0.55	-1.10	1.00
Gender	0.46	0.30	1.54	1.00
Age	0.00	0.01	0.24	1.00
Ethnicity	0.11	0.12	0.93	1.00
Education	0.31	0.17	1.89	0.67
Political Orientation	-0.05	0.11	-0.47	1.00
Income	-0.22	0.15	-1.44	1.00
Numeracy (BNT-S)	0.01	0.10	0.14	1.00
Graph Literacy	-0.19	0.06	-3.15	0.03
Transcript Strength (Weak)*Key Intervention v. Control	-0.60	0.79	-0.76	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.09	0.75	-0.12	1.00

Note: Overall model fit, F (13, 173) = 5.61, $p < .0001$, adjusted $r^2 = .24$

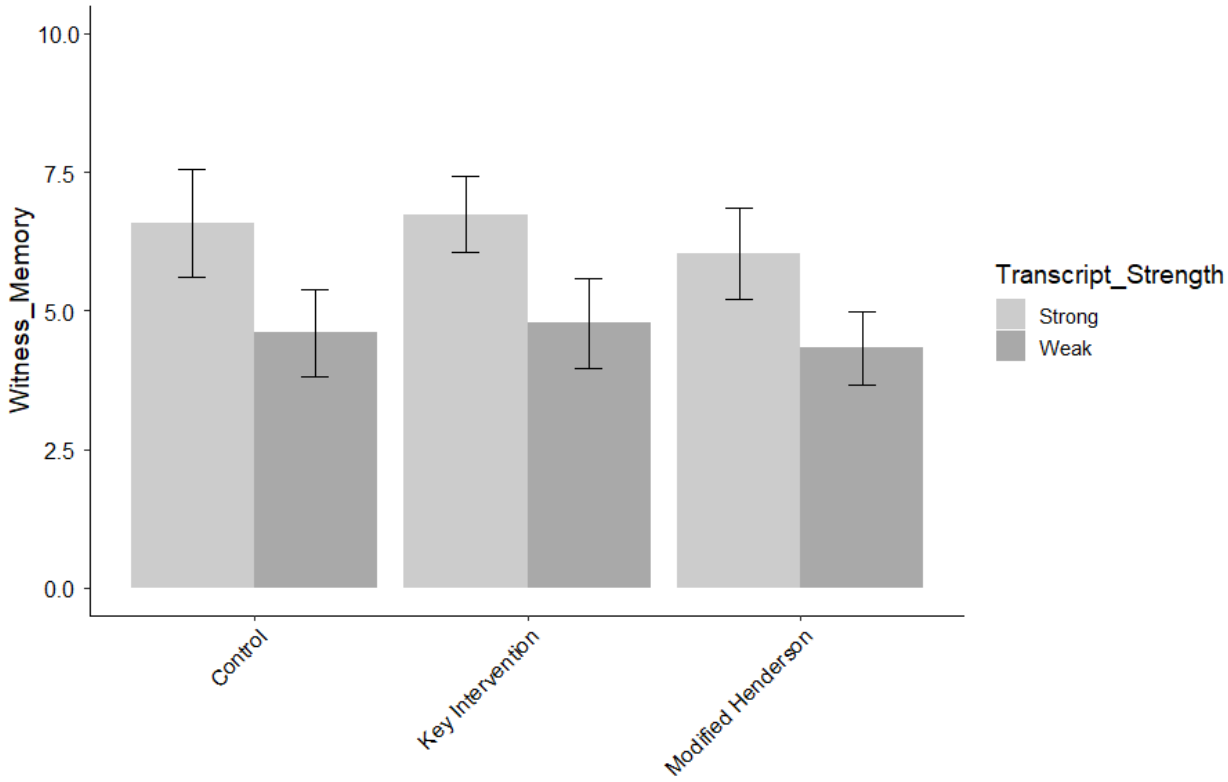


Figure J.7. Average WITNESS MEMORY as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.14.
EYEWITNESS ACCURACY GENERAL

	B	*SE*	*t*	*p*
(Intercept)	8.25	0.79	10.38	0.00
Transcript Strength (Weak)	-1.16	0.46	-2.54	0.15
Key Intervention v. Control	-0.59	0.47	-1.26	1.00
Key Intervention v. modified Henderson	-0.17	0.47	-0.37	1.00
Gender	-0.25	0.25	-0.99	1.00
Age	0.02	0.01	1.45	1.00
Ethnicity	0.25	0.10	2.45	0.19
Education	-0.14	0.14	-0.96	1.00
Political Orientation	-0.14	0.09	-1.50	1.00
Income	0.18	0.13	1.38	1.00
Numeracy (BNT-S)	-0.15	0.09	-1.68	0.94
Graph Literacy	-0.10	0.05	-2.03	0.48
Transcript Strength (Weak)*Key Intervention v. Control	0.89	0.67	1.32	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.12	0.64	0.19	1.00

Note: Overall model fit, F (13, 173) = 3.94, $p < .0001$, adjusted $r^2 = .17$

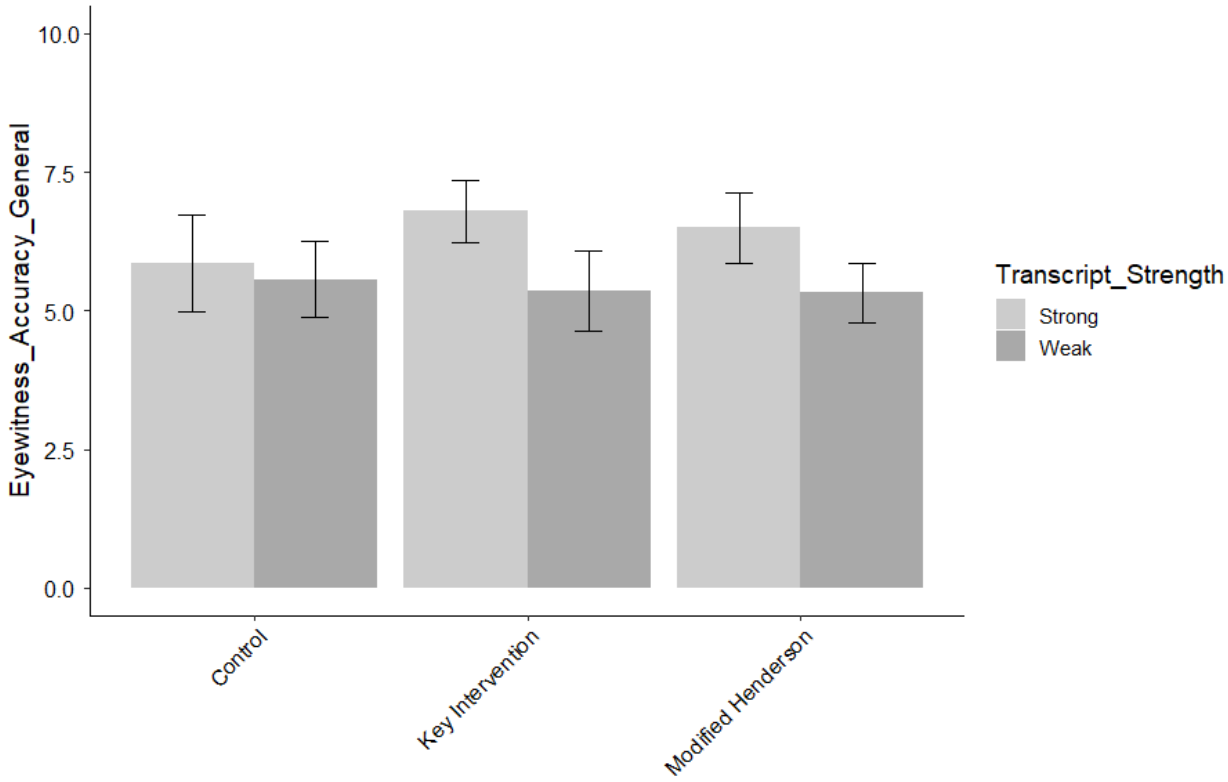


Figure J.8. Average EYEWITNESS ACCURACY GENERAL as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.15.
CONFIDENCE INDICATOR ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	8.19	0.95	8.66	0.00
Transcript Strength (Weak)	-1.01	0.54	-1.87	0.77
Key Intervention v. Control	-0.38	0.56	-0.67	1.00
Key Intervention v. modified Henderson	-0.06	0.56	-0.11	1.00
Gender	0.13	0.30	0.44	1.00
Age	0.02	0.02	1.53	1.00
Ethnicity	0.15	0.12	1.22	1.00
Education	0.19	0.17	1.15	1.00
Political Orientation	-0.09	0.11	-0.84	1.00
Income	-0.09	0.15	-0.60	1.00
Numeracy (BNT-S)	-0.13	0.10	-1.23	1.00
Graph Literacy	-0.17	0.06	-2.90	0.06
Transcript Strength (Weak)*Key Intervention v. Control	0.28	0.80	0.35	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.06	0.76	0.07	1.00

Note: Overall model fit, F (13, 173) = 3.49, $p < .0001$, adjusted $r^2 = .15$

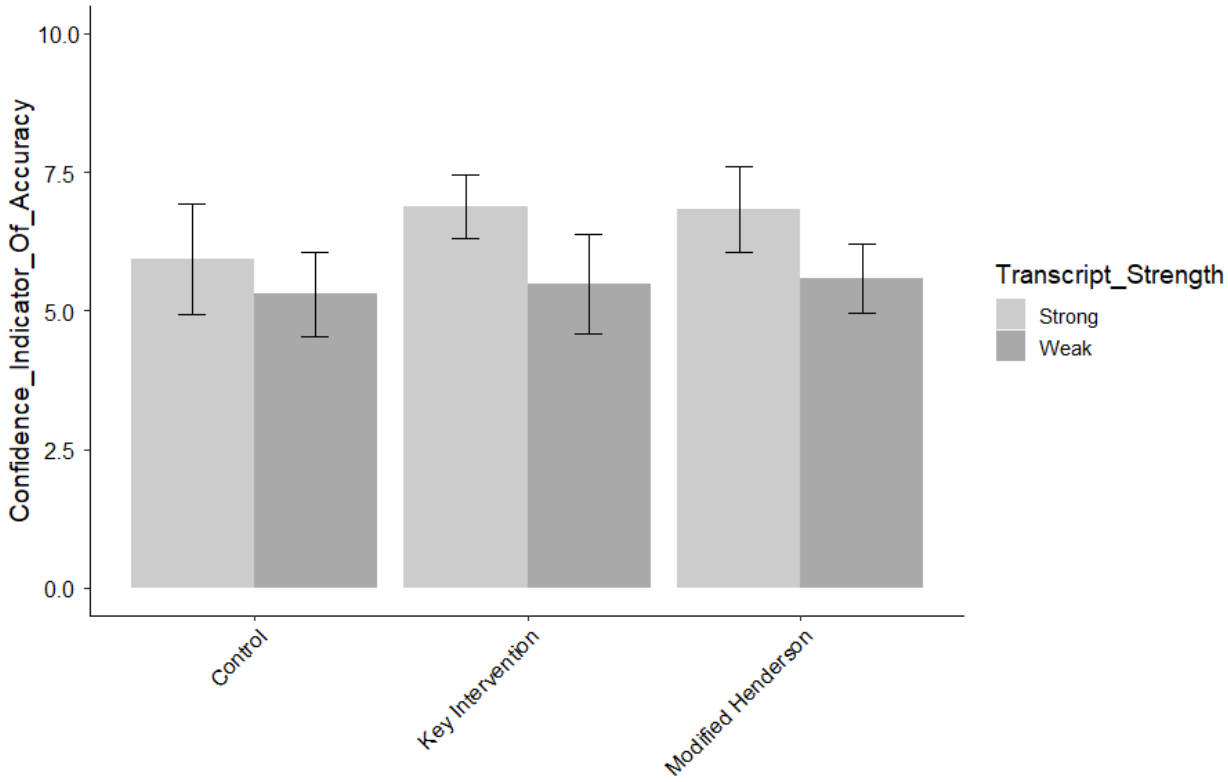


Figure J.9. Average CONFIDENCE INDICATOR OF ACCURACY as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.16.

CONFIDENCE INFLATION OCCURS

	B	*SE*	*t*	*p*
(Intercept)	7.29	1.22	5.98	0.00
Transcript Strength (Weak)	-1.14	0.70	-1.63	1.00
Key Intervention v. Control	-0.95	0.72	-1.31	1.00
Key Intervention v. modified Henderson	-0.12	0.72	-0.17	1.00
Gender	-0.40	0.39	-1.04	1.00
Age	0.01	0.02	0.72	1.00
Ethnicity	0.32	0.15	2.11	0.43
Education	0.12	0.22	0.56	1.00
Political Orientation	0.10	0.15	0.68	1.00
Income	-0.43	0.20	-2.11	0.43
Numeracy (BNT-S)	-0.04	0.14	-0.32	1.00
Graph Literacy	-0.12	0.08	-1.51	1.00
Transcript Strength (Weak)*Key Intervention v. Control	2.29	1.03	2.22	0.36
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.59	0.98	0.61	1.00

Note: Overall model fit, F (13, 172) = 1.96, p = .03, adjusted r^2 = .06

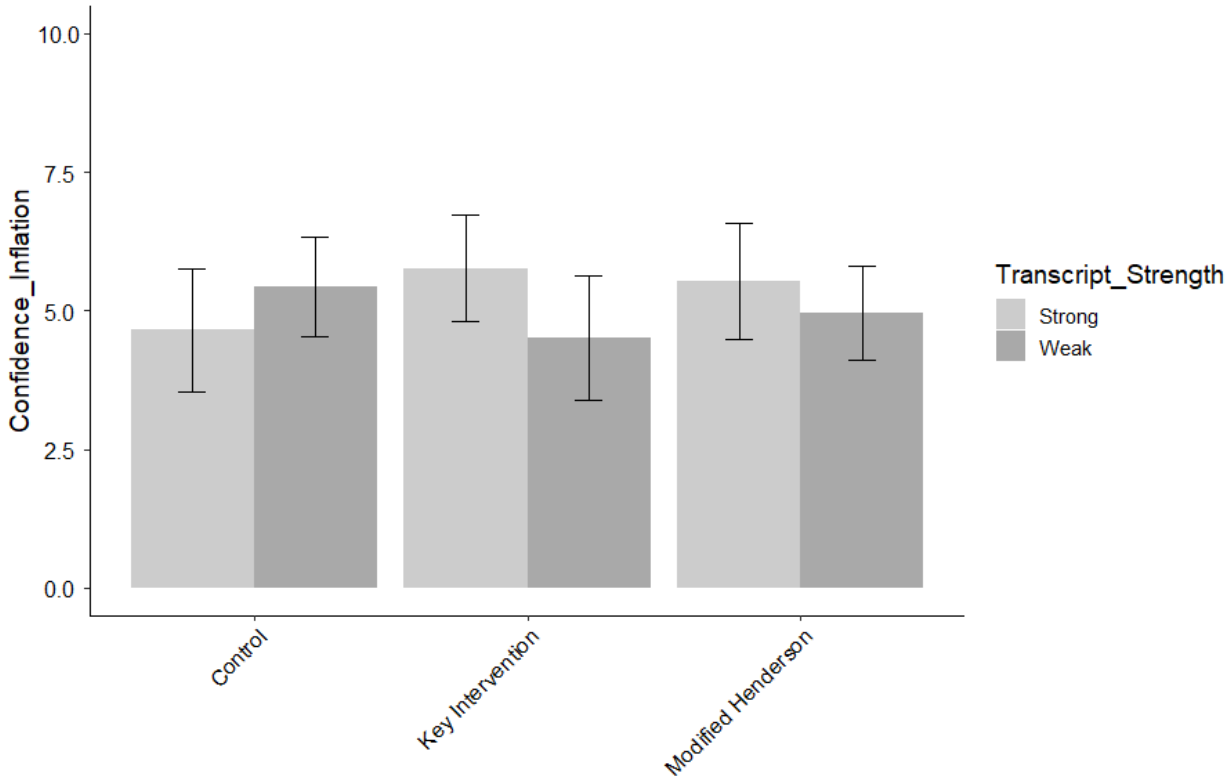


Figure J.10. Average CONFIDENCE INFLATION OCCURS as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.17.
CONFIDENCE INFLATION EQUALS ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	7.82	1.09	7.20	0.00
Transcript Strength (Weak)	-0.15	0.62	-0.24	1.00
Key Intervention v. Control	-0.07	0.64	-0.11	1.00
Key Intervention v. modified Henderson	-0.46	0.64	-0.72	1.00
Gender	0.22	0.34	0.63	1.00
Age	-0.02	0.02	-1.24	1.00
Ethnicity	0.20	0.14	1.48	1.00
Education	0.02	0.19	0.12	1.00
Political Orientation	0.02	0.13	0.20	1.00
Income	-0.04	0.18	-0.20	1.00
Numeracy (BNT-S)	-0.04	0.12	-0.31	1.00
Graph Literacy	-0.23	0.07	-3.37	0.01
Transcript Strength (Weak)*Key Intervention v. Control	0.73	0.92	0.80	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.10	0.87	0.11	1.00

Note: Overall model fit, $F(13, 173) = 2.49$, $p = .0004$, adjusted $r^2 = .09$

Table J.18.
COMPREHENSION CHECK QUESTIONS

	B	*SE*	*t*	*p*
(Intercept)	3.08	0.18	17.02	0.00
Transcript Strength (Weak)	0.08	0.10	0.77	1.00
Key Intervention v. Control	-0.04	0.11	-0.39	1.00
Key Intervention v. modified Henderson	0.12	0.11	1.12	1.00
Gender	0.16	0.06	2.85	0.06
Age	0.00	0.00	0.61	1.00
Ethnicity	0.04	0.02	1.57	1.00
Education	-0.02	0.03	-0.55	1.00
Political Orientation	-0.03	0.02	-1.24	1.00
Income	-0.03	0.03	-1.05	1.00
Numeracy (BNT-S)	0.03	0.02	1.70	0.92
Graph Literacy	0.04	0.01	3.89	0.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.15	0.15	-0.97	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.28	0.15	-1.95	0.58

Note: Overall model fit, F (13, 173) = 3.82, $p < .0001$, adjusted $r^2 = .16$

Table J.19.
INITIAL CONFIDENCE PERCENTAGE

	B	*SE*	*t*	*p*
(Intercept)	10.44	0.69	15.13	0.00
Transcript Strength (Weak)	-6.17	0.41	-15.20	0.00
Key Intervention v. Control	0.17	0.41	0.43	1.00
Key Intervention v. modified Henderson	0.20	0.40	0.51	1.00
Gender	-0.22	0.22	-0.99	1.00
Age	0.00	0.01	-0.25	1.00
Ethnicity	0.24	0.09	2.77	0.08
Education	-0.12	0.12	-0.96	1.00
Political Orientation	-0.04	0.08	-0.45	1.00
Income	0.01	0.12	0.11	1.00
Numeracy (BNT-S)	-0.13	0.08	-1.66	1.00
Graph Literacy	-0.07	0.04	-1.69	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.32	0.59	0.53	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.30	0.56	0.53	1.00

Note: Overall model fit, F (13, 160) = 57.62, $p < .0001$, adjusted $r^2 = .81$

Table J.20.
COURTROOM CONFIDENCE PERCENTAGE

	B	*SE*	*t*	*p*
(Intercept)	6.84	0.51	13.29	0.00
Transcript Strength (Weak)	-0.30	0.31	-0.97	1.00
Key Intervention v. Control	-0.22	0.31	-0.69	1.00
Key Intervention v. modified Henderson	0.08	0.30	0.27	1.00
Gender	0.25	0.17	1.53	1.00
Age	0.01	0.01	1.52	1.00
Ethnicity	0.10	0.07	1.45	1.00
Education	-0.16	0.09	-1.76	0.88
Political Orientation	-0.02	0.06	-0.36	1.00
Income	0.04	0.09	0.49	1.00
Numeracy (BNT-S)	0.14	0.06	2.36	0.23
Graph Literacy	0.11	0.03	3.27	0.02
Transcript Strength (Weak)*Key Intervention v. Control	-0.43	0.44	-0.97	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.30	0.42	-0.73	1.00

Note: Overall model fit, F (13, 160) = 4.39, $p < .0001$, adjusted $r^2 = .20$

Table J.21.
CONFIDENCE RANKING

	B	*SE*	*t*	*p*
(Intercept)	2.22	0.21	10.46	0.00
Transcript Strength (Weak)	-0.85	0.12	-6.83	0.00
Key Intervention v. Control	0.04	0.12	0.29	1.00
Key Intervention v. modified Henderson	0.06	0.12	0.48	1.00
Gender	-0.11	0.07	-1.61	1.00
Age	0.00	0.00	0.26	1.00
Ethnicity	-0.01	0.03	-0.36	1.00
Education	0.01	0.04	0.35	1.00
Political Orientation	-0.02	0.02	-0.77	1.00
Income	0.07	0.04	1.76	0.98
Numeracy (BNT-S)	-0.04	0.02	-1.72	0.98
Graph Literacy	-0.01	0.01	-0.70	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.13	0.18	0.73	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.03	0.17	-0.16	1.00

Note: Overall model fit, F (13, 123) = 11.76, $p < .0001$, adjusted $r^2 = .51$

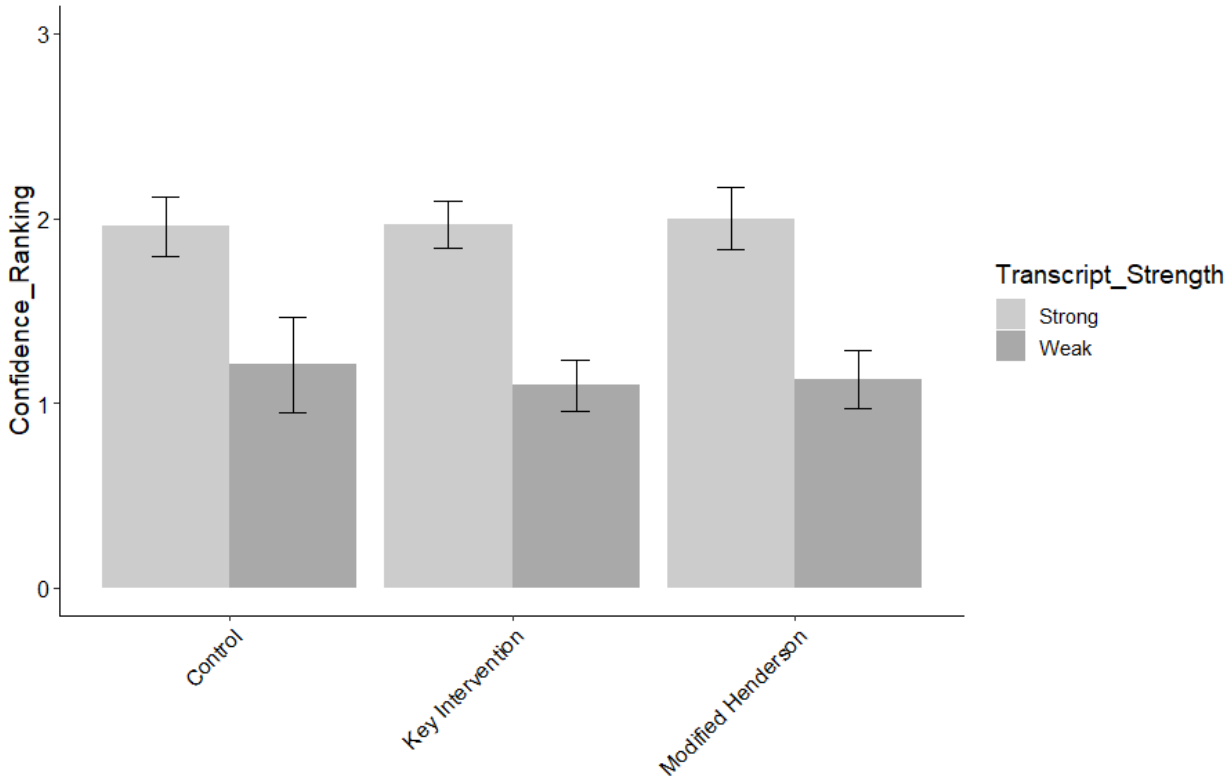


Figure J.11. Average CONFIDENCE RANKING as a function of Transcript Strength and Intervention type. Error bars represent 95% confidence intervals.

Table J.22.
USABILITY

	B	*SE*	*t*	*p*
(Intercept)	7.165	0.737	9.725	0
Transcript Strength (Weak)	-0.18	0.42	-0.43	1.00
Key Intervention v. Control	0.20	0.44	0.47	1.00
Key Intervention v. modified Henderson	0.00	0.43	0.01	1.00
Gender	0.66	0.23	2.85	0.06
Age	0.02	0.01	1.68	0.95
Ethnicity	-0.02	0.09	-0.23	1.00
Education	-0.34	0.13	-2.63	0.10
Political Orientation	-0.11	0.09	-1.29	1.00
Income	-0.02	0.12	-0.16	1.00
Numeracy (BNT-S)	0.11	0.08	1.36	1.00
Graph Literacy	0.15	0.05	3.14	0.03
Transcript Strength (Weak)*Key Intervention v. Control	-0.49	0.62	-0.79	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.61	0.59	-1.03	1.00

Note: Overall model fit, F (13, 173) = 4.66, $p < .0001$, adjusted $r^2 = .20$

Table J.23.
WORKLOAD

	B	*SE*	*t*	*p*
(Intercept)	6.39	0.75	8.50	0.00
Transcript Strength (Weak)	0.23	0.43	0.52	1.00
Key Intervention v. Control	0.00	0.45	-0.01	1.00
Key Intervention v. modified Henderson	-0.31	0.44	-0.71	1.00
Gender	-0.17	0.24	-0.72	1.00
Age	0.00	0.01	0.01	1.00
Ethnicity	0.08	0.09	0.84	1.00
Education	0.17	0.13	1.26	1.00
Political Orientation	0.06	0.09	0.69	1.00
Income	0.13	0.12	1.06	1.00
Numeracy (BNT-S)	-0.05	0.08	-0.54	1.00
Graph Literacy	-0.19	0.05	-4.00	0.00
Transcript Strength (Weak)*Key Intervention v. Control	0.36	0.63	0.57	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.65	0.60	1.07	1.00

Note: Overall model fit, F (13, 173) = 3.34 p = .0001, adjusted r^2 = .14

Appendix K

Regression models and graphs, Experiment 2, *n* = 790

Table K.1.

VERDICT

	B	*SE*	*z*	*p*
(Intercept)	3.66	0.58	6.32	0.00
Transcript Strength (Weak)	-1.50	0.41	-3.68	0.00
Memory Test (Unfair)	-0.37	0.42	-0.88	1.00
Key Intervention v. Control	-0.79	0.39	-2.04	0.58
Key Intervention v. modified Henderson	-0.83	0.41	-2.00	0.59
Gender	-0.09	0.16	-0.59	1.00
Age	-0.02	0.01	-2.31	0.33
Ethnicity	0.02	0.06	0.36	1.00
Education	0.08	0.09	0.99	1.00
Political Orientation	-0.16	0.06	-2.67	0.13
Income	-0.13	0.08	-1.69	0.99
Numeracy (BNT-S)	-0.05	0.05	-1.08	1.00
Graph Literacy	-0.13	0.03	-4.78	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.56	0.58	0.96	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.20	0.55	2.20	0.42
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.10	0.58	1.88	0.71
Memory Test (Unfair)*Key Intervention v. Control	0.33	0.55	0.60	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.08	0.59	-0.13	1.00
Weak*Unfair*Key Intervention v. Control	-0.75	0.78	-0.96	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.79	0.83	-0.95	1.00

Note: Overall model fit, *x2* (df = 20) = -445.62, $p < .0001$, $r^2 = .11$, AIC = 931.23

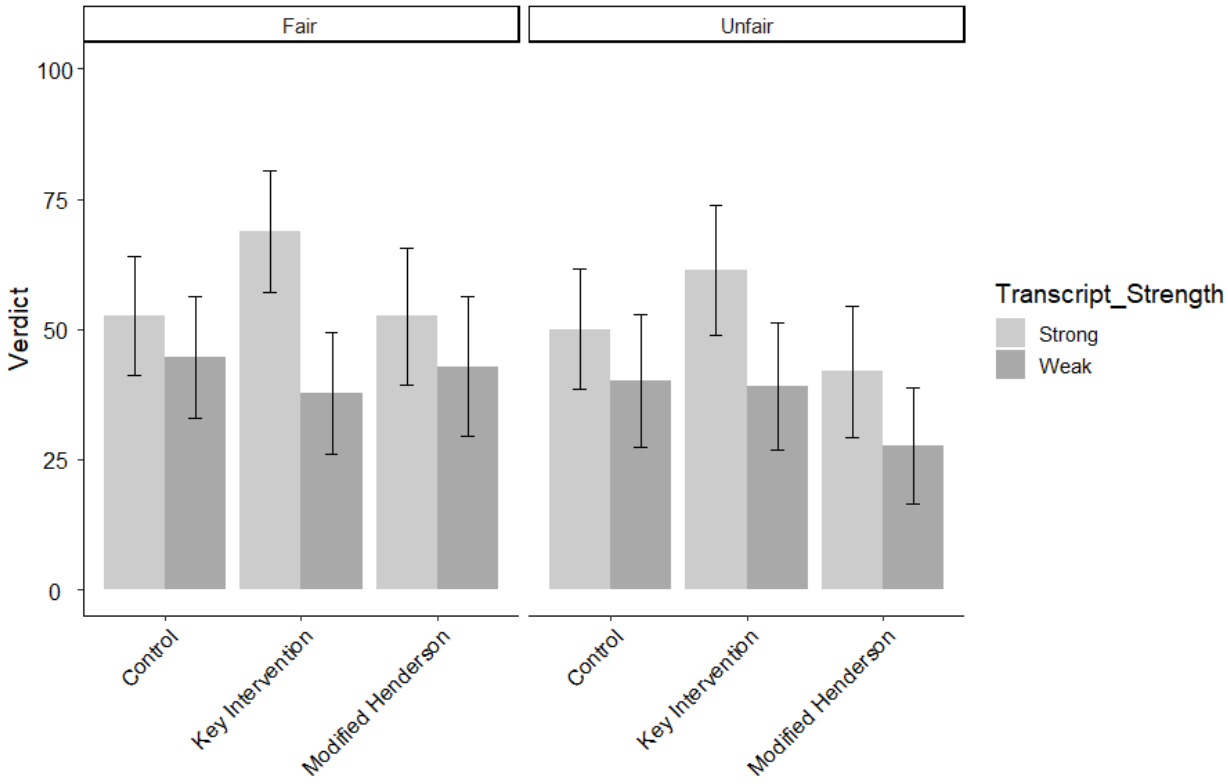


Figure K.1. Average VERDICT percentage as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.2.
CONFIDENCE IN VERDICT

	B	*SE*	*t*	*p*
(Intercept)	9.82	0.47	20.96	0.00
Transcript Strength (Weak)	-0.40	0.34	-1.17	1.00
Memory Test (Unfair)	-0.46	0.36	-1.28	1.00
Key Intervention v. Control	-0.92	0.33	-2.78	0.10
Key Intervention v. modified Henderson	-0.53	0.36	-1.46	1.00
Gender	-0.15	0.14	-1.09	1.00
Age	0.01	0.01	0.97	1.00
Ethnicity	-0.07	0.05	-1.47	1.00
Education	0.00	0.07	-0.03	1.00
Political Orientation	-0.04	0.05	-0.87	1.00
Income	0.00	0.07	0.03	1.00
Numeracy (BNT-S)	-0.16	0.04	-3.82	0.00
Graph Literacy	-0.09	0.02	-3.73	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.50	0.49	1.02	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.12	0.47	0.25	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.47	0.50	0.93	1.00
Memory Test (Unfair)*Key Intervention v. Control	1.06	0.48	2.21	0.44
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.35	0.51	0.69	1.00
Weak*Unfair*Key Intervention v. Control	-0.57	0.68	-0.84	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.87	0.70	-1.24	1.00

Note: Overall model fit, F (19, 707) = 4.91, $p < .0001$, adjusted $r^2 = .09$

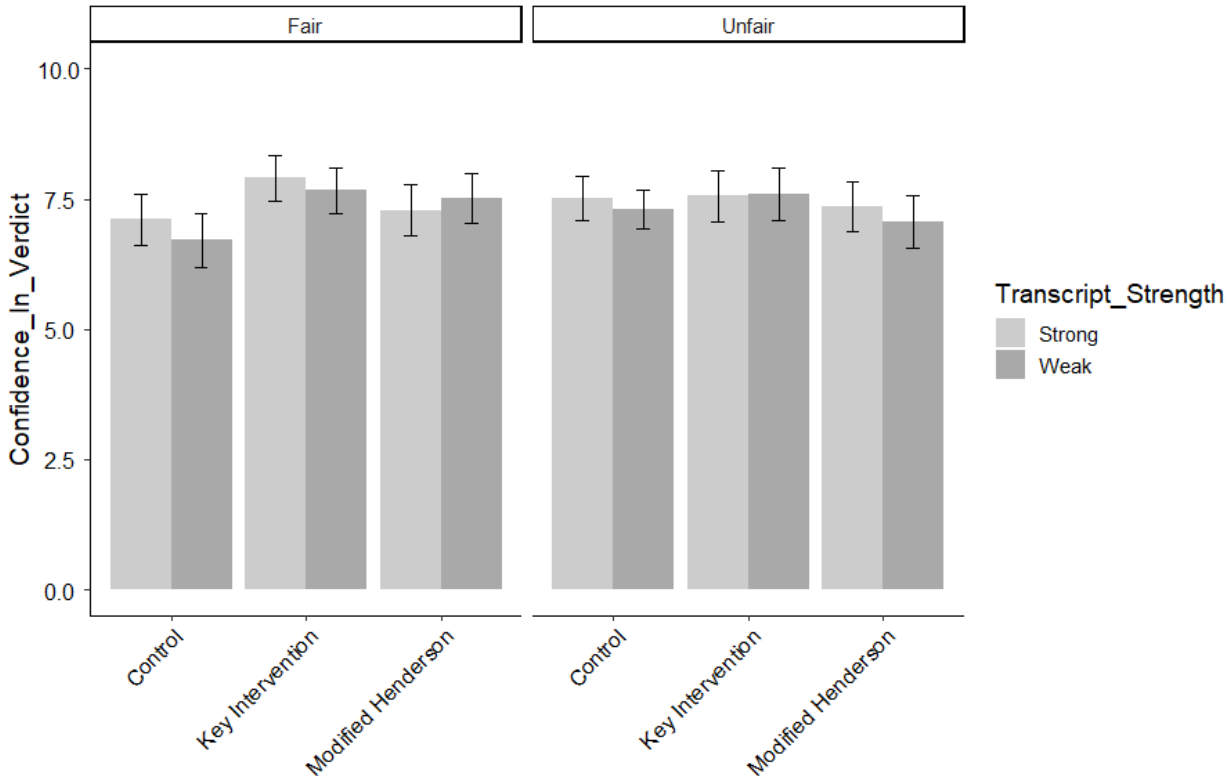


Figure K.2. Average CONFIDENCE IN VERDICT as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.3.
LIKELIHOOD OF GUILT

	B	*SE*	*t*	*p*
(Intercept)	87.26	6.77	12.89	0.00
Transcript Strength (Weak)	-7.49	4.94	-1.52	1.00
Memory Test (Unfair)	1.29	5.16	0.25	1.00
Key Intervention v. Control	1.26	4.80	0.26	1.00
Key Intervention v. modified Henderson	-4.82	5.18	-0.93	1.00
Gender	0.59	2.00	0.30	1.00
Age	-0.10	0.09	-1.17	1.00
Ethnicity	-1.64	0.72	-2.29	0.38
Education	1.58	1.05	1.49	1.00
Political Orientation	-2.78	0.73	-3.82	0.00
Income	-0.61	0.98	-0.62	1.00
Numeracy (BNT-S)	-0.11	0.60	-0.19	1.00
Graph Literacy	-0.90	0.34	-2.68	0.14
Transcript Strength (Weak)*Memory Test (Unfair)	-6.98	7.10	-0.98	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.49	6.76	-0.07	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	5.70	7.20	0.79	1.00
Memory Test (Unfair)*Key Intervention v. Control	-6.90	6.95	-0.99	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-5.28	7.36	-0.72	1.00
Weak*Unfair*Key Intervention v. Control	12.47	9.76	1.28	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-2.33	10.17	-0.23	1.00

Note: Overall model fit, F (19, 706) = 4.19, $p < .0001$, adjusted $r^2 = .08$

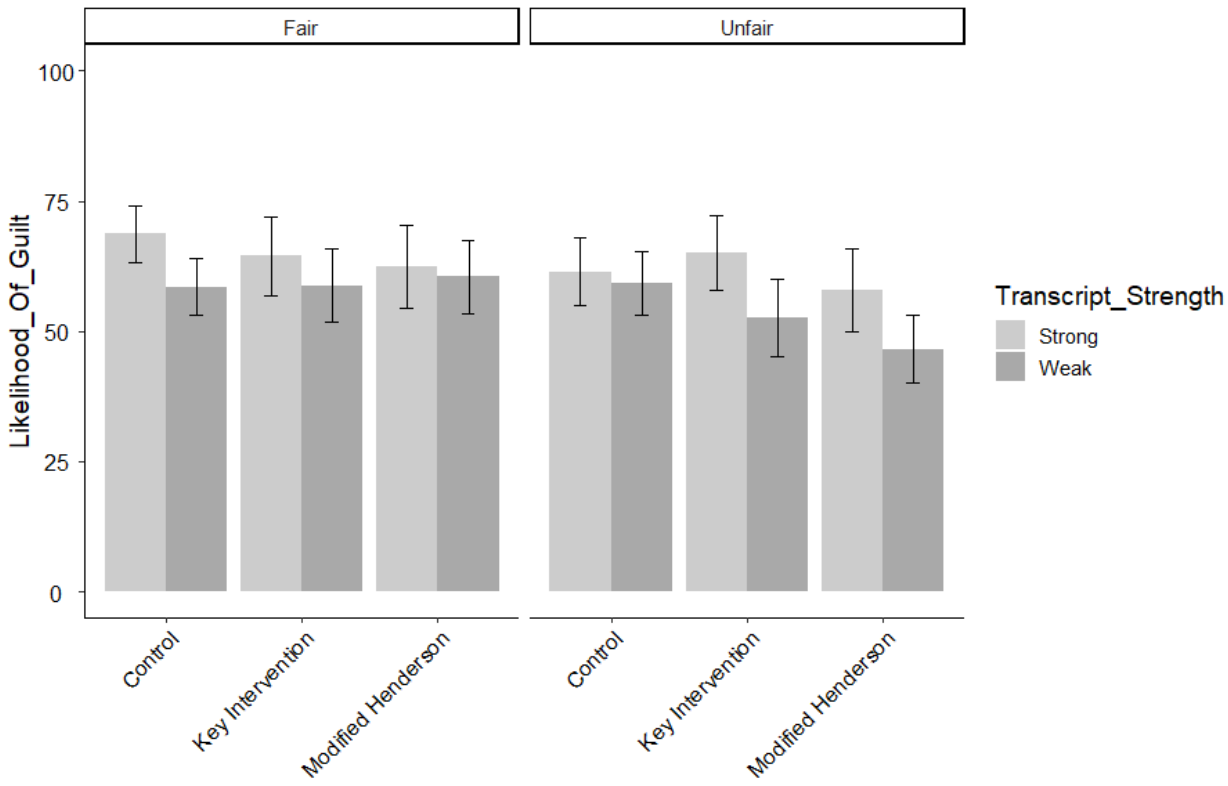


Figure K.3. Average LIKELIHOOD OF GUILT as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.4.
WITNESS ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	9.87	0.62	15.98	0.00
Transcript Strength (Weak)	-1.87	0.45	-4.13	0.00
Memory Test (Unfair)	-0.76	0.47	-1.62	1.00
Key Intervention v. Control	-0.58	0.44	-1.31	1.00
Key Intervention v. modified Henderson	-1.14	0.47	-2.42	0.23
Gender	-0.07	0.18	-0.36	1.00
Age	-0.01	0.01	-0.67	1.00
Ethnicity	-0.11	0.07	-1.70	1.00
Education	0.14	0.10	1.47	1.00
Political Orientation	-0.18	0.07	-2.65	0.14
Income	-0.07	0.09	-0.83	1.00
Numeracy (BNT-S)	-0.06	0.05	-1.03	1.00
Graph Literacy	-0.13	0.03	-4.26	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.76	0.65	1.17	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.06	0.62	1.72	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.70	0.66	2.59	0.16
Memory Test (Unfair)*Key Intervention v. Control	0.25	0.63	0.39	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.47	0.67	0.70	1.00
Weak*Unfair*Key Intervention v. Control	-0.65	0.89	-0.73	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-1.34	0.93	-1.45	1.00

Note: Overall model fit, F (19, 704) = 5.35, $p < .0001$, adjusted $r^2 = .10$

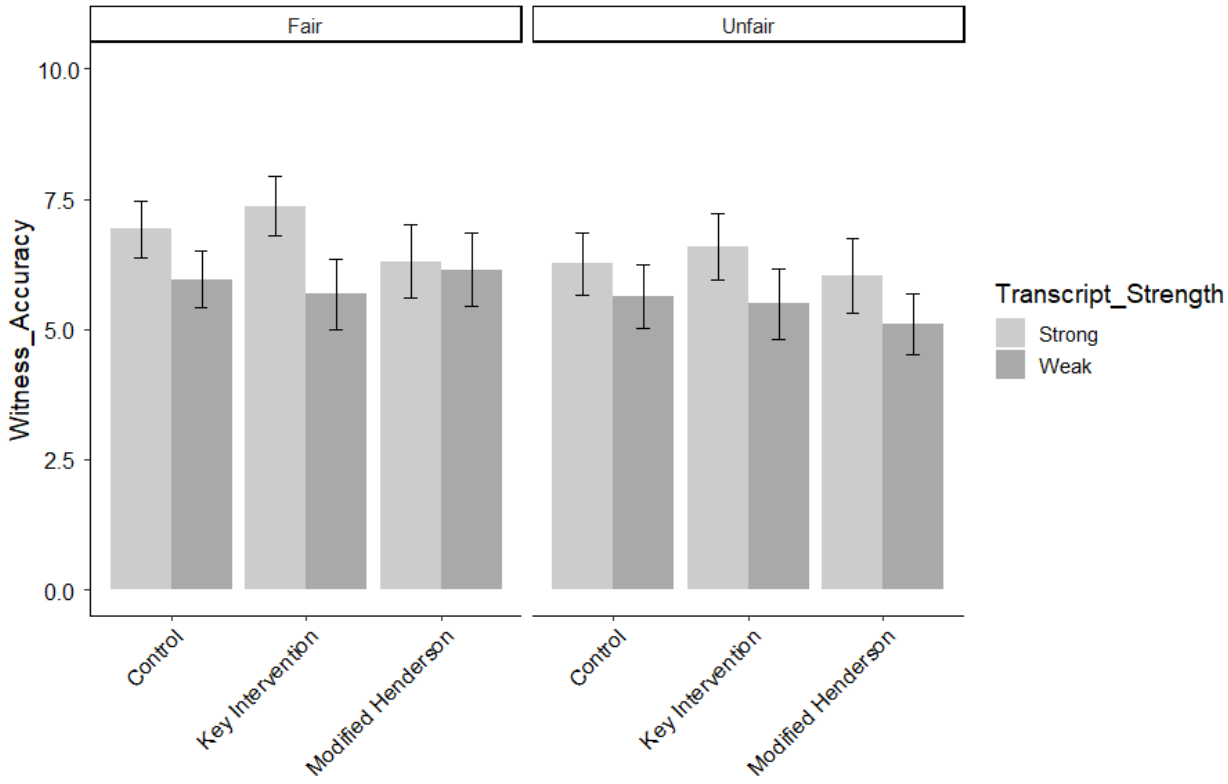


Figure K.4. Average WITNESS ACCURACY as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.5.
CASE STRENGTH

	B	*SE*	*t*	*p*
(Intercept)	9.85	0.63	15.53	0.00
Transcript Strength (Weak)	-1.66	0.46	-3.58	0.01
Memory Test (Unfair)	-1.00	0.48	-2.06	0.64
Key Intervention v. Control	-0.81	0.45	-1.80	1.00
Key Intervention v. modified Henderson	-0.70	0.49	-1.44	1.00
Gender	-0.08	0.19	-0.44	1.00
Age	0.00	0.01	-0.41	1.00
Ethnicity	-0.02	0.07	-0.23	1.00
Education	0.20	0.10	1.99	0.70
Political Orientation	-0.23	0.07	-3.34	0.02
Income	-0.12	0.09	-1.33	1.00
Numeracy (BNT-S)	-0.07	0.06	-1.30	1.00
Graph Literacy	-0.16	0.03	-5.18	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.84	0.67	1.26	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.14	0.63	1.80	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.01	0.68	1.49	1.00
Memory Test (Unfair)*Key Intervention v. Control	1.17	0.65	1.80	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.31	0.69	0.44	1.00
Weak*Unfair*Key Intervention v. Control	-1.20	0.91	-1.31	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-1.07	0.95	-1.12	1.00

Note: Overall model fit, F (19, 707) = 6.47, $p < .0001$, adjusted $r^2 = .13$

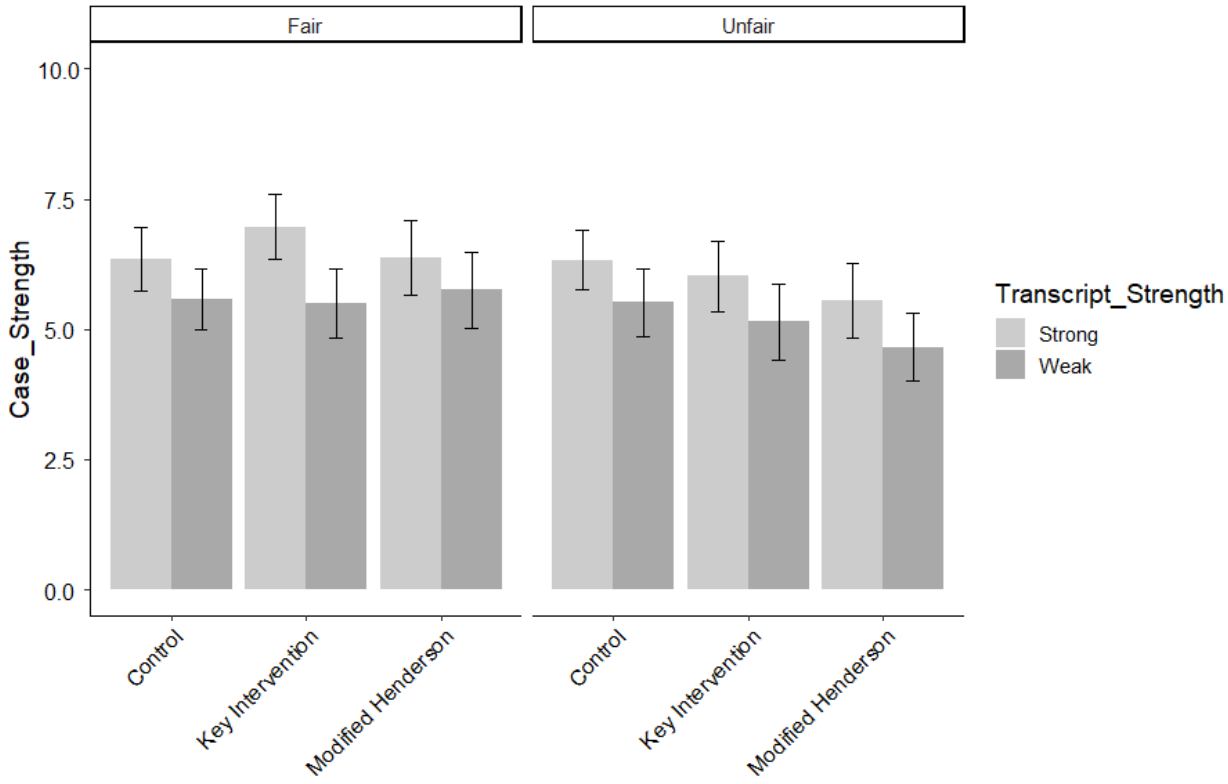


Figure K.5. Average CASE STRENGTH as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.6.
SENTENCE LENGTH RECOMMENDATION

	B	*SE*	*t*	*p*
(Intercept)	8.83	0.57	15.40	0.00
Transcript Strength (Weak)	-0.60	0.42	-1.42	1.00
Memory Test (Unfair)	-0.06	0.44	-0.15	1.00
Key Intervention v. Control	-0.11	0.41	-0.27	1.00
Key Intervention v. modified Henderson	-0.65	0.44	-1.47	1.00
Gender	-0.19	0.17	-1.10	1.00
Age	0.00	0.01	-0.52	1.00
Ethnicity	0.13	0.06	2.13	0.50
Education	0.24	0.09	2.66	0.14
Political Orientation	-0.27	0.06	-4.31	0.00
Income	-0.13	0.08	-1.55	1.00
Numeracy (BNT-S)	-0.06	0.05	-1.09	1.00
Graph Literacy	-0.19	0.03	-6.65	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	-0.13	0.60	-0.21	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.11	0.57	0.20	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.10	0.61	1.80	1.00
Memory Test (Unfair)*Key Intervention v. Control	-0.48	0.59	-0.82	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.68	0.62	1.09	1.00
Weak*Unfair*Key Intervention v. Control	0.68	0.83	0.82	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-1.96	0.86	-2.28	0.37

Note: Overall model fit, F (19, 707) = 9.04, $p < .0001$, adjusted $r^2 = .17$

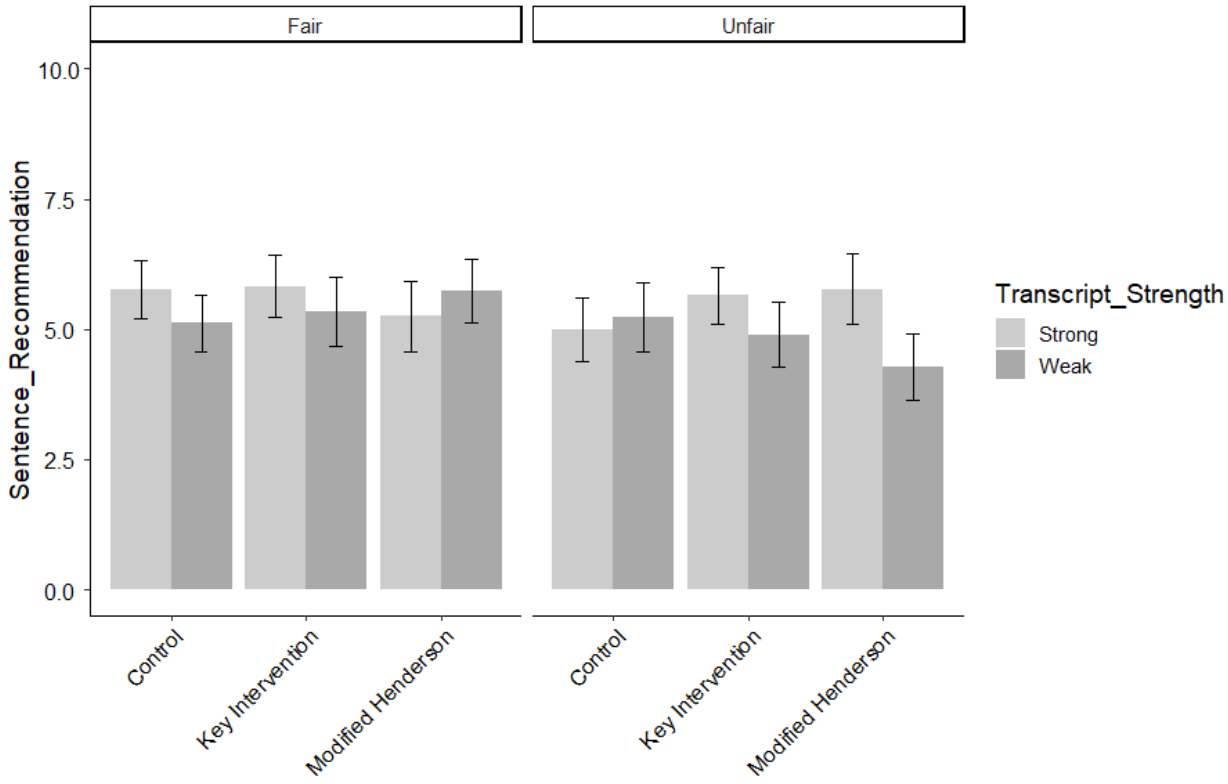


Figure K.6. Average SENTENCE LENGTH RECOMMENDATION as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.7.
CONFIDENCE IN WITNESS

	B	*SE*	*t*	*p*
(Intercept)	10.83	0.71	15.17	0.00
Transcript Strength (Weak)	-1.98	0.52	-3.80	0.00
Memory Test (Unfair)	-0.28	0.55	-0.52	1.00
Key Intervention v. Control	-0.43	0.51	-0.84	1.00
Key Intervention v. modified Henderson	-0.96	0.55	-1.76	1.00
Gender	-0.18	0.21	-0.83	1.00
Age	-0.01	0.01	-0.83	1.00
Ethnicity	0.03	0.08	0.37	1.00
Education	0.15	0.11	1.33	1.00
Political Orientation	-0.29	0.08	-3.73	0.00
Income	-0.08	0.10	-0.73	1.00
Numeracy (BNT-S)	-0.06	0.06	-0.93	1.00
Graph Literacy	-0.15	0.04	-4.18	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.25	0.75	0.33	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.61	0.71	0.86	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.45	0.76	1.90	0.92
Memory Test (Unfair)*Key Intervention v. Control	-0.37	0.73	-0.51	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.05	0.78	-0.07	1.00
Weak*Unfair*Key Intervention v. Control	0.55	1.03	0.53	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-1.37	1.07	-1.28	1.00

Note: Overall model fit, F (19, 707) = 6.45, $p < .0001$, adjusted $r^2 = .12$

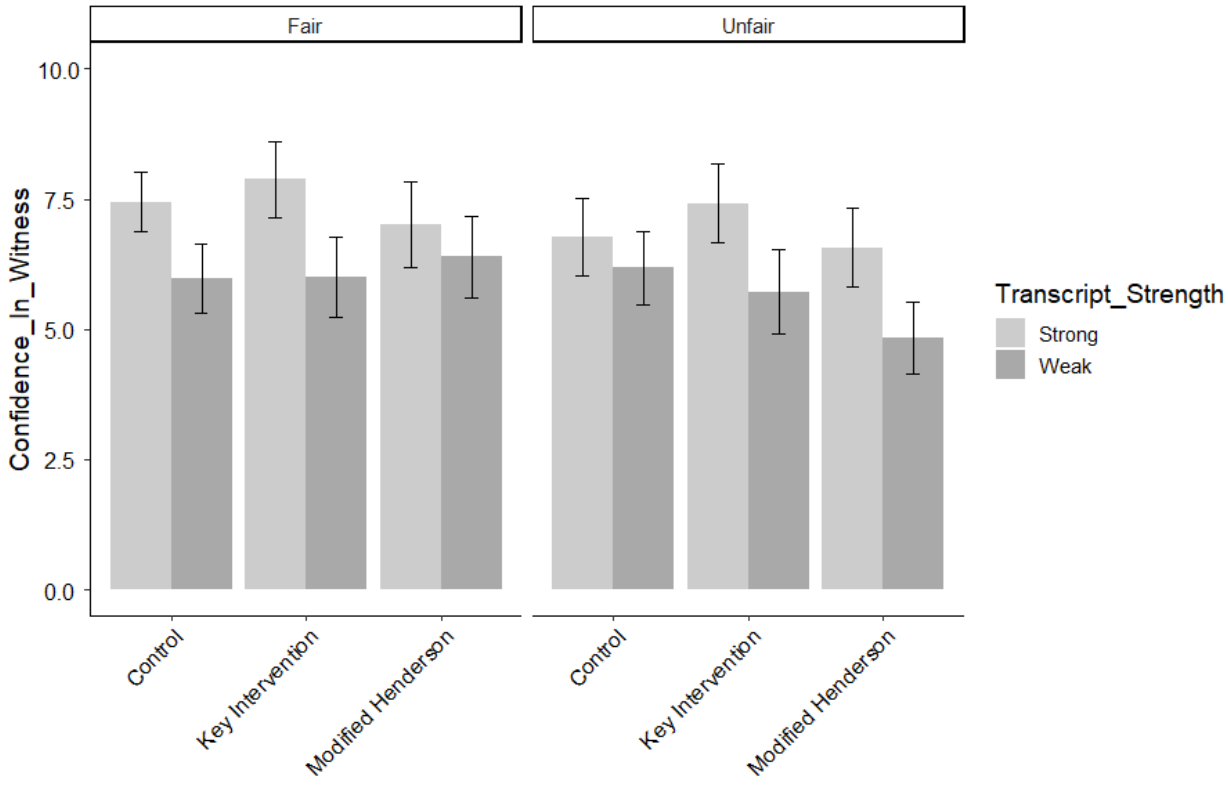


Figure K.7. Average CONFIDENCE IN WITNESS as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.8.
INITIAL CONFIDENCE INFLUENCE VERDICT

	B	*SE*	*t*	*p*
(Intercept)	9.07	0.65	13.94	0.00
Transcript Strength (Weak)	0.16	0.48	0.34	1.00
Memory Test (Unfair)	-0.76	0.49	-1.53	1.00
Key Intervention v. Control	-1.40	0.45	-3.07	0.04
Key Intervention v. modified Henderson	-0.85	0.49	-1.74	1.00
Gender	-0.04	0.19	-0.23	1.00
Age	0.01	0.01	1.50	1.00
Ethnicity	-0.15	0.07	-2.19	0.47
Education	0.16	0.10	1.54	1.00
Political Orientation	-0.21	0.07	-2.89	0.07
Income	-0.04	0.10	-0.40	1.00
Numeracy (BNT-S)	0.12	0.06	1.96	0.71
Graph Literacy	-0.11	0.03	-3.42	0.01
Transcript Strength (Weak)*Memory Test (Unfair)	0.19	0.69	0.27	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.74	0.65	1.14	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.74	0.69	1.08	1.00
Memory Test (Unfair)*Key Intervention v. Control	1.34	0.66	2.02	0.65
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.17	0.71	-0.25	1.00
Weak*Unfair*Key Intervention v. Control	-0.76	0.95	-0.80	1.00
Weak*Unfair*Key Intervention v. modified Henderson	0.11	0.99	0.11	1.00

Note: Overall model fit, F (19, 624) = 3.86, $p < .0001$, adjusted $r^2 = .08$

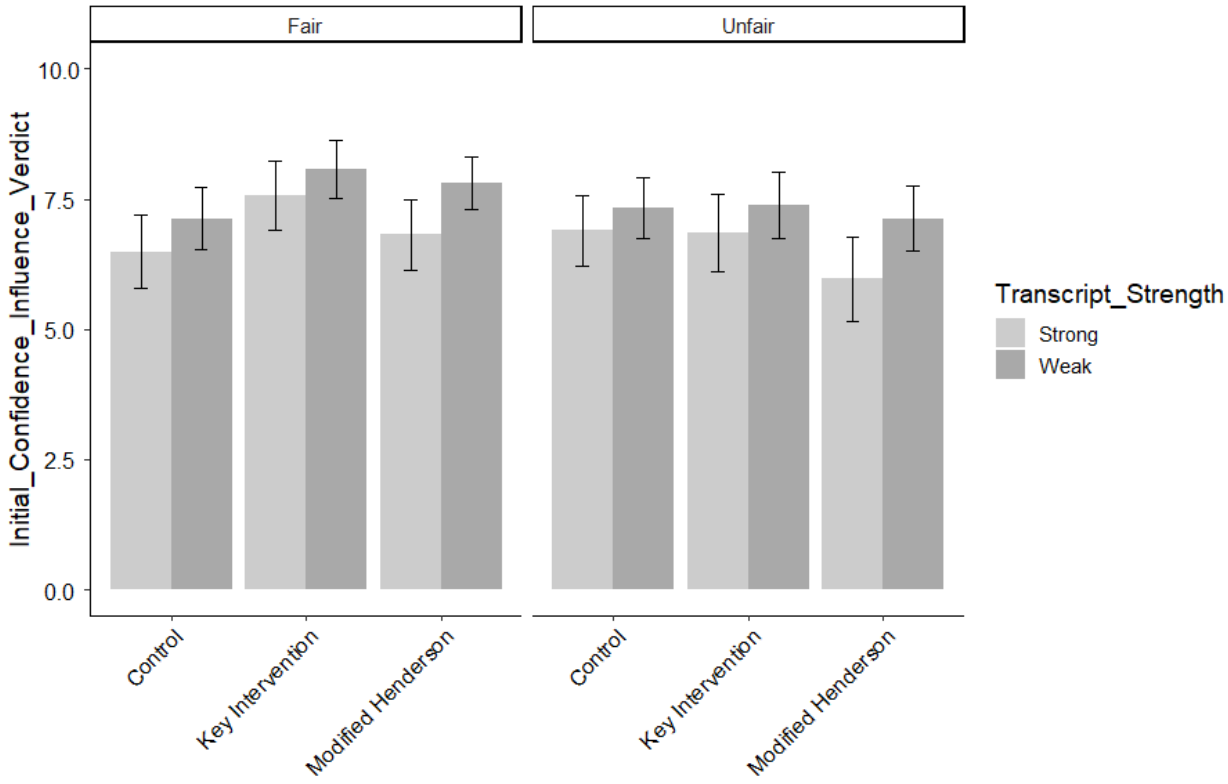


Figure K.8. Average INITIAL CONFIDENCE INFLUENCED VERDICT as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.9.
COURTROOM CONFIDENCE INFLUENCE VERDICT

	B	*SE*	*t*	*p*
(Intercept)	10.19	0.68	14.95	0.00
Transcript Strength (Weak)	-1.25	0.50	-2.53	0.20
Memory Test (Unfair)	-0.44	0.52	-0.85	1.00
Key Intervention v. Control	-0.19	0.48	-0.40	1.00
Key Intervention v. modified Henderson	0.00	0.53	0.00	1.00
Gender	-0.17	0.20	-0.87	1.00
Age	-0.01	0.01	-1.07	1.00
Ethnicity	0.02	0.07	0.24	1.00
Education	0.10	0.11	0.88	1.00
Political Orientation	-0.31	0.07	-4.19	0.00
Income	-0.02	0.10	-0.16	1.00
Numeracy (BNT-S)	-0.01	0.06	-0.13	1.00
Graph Literacy	-0.19	0.03	-5.44	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.23	0.71	0.32	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.16	0.68	1.71	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.80	0.74	1.09	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.54	0.70	0.77	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.70	0.75	-0.93	1.00
Weak*Unfair*Key Intervention v. Control	-0.52	0.98	-0.53	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.29	1.04	-0.28	1.00

Note: Overall model fit, F (19, 653) = 6.03, $p < .0001$, adjusted $r^2 = .12$

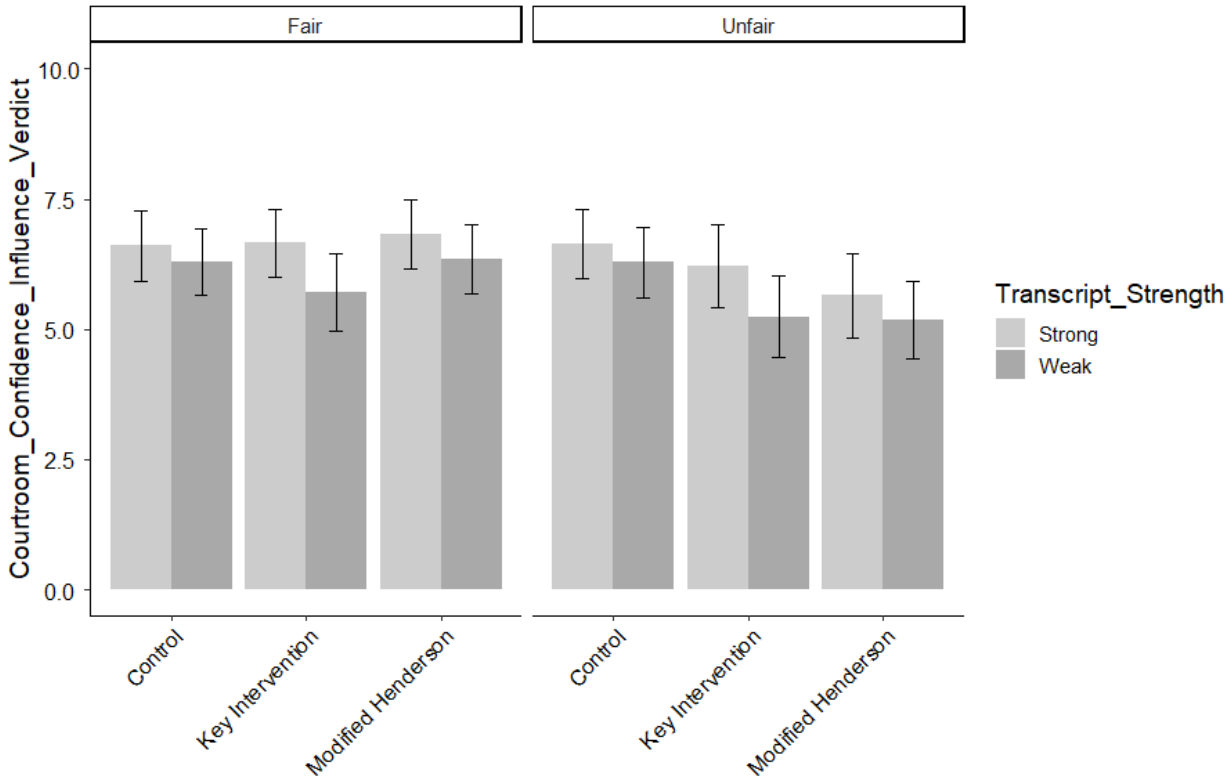


Figure K.9. Average COURTROOM CONFIDENCE INFLUENCED VERDICT as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.10.
GOOD LOOK

	B	*SE*	*t*	*p*
(Intercept)	9.75	0.64	15.13	0.00
Transcript Strength (Weak)	-1.76	0.47	-3.73	0.00
Memory Test (Unfair)	-0.61	0.49	-1.24	1.00
Key Intervention v. Control	-0.97	0.46	-2.11	0.46
Key Intervention v. modified Henderson	-1.55	0.49	-3.14	0.03
Gender	-0.16	0.19	-0.86	1.00
Age	0.00	0.01	-0.04	1.00
Ethnicity	-0.05	0.07	-0.67	1.00
Education	0.24	0.10	2.34	0.27
Political Orientation	-0.28	0.07	-3.97	0.00
Income	-0.09	0.09	-0.98	1.00
Numeracy (BNT-S)	-0.04	0.06	-0.71	1.00
Graph Literacy	-0.17	0.03	-5.18	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.66	0.68	0.97	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.18	0.64	1.84	0.79
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.75	0.69	2.56	0.16
Memory Test (Unfair)*Key Intervention v. Control	0.27	0.66	0.41	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.66	0.70	0.94	1.00
Weak*Unfair*Key Intervention v. Control	-0.58	0.93	-0.62	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-1.52	0.97	-1.57	1.00

Note: Overall model fit, F (19, 707) = 6.55, $p < .0001$, adjusted $r^2 = .13$

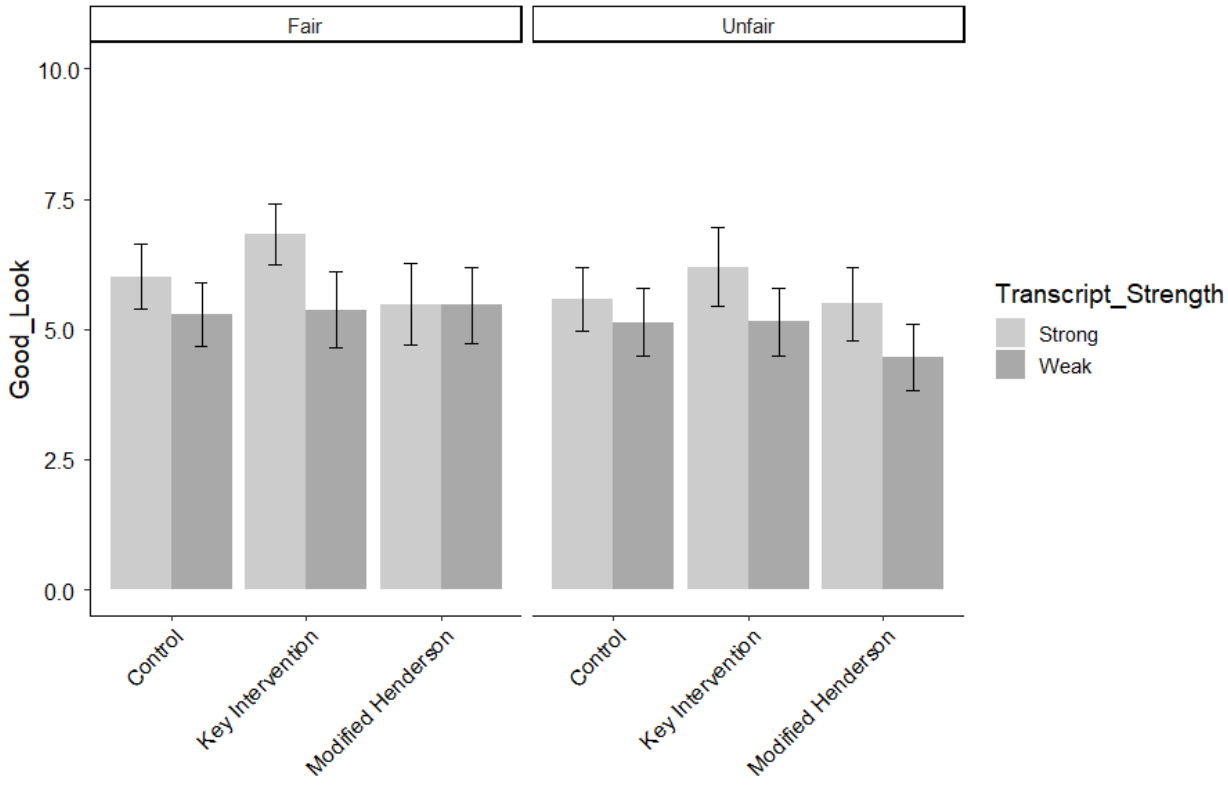


Figure K.10. Average GOOD LOOK as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.11.
ATTENTION

	B	*SE*	*t*	*p*
(Intercept)	9.21	0.63	14.65	0.00
Transcript Strength (Weak)	-1.55	0.46	-3.38	0.01
Memory Test (Unfair)	-0.47	0.48	-0.98	1.00
Key Intervention v. Control	-0.62	0.45	-1.38	1.00
Key Intervention v. modified Henderson	-1.12	0.48	-2.34	0.31
Gender	0.09	0.19	0.46	1.00
Age	0.00	0.01	0.56	1.00
Ethnicity	-0.06	0.07	-0.85	1.00
Education	0.23	0.10	2.34	0.31
Political Orientation	-0.24	0.07	-3.48	0.01
Income	-0.08	0.09	-0.83	1.00
Numeracy (BNT-S)	-0.07	0.06	-1.28	1.00
Graph Literacy	-0.18	0.03	-5.64	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.64	0.66	0.98	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.91	0.63	1.45	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.08	0.67	1.61	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.41	0.64	0.63	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.13	0.68	0.19	1.00
Weak*Unfair*Key Intervention v. Control	-0.86	0.91	-0.95	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.48	0.95	-0.51	1.00

Note: Overall model fit, F (19, 707) = 6.67, $p < .0001$, adjusted $r^2 = .13$

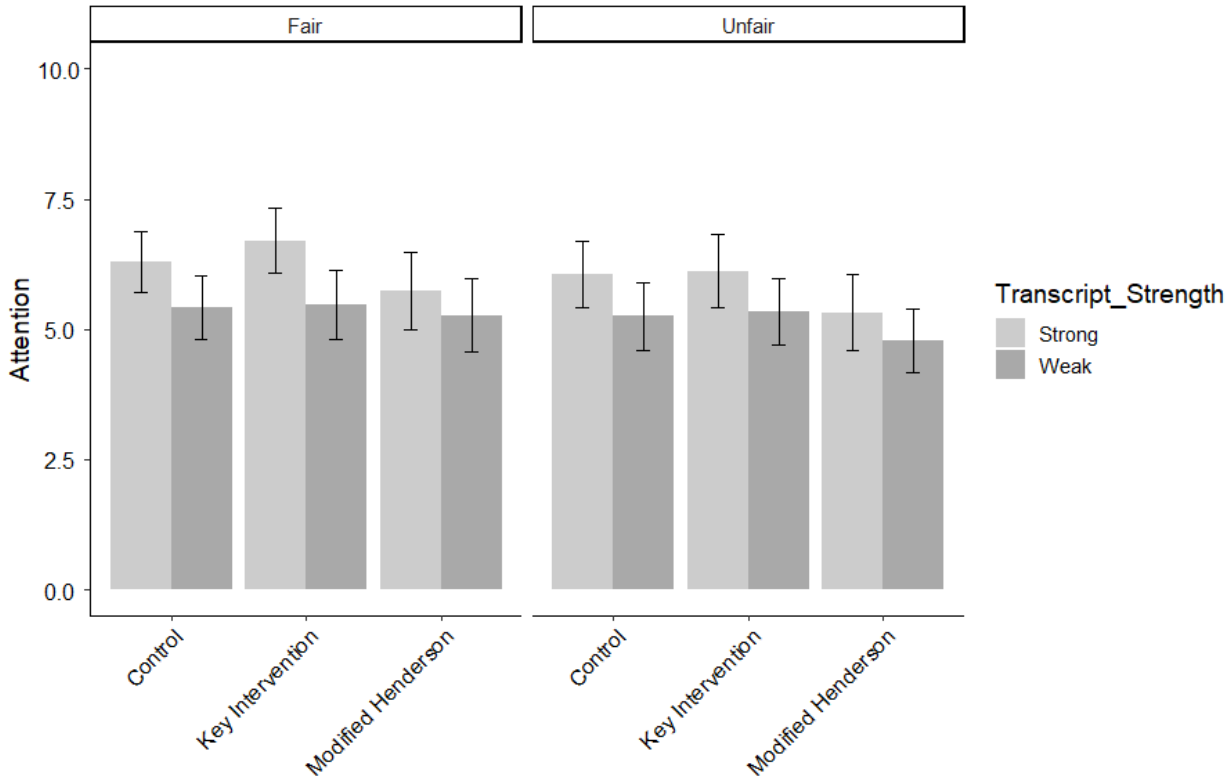


Figure K.11. Average ATTENTION as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.12.
GOOD BASIS

	B	*SE*	*t*	*p*
(Intercept)	10.02	0.62	16.13	0.00
Transcript Strength (Weak)	-1.49	0.45	-3.27	0.02
Memory Test (Unfair)	-0.42	0.47	-0.88	1.00
Key Intervention v. Control	-0.65	0.44	-1.48	1.00
Key Intervention v. modified Henderson	-1.00	0.48	-2.10	0.58
Gender	-0.13	0.18	-0.72	1.00
Age	0.00	0.01	-0.13	1.00
Ethnicity	-0.10	0.07	-1.53	1.00
Education	0.18	0.10	1.87	0.87
Political Orientation	-0.25	0.07	-3.68	0.01
Income	-0.07	0.09	-0.82	1.00
Numeracy (BNT-S)	-0.08	0.06	-1.50	1.00
Graph Literacy	-0.18	0.03	-5.98	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.39	0.65	0.60	1.00
Transcript Strength (Weak)*Key Intervention v. Control	1.04	0.62	1.68	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.36	0.66	2.06	0.60
Memory Test (Unfair)*Key Intervention v. Control	0.23	0.64	0.36	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.21	0.68	0.32	1.00
Weak*Unfair*Key Intervention v. Control	-0.51	0.90	-0.57	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.85	0.93	-0.91	1.00

Note: Overall model fit, F (19, 707) = 7.28, $p < .0001$, adjusted $r^2 = .14$

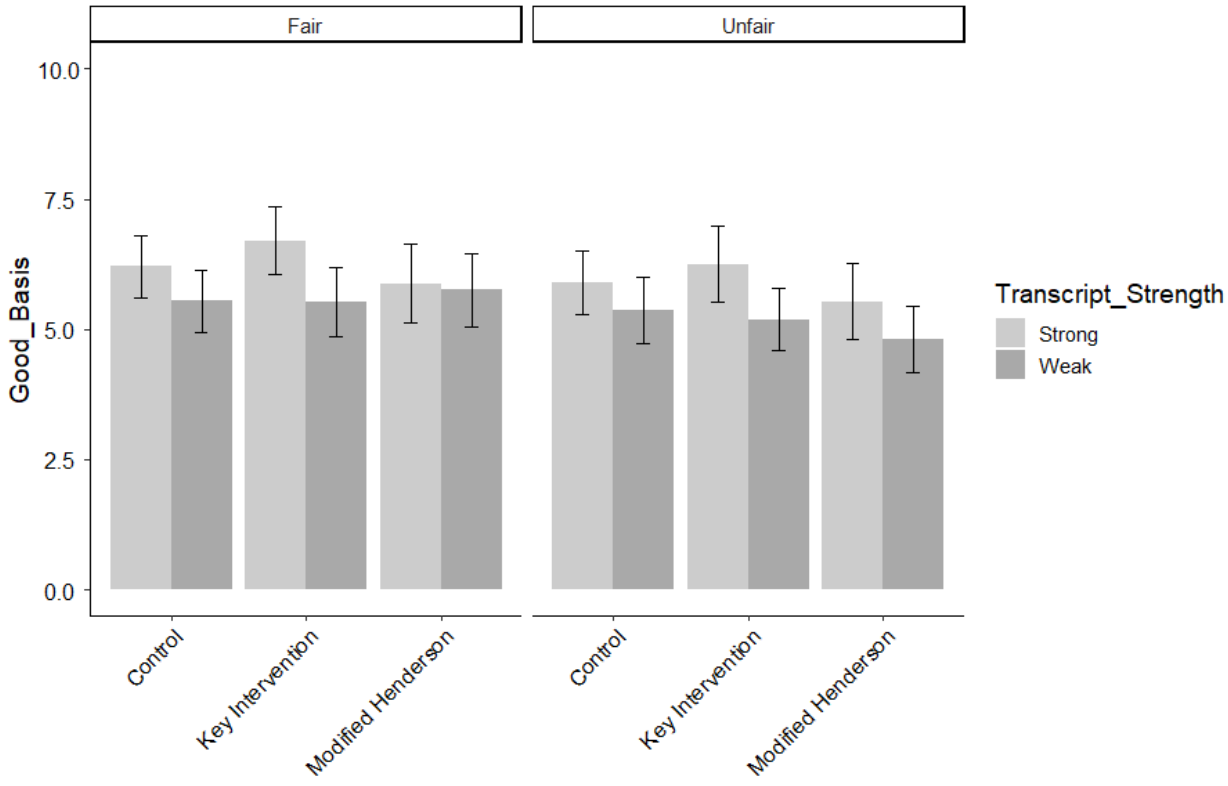


Figure K.12. Average GOOD BASIS as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.13.
WITNESS MEMORY

	B	*SE*	*t*	*p*
(Intercept)	9.04	0.59	15.34	0.00
Transcript Strength (Weak)	-1.54	0.43	-3.61	0.01
Memory Test (Unfair)	-0.26	0.45	-0.57	1.00
Key Intervention v. Control	-0.26	0.42	-0.62	1.00
Key Intervention v. modified Henderson	-1.15	0.45	-2.56	0.15
Gender	-0.01	0.17	-0.07	1.00
Age	0.00	0.01	-0.05	1.00
Ethnicity	-0.03	0.06	-0.46	1.00
Education	0.24	0.09	2.65	0.13
Political Orientation	-0.20	0.06	-3.17	0.03
Income	-0.09	0.09	-1.00	1.00
Numeracy (BNT-S)	-0.09	0.05	-1.73	1.00
Graph Literacy	-0.16	0.03	-5.46	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.56	0.62	0.91	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.50	0.59	0.85	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	1.66	0.62	2.66	0.13
Memory Test (Unfair)*Key Intervention v. Control	-0.15	0.60	-0.25	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.48	0.64	0.75	1.00
Weak*Unfair*Key Intervention v. Control	-0.23	0.85	-0.27	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-1.54	0.88	-1.74	1.00

Note: Overall model fit, F (19, 705) = 7.27, $p < .0001$, adjusted $r^2 = .14$

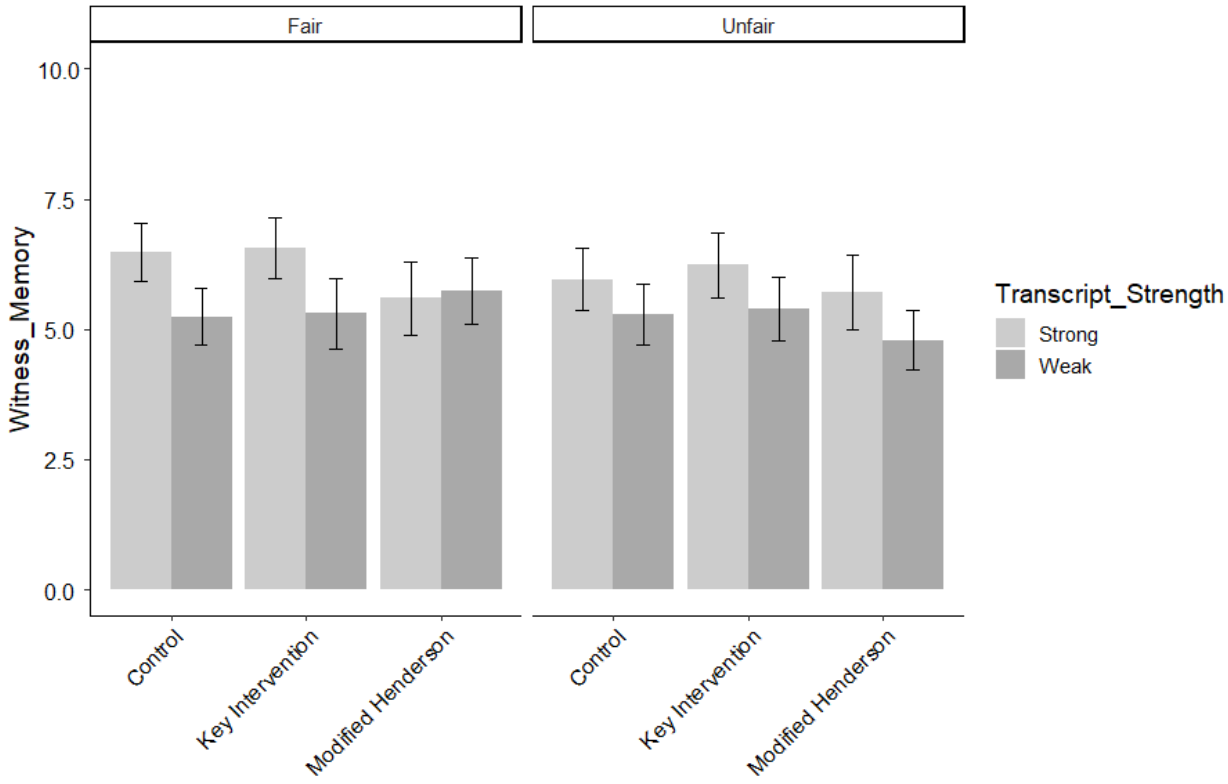


Figure K.13. Average WITNESS MEMORY as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.14.
EYEWITNESS ACCURACY GENERAL

	B	*SE*	*t*	*p*
(Intercept)	8.53	0.42	20.13	0.00
Transcript Strength (Weak)	-0.39	0.31	-1.27	1.00
Memory Test (Unfair)	0.00	0.32	0.00	1.00
Key Intervention v. Control	-0.26	0.30	-0.87	1.00
Key Intervention v. modified Henderson	-0.53	0.32	-1.64	1.00
Gender	-0.07	0.12	-0.53	1.00
Age	0.00	0.01	0.25	1.00
Ethnicity	0.03	0.04	0.65	1.00
Education	0.17	0.07	2.63	0.15
Political Orientation	-0.24	0.05	-5.35	0.00
Income	-0.01	0.06	-0.16	1.00
Numeracy (BNT-S)	-0.06	0.04	-1.56	1.00
Graph Literacy	-0.12	0.02	-5.90	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	-0.07	0.44	-0.15	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.02	0.42	0.04	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.28	0.45	0.63	1.00
Memory Test (Unfair)*Key Intervention v. Control	-0.34	0.43	-0.77	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.21	0.46	0.46	1.00
Weak*Unfair*Key Intervention v. Control	0.72	0.61	1.18	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.48	0.64	-0.76	1.00

Note: Overall model fit, F (19, 707) = 8.18, $p < .0001$, adjusted $r^2 = .16$

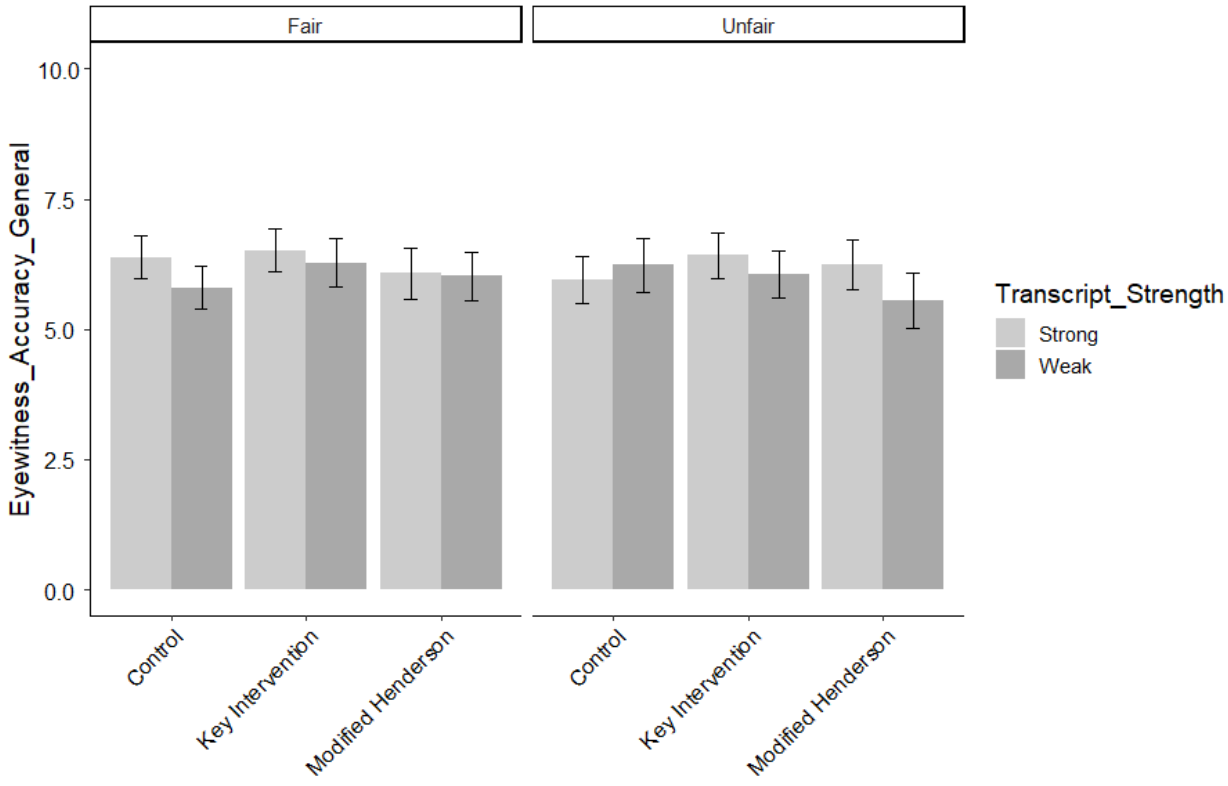


Figure K.14. Average EYEWITNESS ACCURACY GENERAL as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.15.
CONFIDENCE INDICATOR OF ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	8.22	0.49	16.71	0.00
Transcript Strength (Weak)	-0.27	0.36	-0.75	1.00
Memory Test (Unfair)	-0.36	0.37	-0.97	1.00
Key Intervention v. Control	-0.45	0.35	-1.29	1.00
Key Intervention v. modified Henderson	-0.50	0.38	-1.32	1.00
Gender	0.22	0.15	1.52	1.00
Age	0.01	0.01	1.73	1.00
Ethnicity	0.03	0.05	0.53	1.00
Education	0.14	0.08	1.85	1.00
Political Orientation	-0.27	0.05	-5.03	0.00
Income	0.01	0.07	0.14	1.00
Numeracy (BNT-S)	-0.05	0.04	-1.09	1.00
Graph Literacy	-0.12	0.02	-4.81	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.15	0.52	0.29	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.38	0.49	-0.77	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.00	0.52	0.01	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.08	0.50	0.15	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.51	0.53	0.95	1.00
Weak*Unfair*Key Intervention v. Control	0.72	0.71	1.01	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.32	0.74	-0.43	1.00

Note: Overall model fit, F (19, 705) = 6.00, $p < .0001$, adjusted $r^2 = .12$

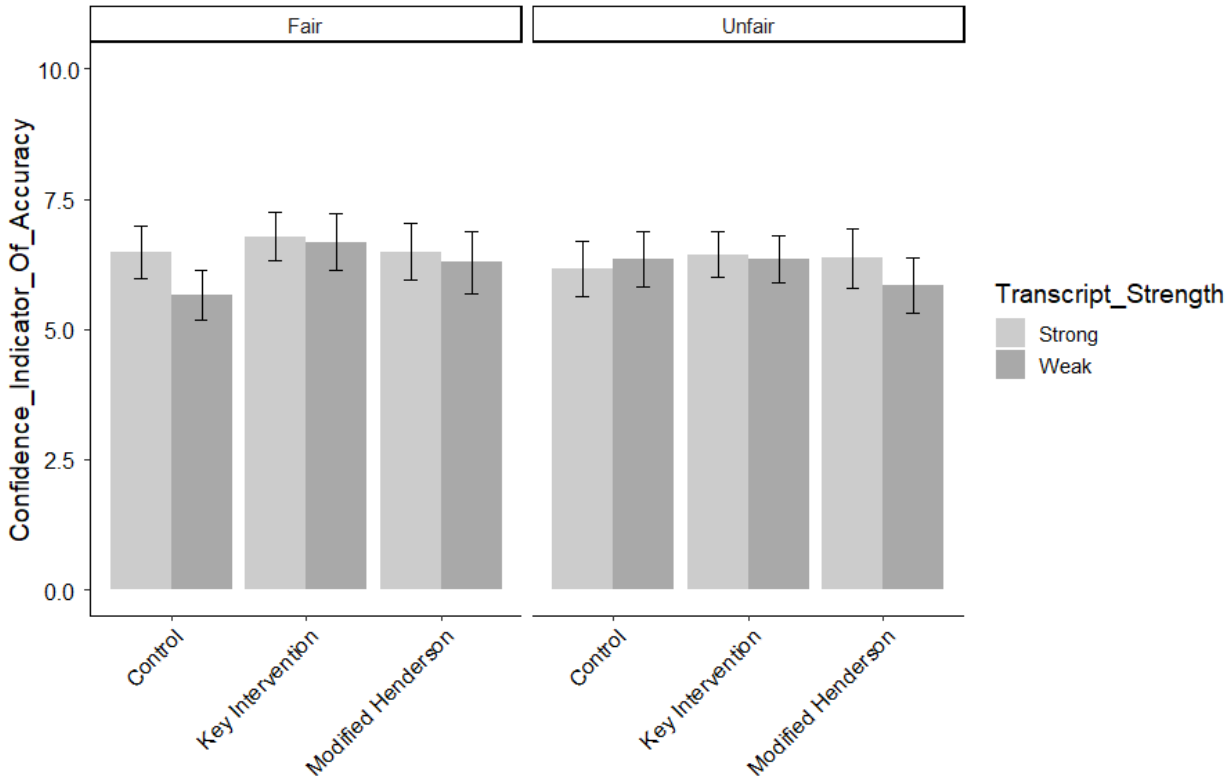


Figure K.15. Average CONFIDENCE INDICATOR OF ACCURACY as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.16.
CONFIDENCE INFLATION OCCURS

	B	*SE*	*t*	*p*
(Intercept)	7.94	0.65	12.22	0.00
Transcript Strength (Weak)	-0.29	0.47	-0.62	1.00
Memory Test (Unfair)	-0.19	0.50	-0.39	1.00
Key Intervention v. Control	0.14	0.46	0.30	1.00
Key Intervention v. modified Henderson	-0.38	0.50	-0.76	1.00
Gender	-0.14	0.19	-0.75	1.00
Age	0.01	0.01	0.81	1.00
Ethnicity	-0.01	0.07	-0.08	1.00
Education	0.28	0.10	2.79	0.10
Political Orientation	-0.14	0.07	-1.95	0.88
Income	-0.11	0.09	-1.20	1.00
Numeracy (BNT-S)	-0.01	0.06	-0.22	1.00
Graph Literacy	-0.16	0.03	-5.03	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.50	0.68	0.73	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.17	0.65	0.26	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.98	0.69	1.42	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.44	0.67	0.66	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.60	0.71	0.85	1.00
Weak*Unfair*Key Intervention v. Control	-0.10	0.94	-0.10	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.89	0.98	-0.91	1.00

Note: Overall model fit, F (19, 706) = 3.90, $p < .0001$, adjusted $r^2 = .07$

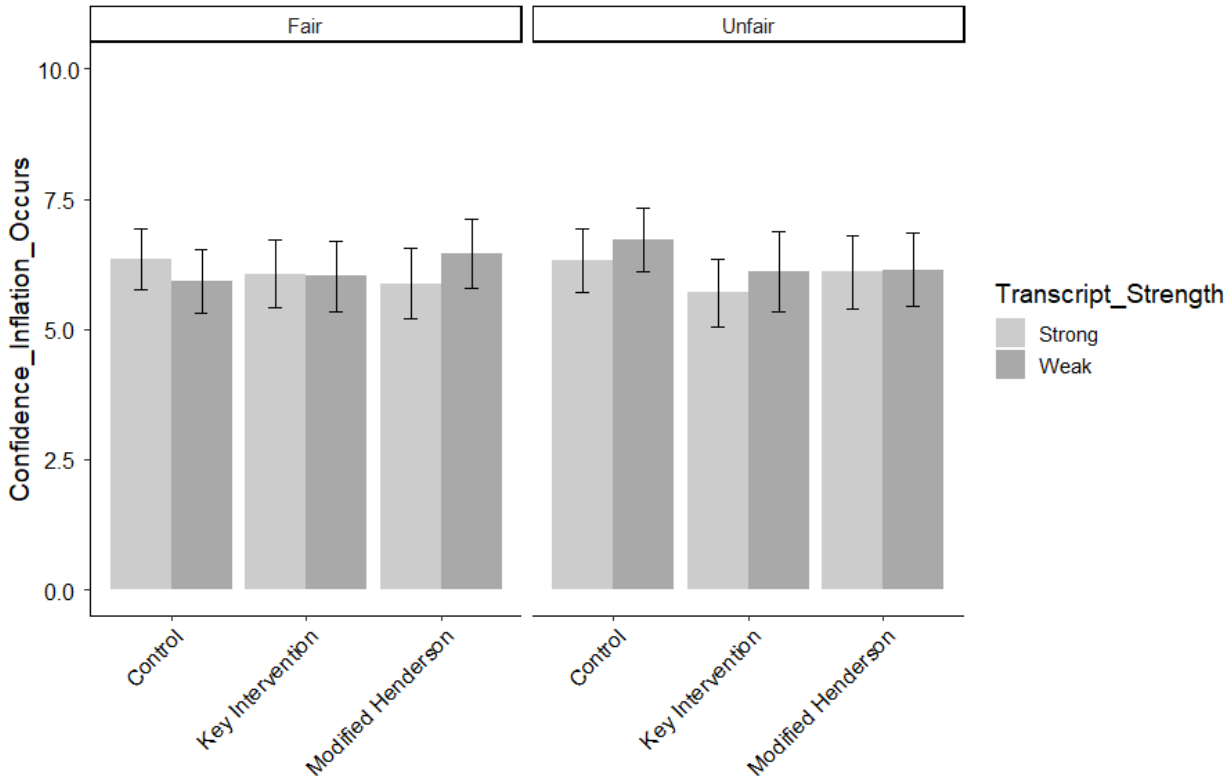


Figure K.16. Average CONFIDENCE INFLATION OCCURS as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.17.
CONFIDENCE INFLATION EQUALS ACCURACY

	B	*SE*	*t*	*p*
(Intercept)	7.49	0.58	12.93	0.00
Transcript Strength (Weak)	0.54	0.42	1.28	1.00
Memory Test (Unfair)	0.33	0.44	0.75	1.00
Key Intervention v. Control	0.81	0.41	1.97	0.79
Key Intervention v. modified Henderson	-0.01	0.44	-0.02	1.00
Gender	-0.10	0.17	-0.61	1.00
Age	0.00	0.01	-0.09	1.00
Ethnicity	-0.01	0.06	-0.15	1.00
Education	0.25	0.09	2.72	0.11
Political Orientation	-0.21	0.06	-3.42	0.01
Income	0.03	0.08	0.42	1.00
Numeracy (BNT-S)	-0.08	0.05	-1.61	1.00
Graph Literacy	-0.18	0.03	-6.08	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	-0.52	0.61	-0.85	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.76	0.58	-1.32	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.23	0.62	0.37	1.00
Memory Test (Unfair)*Key Intervention v. Control	-0.76	0.59	-1.28	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.04	0.63	-0.07	1.00
Weak*Unfair*Key Intervention v. Control	1.53	0.83	1.84	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.20	0.87	-0.23	1.00

Note: Overall model fit, F (19, 706) = 7.22, $p < .0001$, adjusted $r^2 = .14$

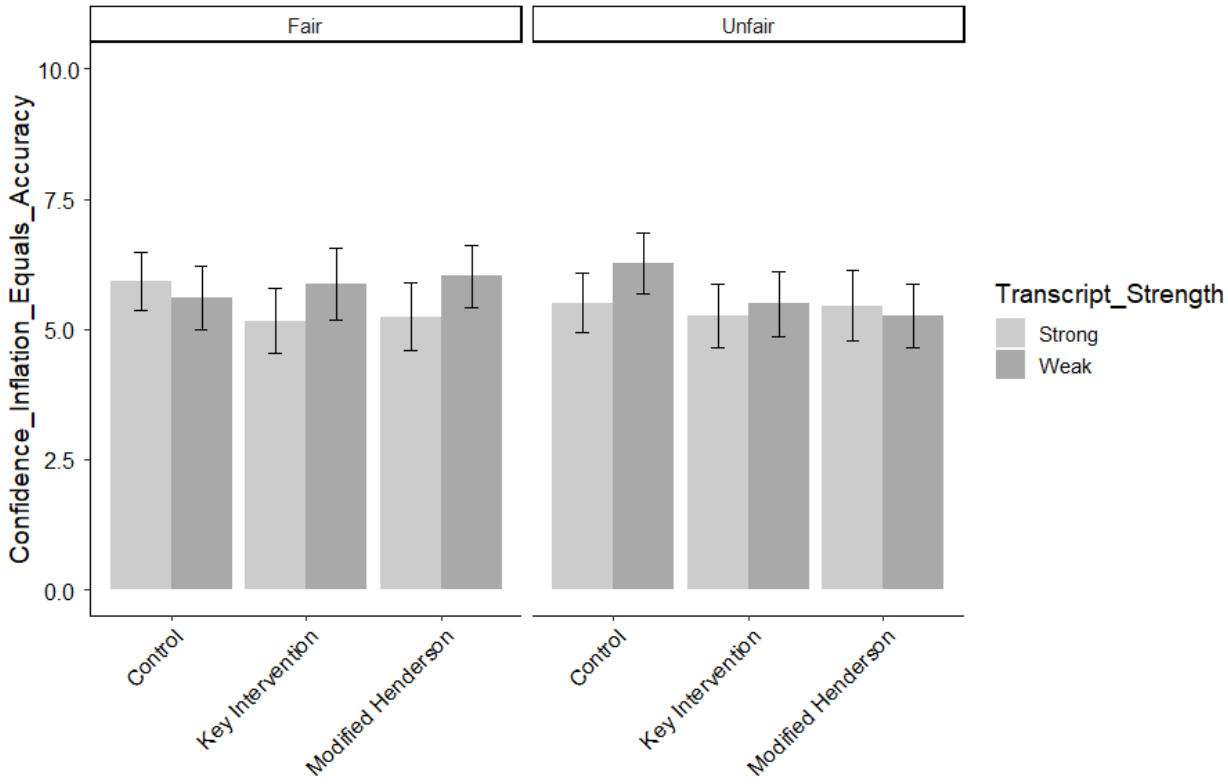


Figure K.17. Average CONFIDENCE INFLATION EQUALS ACCURACY as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.18.
COMPREHENSION CHECK QUESTIONS

	B	*SE*	*t*	*p*
(Intercept)	2.89	0.12	24.09	0.00
Transcript Strength (Weak)	-0.14	0.09	-1.61	1.00
Memory Test (Unfair)	-0.01	0.09	-0.12	1.00
Key Intervention v. Control	-0.10	0.09	-1.20	1.00
Key Intervention v. modified Henderson	-0.07	0.09	-0.76	1.00
Gender	0.10	0.04	2.73	0.10
Age	0.00	0.00	1.62	1.00
Ethnicity	-0.02	0.01	-1.31	1.00
Education	-0.07	0.02	-3.75	0.00
Political Orientation	0.04	0.01	2.83	0.08
Income	0.02	0.02	1.41	1.00
Numeracy (BNT-S)	0.03	0.01	2.58	0.15
Graph Literacy	0.06	0.01	9.70	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.01	0.13	0.04	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.18	0.12	1.52	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.13	0.13	1.06	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.15	0.12	1.23	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.00	0.13	-0.02	1.00
Weak*Unfair*Key Intervention v. Control	-0.05	0.17	-0.28	1.00
Weak*Unfair*Key Intervention v. modified Henderson	0.04	0.18	0.22	1.00

Note: Overall model fit, F (19, 707) = 15.34, $p < .0001$, adjusted $r^2 = .27$

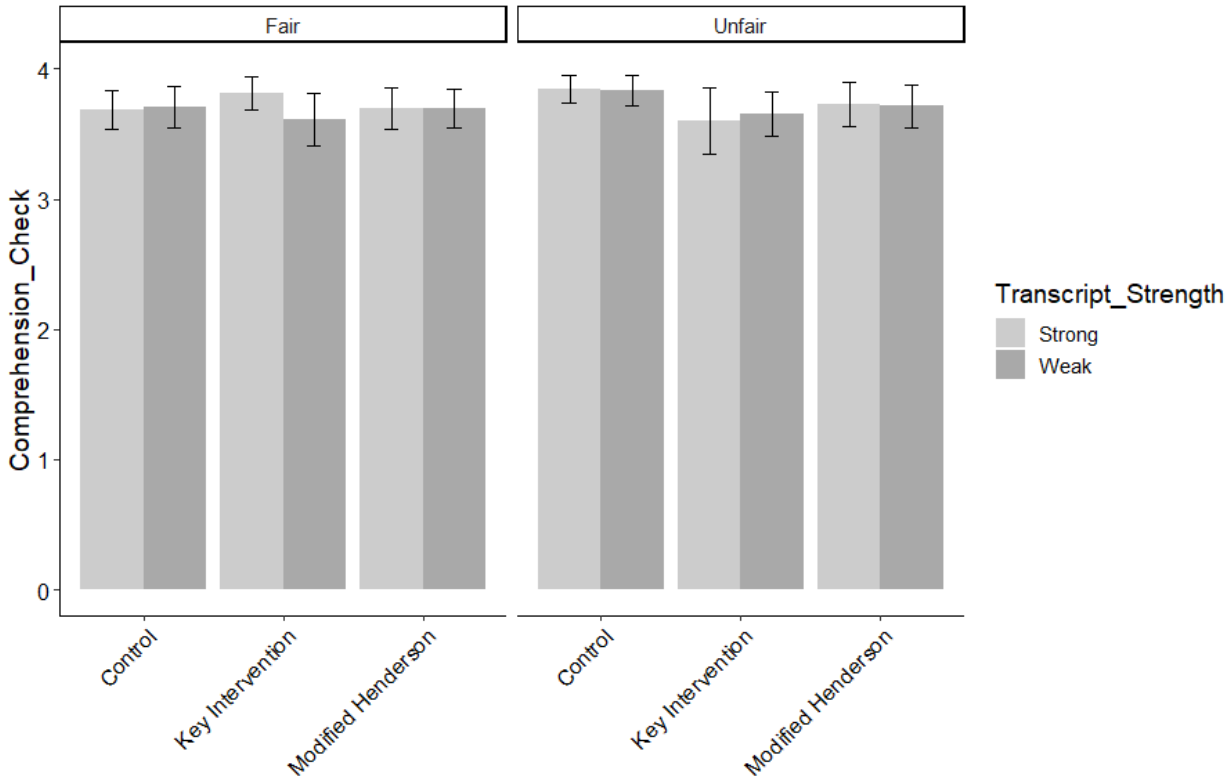


Figure K.18. Average COMPREHENSION CHECK QUESTIONS as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.19.
INITIAL CONFIDENCE PERCENTAGE

	B	*SE*	*t*	*p*
(Intercept)	9.89	0.28	34.75	0.00
Transcript Strength (Weak)	-2.94	0.21	-14.07	0.00
Memory Test (Unfair)	0.05	0.22	0.24	1.00
Key Intervention v. Control	-0.23	0.20	-1.17	1.00
Key Intervention v. modified Henderson	0.17	0.21	0.80	1.00
Gender	-0.02	0.08	-0.23	1.00
Age	-0.01	0.00	-2.50	0.20
Ethnicity	-0.06	0.03	-1.96	0.76
Education	0.12	0.04	2.60	0.16
Political Orientation	-0.04	0.03	-1.24	1.00
Income	-0.06	0.04	-1.42	1.00
Numeracy (BNT-S)	-0.02	0.03	-0.78	1.00
Graph Literacy	-0.06	0.01	-4.20	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	-0.42	0.30	-1.38	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.10	0.29	-0.35	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.33	0.30	-1.10	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.25	0.29	0.87	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.28	0.31	-0.90	1.00
Weak*Unfair*Key Intervention v. Control	0.42	0.42	1.01	1.00
Weak*Unfair*Key Intervention v. modified Henderson	0.47	0.43	1.09	1.00

Note: Overall model fit, F (19, 625) = 75.17, $p < .0001$, adjusted $r^2 = .69$

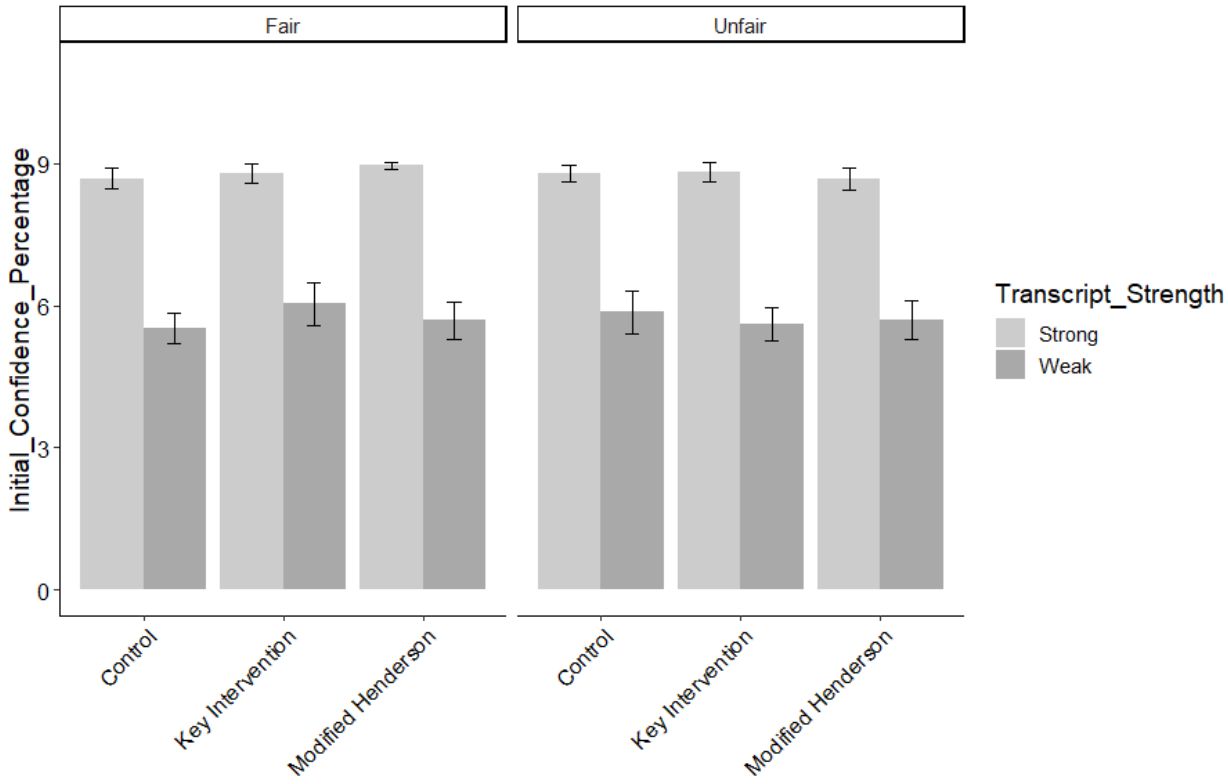


Figure K.19. Average INITIAL CONFIDENCE PERCENTAGE as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.20.
COURTROOM CONFIDENCE PERCENTAGE

	B	*SE*	*t*	*p*
(Intercept)	7.84	0.19	40.74	0.00
Transcript Strength (Weak)	0.07	0.14	0.48	1.00
Memory Test (Unfair)	0.03	0.15	0.23	1.00
Key Intervention v. Control	-0.11	0.14	-0.85	1.00
Key Intervention v. modified Henderson	0.22	0.15	1.44	1.00
Gender	0.16	0.06	2.74	0.11
Age	0.00	0.00	1.30	1.00
Ethnicity	-0.05	0.02	-2.29	0.36
Education	-0.05	0.03	-1.47	1.00
Political Orientation	0.01	0.02	0.58	1.00
Income	0.01	0.03	0.40	1.00
Numeracy (BNT-S)	0.05	0.02	2.61	0.16
Graph Literacy	0.05	0.01	5.26	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	-0.20	0.20	-1.01	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.03	0.19	0.15	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	-0.25	0.21	-1.19	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.11	0.20	0.55	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.26	0.21	-1.23	1.00
Weak*Unfair*Key Intervention v. Control	0.04	0.28	0.13	1.00
Weak*Unfair*Key Intervention v. modified Henderson	0.44	0.29	1.52	1.00

Note: Overall model fit, F (19, 653) = 6.12, $p < .0001$, adjusted $r^2 = .13$

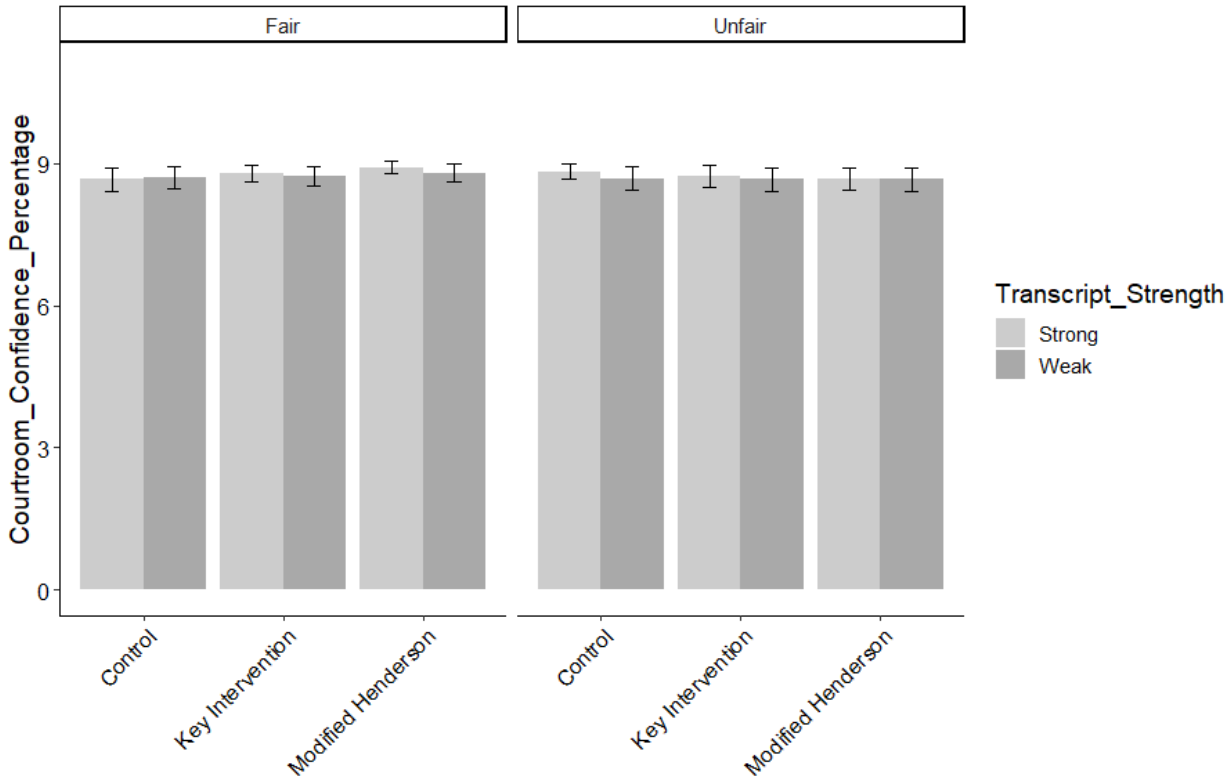


Figure K.20. Average COURTROOM CONFIDENCE PERCENTAGE as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.21.
CONFIDENCE RANKING

	B	*SE*	*t*	*p*
(Intercept)	2.20	0.14	15.29	0.00
Transcript Strength (Weak)	-0.65	0.10	-6.36	0.00
Memory Test (Unfair)	0.03	0.11	0.30	1.00
Key Intervention v. Control	0.01	0.11	0.11	1.00
Key Intervention v. modified Henderson	-0.05	0.11	-0.42	1.00
Gender	-0.04	0.04	-0.89	1.00
Age	0.00	0.00	-1.25	1.00
Ethnicity	0.02	0.02	1.23	1.00
Education	0.00	0.02	0.14	1.00
Political Orientation	0.03	0.02	1.66	1.00
Income	0.02	0.02	0.90	1.00
Numeracy (BNT-S)	-0.03	0.01	-2.66	0.15
Graph Literacy	-0.01	0.01	-2.04	0.71
Transcript Strength (Weak)*Memory Test (Unfair)	0.00	0.15	0.00	1.00
Transcript Strength (Weak)*Key Intervention v. Control	-0.05	0.15	-0.33	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.03	0.15	0.20	1.00
Memory Test (Unfair)*Key Intervention v. Control	-0.04	0.15	-0.25	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	-0.08	0.16	-0.50	1.00
Weak*Unfair*Key Intervention v. Control	0.03	0.21	0.14	1.00
Weak*Unfair*Key Intervention v. modified Henderson	0.07	0.22	0.31	1.00

Note: Overall model fit, F (19, 526) = 13.91, $p < .0001$, adjusted $r^2 = .31$

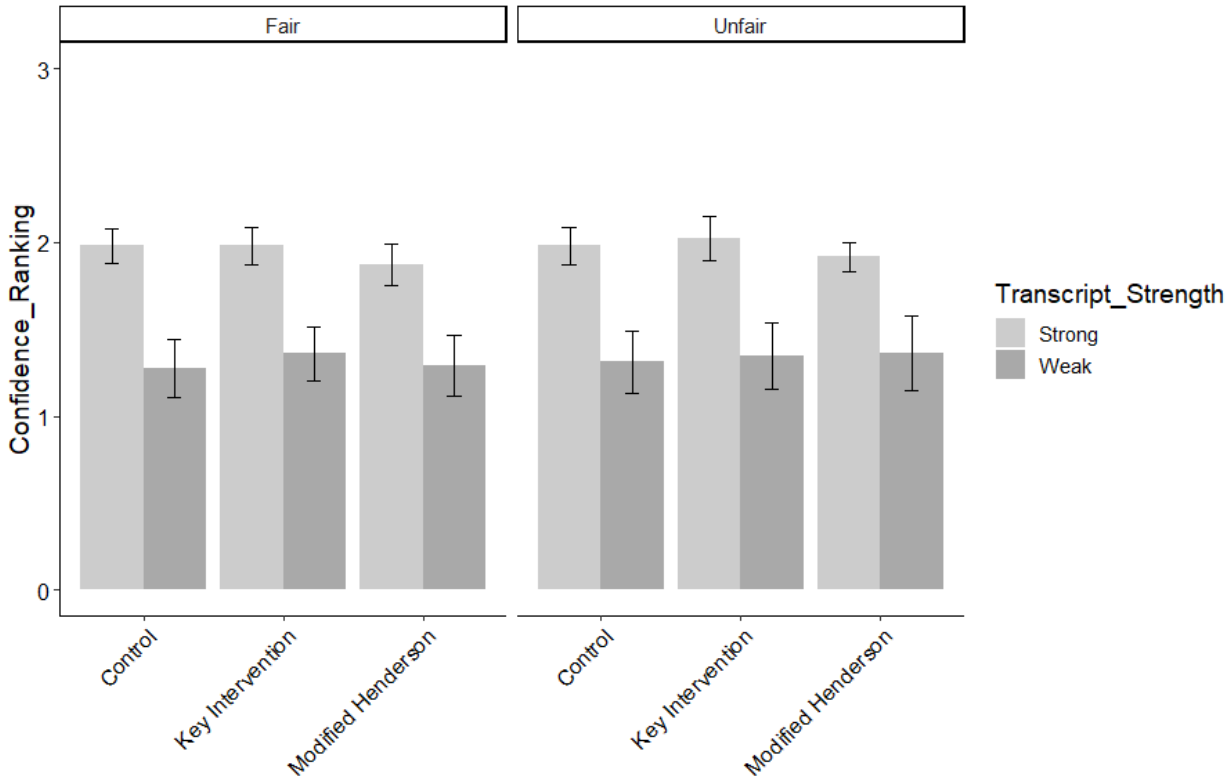


Figure K.21. Average CONFIDENCE RANKING as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.22.
USABILITY

	B	*SE*	*t*	*p*
(Intercept)	6.28	0.38	16.75	0.00
Transcript Strength (Weak)	-0.59	0.27	-2.15	0.41
Memory Test (Unfair)	-0.40	0.29	-1.41	0.95
Key Intervention v. Control	-0.57	0.27	-2.15	0.41
Key Intervention v. modified Henderson	-0.34	0.29	-1.20	1.00
Gender	0.31	0.11	2.78	0.09
Age	0.02	0.00	3.97	0.00
Ethnicity	-0.14	0.04	-3.50	0.01
Education	-0.28	0.06	-4.76	0.00
Political Orientation	0.09	0.04	2.34	0.28
Income	0.05	0.05	1.01	1.00
Numeracy (BNT-S)	0.07	0.03	2.06	0.41
Graph Literacy	0.20	0.02	10.48	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.84	0.39	2.13	0.41
Transcript Strength (Weak)*Key Intervention v. Control	0.63	0.37	1.67	0.76
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.35	0.40	0.87	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.64	0.38	1.67	0.76
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.27	0.41	0.66	1.00
Weak*Unfair*Key Intervention v. Control	-1.11	0.54	-2.06	0.41
Weak*Unfair*Key Intervention v. modified Henderson	-0.51	0.56	-0.90	1.00

Note: Overall model fit, F (19, 707) = 18.63, $p < .0001$, adjusted $r^2 = .32$

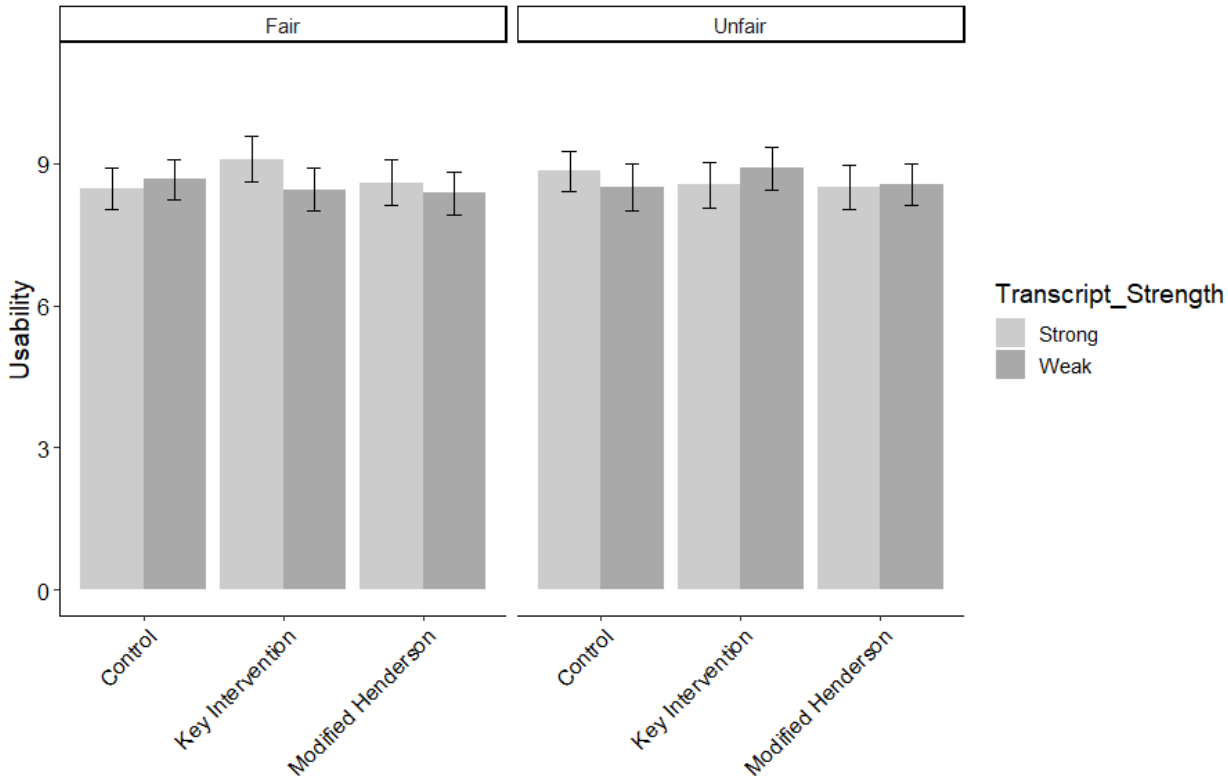


Figure K.22. Average USABILITY as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.

Table K.23.
WORKLOAD

	B	*SE*	*t*	*p*
(Intercept)	8.59	0.38	22.78	0.00
Transcript Strength (Weak)	-0.32	0.28	-1.15	1.00
Memory Test (Unfair)	-0.35	0.29	-1.21	1.00
Key Intervention v. Control	-0.24	0.27	-0.90	1.00
Key Intervention v. modified Henderson	-0.26	0.29	-0.89	1.00
Gender	-0.07	0.11	-0.67	1.00
Age	-0.01	0.00	-3.02	0.04
Ethnicity	0.14	0.04	3.51	0.01
Education	0.38	0.06	6.41	0.00
Political Orientation	-0.25	0.04	-6.12	0.00
Income	-0.10	0.05	-1.88	0.85
Numeracy (BNT-S)	-0.06	0.03	-1.79	0.97
Graph Literacy	-0.22	0.02	-11.50	0.00
Transcript Strength (Weak)*Memory Test (Unfair)	0.17	0.40	0.43	1.00
Transcript Strength (Weak)*Key Intervention v. Control	0.31	0.38	0.82	1.00
Transcript Strength (Weak)*Key Intervention v. modified Henderson	0.52	0.40	1.30	1.00
Memory Test (Unfair)*Key Intervention v. Control	0.11	0.39	0.28	1.00
Memory Test (Unfair)*Key Intervention v. modified Henderson	0.66	0.41	1.61	1.00
Weak*Unfair*Key Intervention v. Control	-0.18	0.54	-0.33	1.00
Weak*Unfair*Key Intervention v. modified Henderson	-0.92	0.57	-1.63	1.00

Note: Overall model fit, F (19, 707) = 23.63, $p < .0001$, adjusted $r^2 = .37$

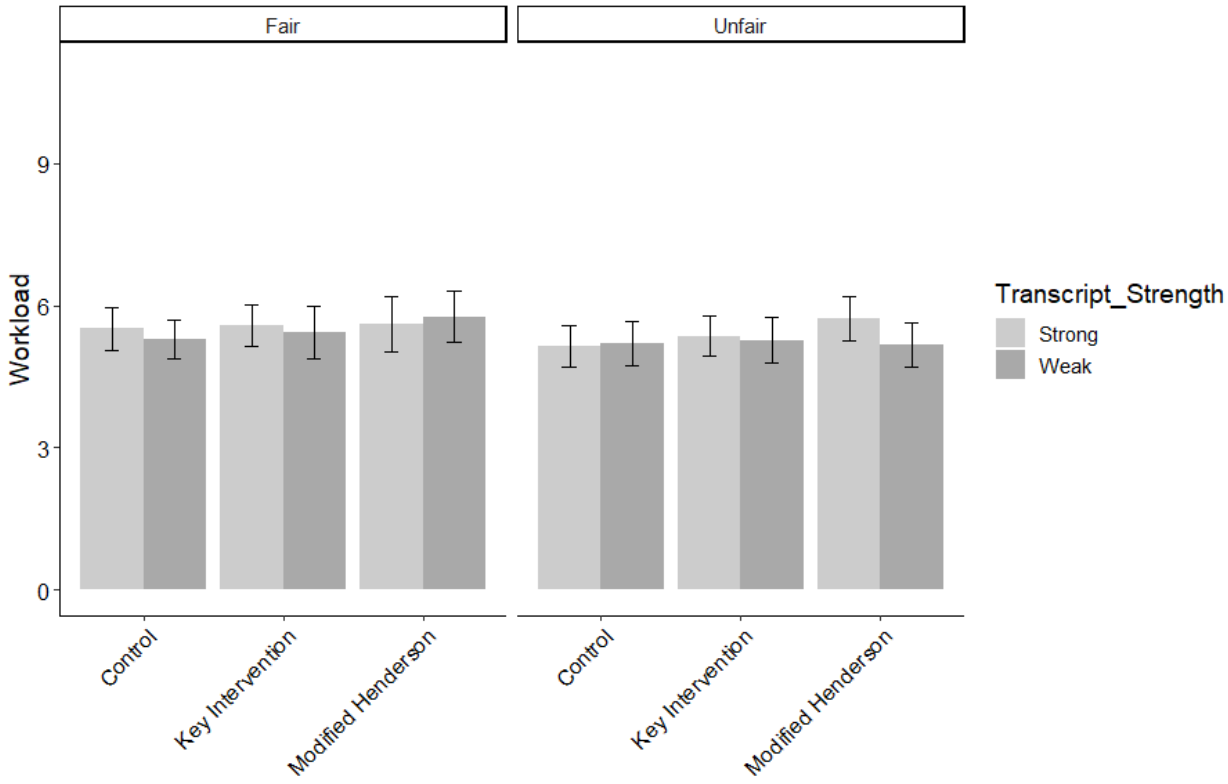


Figure K.23. Average WORKLOAD as a function of Transcript Strength, Intervention type, and Memory Test Fairness. Error bars represent 95% confidence intervals.