

## The flaw in Amendola and Wixted's conclusion on simultaneous versus sequential lineups

Gary L. Wells · Jennifer E. Dysart · Nancy K. Steblay

Published online: 18 January 2015

© Springer Science+Business Media Dordrecht 2015

### Abstract

*Objective* Our objective was to examine how Amendola and Wixted (A&W, 2014) arrived at their conclusion that eyewitness identifications of suspects from simultaneous lineups were supported better by corroborating evidence than were identifications from sequential lineups. Their cases came from a randomized field experiment by Wells et al. (2014).

*Methods* We gathered information from the A&W article, examined an earlier, more complete report by Amendola et al. (2013), and then confirmed our numbers with Amendola.

*Results* We discovered that the small subsample ( $n=52$ ) on which A&W's entire conclusion was based was unrepresentative of the larger set of cases ( $N=236$ ) in a way that was heavily biased in favor of the simultaneous lineup. Specifically, although the larger data set showed that simultaneous and sequential lineups produced the same rate of adjudicated guilt, their small subsample of 52 cases was highly imbalanced: Among the 30 sequential cases selected, 16 were drawn from the adjudicated-guilty set and 14 were from the not-prosecuted set; among the 22 simultaneous cases selected, 17 were drawn from the adjudicated-guilty set and a mere five were from the not-prosecuted set. This problem could not be known from the article itself.

*Conclusions* Because adjudicated guilty cases had more corroborating evidence than not-prosecuted cases and because simultaneous and sequential lineups produced equivalent rates of adjudicated guilty outcomes, the small sub-sample of 52 should have reflected this same equivalence. Instead, the sub-sample was stacked against the sequential and in favor of the simultaneous and A&W's conclusion is not warranted.

---

G. L. Wells (✉)

Department of Psychology, Iowa State University, Ames, IA 50021, USA  
e-mail: glwells@iastate.edu

J. E. Dysart

John Jay College of Criminal Justice, New York, NY, USA

N. K. Steblay

Augsburg College, Minneapolis, MN, USA

**Keywords** Sequential lineups · Eyewitness field studies · Eyewitness identification

It was with great interest that we read Amendola and Wixted's (2014) conclusion that the simultaneous lineup seemed to outperform the sequential lineup when the strength of evidence in cases where the eyewitness identified the suspect was examined. Amendola and Wixted (A&W) analyzed a small subset of the data from an original field experiment of 494 lineups (Wells et al. 2014) that tested double-blind sequential versus double-blind simultaneous lineup procedures. The field test used computer-based random assignment of each lineup type and it is important to note upfront that the random assignment was successful in the original field study as a means to control extraneous factors between the two tested lineup procedures (see Wells et al. for details).

Amendola obtained the lineup data and lineup materials from the field experiment from the current authors (Wells et al. 2014) so that she could collect additional data on case outcomes and strength of corroborating evidence that were not available at the time of our original experiment. Amendola had a very difficult task. Unfortunately, the conclusion that A&W reached is flawed and we will describe the nature of the error that leads to their unsubstantiated conclusion.

It is important to keep in mind that our field experiment had 494 lineups, yet A&W base their entire conclusion on only 52 cases. This small sample is certainly a concern, in that small cell sizes may lead to unstable and misleading results, but the fundamental flaw that undermines their conclusion is not the small size of the final sample per se. Their error is much more definitive.

A&W wished to compare sequential and simultaneous lineup procedures with respect to corroborating evidence in a small sub-sample of the Wells et al. dataset. The challenge for A&W was to make certain that the protection of random assignment achieved in the original study held firmly as they narrowed their sample to a much smaller subset. We show a serious failure to meet this requirement, namely that the final A&W sample of 52 is unrepresentative of the cases from which the sample was drawn and that the specific pattern of that unrepresentative sample biased the result strongly against the sequential procedure and led A&W toward an erroneous conclusion.

There were many legitimate reasons why A&W had to pare down the sample for purposes of analysis. For example, only 236 of the 494 cases had dispositions (i.e., A&W required a dichotomous categorization of lineups such that either the suspect in the lineup was adjudicated to be guilty or the case was not prosecuted). Other reasons for not including cases also existed (see A&W article for a list of these reasons). Also, the time and expense of having experts rate the evidence strength may well have necessitated using only a small subset as the focus of the final analysis. Moreover, A&W state that they attempted to engage in a stratified random sampling approach as they pared their analyses to the final 52. We do not doubt their effort or question their intent, although the article does not make clear exactly what strata they were using.

Whatever the details of the process by which A&W settled on the final 52 cases, the sample ended up being heavily biased, and the result cannot possibly be used to support their argument against the sequential procedure or in favor of the simultaneous procedure. We note as well that the flaw that we discovered could not have been caught by the peer reviewers of the manuscript because key, relevant information was not reported in the article. The problem we uncovered, which we will describe in detail,

is closely related to what is commonly known as Simpson's paradox (sometimes called the amalgamation paradox) in which one collapses across cells of differing sizes in a way that hides a lurking variable and therefore yields a result that fails to reflect the pattern of frequencies in the original cells (e.g., see Hernan et al. 2011). Below we explore the differing cell sizes collapsed by A&W and we identify a critical "lurking variable."

Our argument requires us to point out three uncontested facts about the data on the 236 cases for which dispositions were available. The first and the third (and most critical) of these facts could not be derived from the article. The first uncontested fact comes from the much longer report that was prepared on these data (Amendola et al. 2013) and is not reported in the A&W article. Specifically, cases that were adjudicated guilty were rated as having significantly stronger non-witness evidence (e.g., physical evidence, self-incriminating statements) than were cases that were not prosecuted. As obvious as this fact might seem, this fact is critical to our overall argument and to proper interpretation of the study outcomes. The second uncontested fact is that the simultaneous and sequential lineup cases resulted in the same rate of suspects being adjudicated guilty. This equivalence of simultaneous and sequential in terms of guilty rates is true for the larger set of 236 cases (41 % guilty for sequential versus 36 % guilty for simultaneous, a non-significant difference) and for the subset of 69 cases in which the witness identified the suspect (66 % guilty for sequential, 70 % guilty for simultaneous, also a non-significant difference).

So, the first two facts from the large sample are: (1) adjudicated-guilty cases have more evidence against the suspect than do not-prosecuted cases and (2) the simultaneous and sequential procedures were equivalent in their rates of producing guilty outcomes. These two facts constitute clear scientific expectations regarding necessary properties of A&W's final sub-sample, namely that the proportion of guilty versus not-prosecuted cases needs to be the same for simultaneous as it is for sequential in any smaller sample that was used for further analyses.

A&W's conclusion is based entirely on a final sample of 52 cases, which they use to argue that the evidence supporting the identification of suspects from the simultaneous lineups was stronger than the evidence from the sequential lineups. However, this leads us to our third fact, which could not be discerned from the A&W article, the more extensive earlier report, or a combination of the two. Specifically, the sequential sample ( $N=30$ ) used 16 cases in which the suspect was adjudicated guilty and 14 cases in which the suspect was not prosecuted. The simultaneous sample ( $N=22$ ), in contrast, used 17 cases in which the suspect was adjudicated guilty and a mere five cases in which the suspect was not prosecuted.<sup>1</sup> Remember, however, that adjudicated guilty cases are associated with stronger evidence than cases not prosecuted and that the simultaneous and sequential lineup procedures actually produced equal rates of guilty and not prosecuted in the original, larger sample.

We do not know how that final sub-sample of 52 cases, from which A&W base their entire conclusion, ended up with only 23 % (five of 22) of the simultaneous cases as not prosecuted whereas for the sequential a full 47 % (twice the rate) came from cases not prosecuted, but we now see the "lurking variable" that confounds the A&W analysis,

<sup>1</sup> We were initially able to only estimate these sample sizes by making some mathematical inferences from the data. However, we have now confirmed these numbers with Karen Amendola.

namely the unrepresentative imbalance of guilty and not-prosecuted cases for simultaneous versus sequential lineups. Their samples of sequential and simultaneous lineups profoundly failed to reflect the fact that the rates of guilty versus not-prosecuted cases were equal in the original, larger set of data, and it is clear that the effect of this unrepresentative and unbalanced sample is stacked against the sequential lineups: Cases adjudicated guilty have stronger evidence, and the simultaneous condition was disproportionately stacked with guilty cases.

It is easy to understand why it is not possible to reach a conclusion about the superiority of simultaneous or sequential lineups from this analysis of 52 cases. Consider the following example. Suppose that in a large sample of witnesses those with blue eyes and those with brown eyes produce exactly the same rates of guilty outcomes from their identifications. For a closer look, however, we select two small sub-samples, one of witnesses with blue eyes and the other of witnesses with brown eyes. Somehow, however, in the blue-eyed sample we include 25 % not-guilty outcome cases and 75 % guilty-outcome cases whereas in the brown-eyed sample we include 50 % not-guilty outcome cases and 50 % guilty-outcome cases. Then we look at the average strength of corroborating evidence for the identifications by the sub-sample of blue-eyed versus brown-eyed witnesses. Would anyone be surprised that the sub-sample of blue-eyed witnesses had more evidence corroborating their identifications than did the sub-sample of brown-eyed witnesses? Would we conclude from this that blue-eyed witnesses are superior? No.

Having quite small samples for comparing two procedures is a serious problem in and of itself, but when those two samples involve vastly unequal representation for a critical lurking variable over which the data are aggregated, the result can be profoundly misleading. The sample bias against the sequential lineup is not the only problem we have with the Amendola and Wixted (this issue) article, but our other concerns pale in comparison to the methodological confound (Simpson's paradox) that makes their main conclusion a specious one.

## References

- Amendola, K., Valdovinos, M. D., Slipka, M. G., Hamilton, E., Sigler, M., & Kaufman, A. (2013). *Photo Arrays in Eyewitness Identification Procedures: Follow-up on the Test of Sequential versus Simultaneous Procedures (Study One) and An Experimental Study of the Effect of Photo Arrays on Evaluations of Evidentiary Strength by Key Criminal Justice Decision Makers*. Unpublished manuscript, Police Foundation, Washington, DC.
- Amendola, K., & Wixted, J. T. (2014). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology*. doi:10.1007/s11292-014-9219-2.
- Heman, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40, 780–785. doi:10.1093/ije/dyr041.
- Wells, G. L., Steblay, N. K. & Dysart, J. E. (2014). Double-blind photo-lineups using actual eyewitnesses: An empirical test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior*, in press.

**Gary L. Wells** is Professor of Psychology and the Mark and Wendy Stavish Chair in the Social Sciences at Iowa State University.

---

**Jennifer E. Dysart** is an Associate Professor of Psychology at John Jay College of Criminal Justice, City University of New York.

**Nancy K. Steblay** is a Professor of Psychology at Augsburg College.