Supplementary online material

In the main article we have reported the results from CAC plots. Here we compare identification responses for target present (TP) and target absent (TA) lineups as a function of scale range and scale type using chi-square tests, and we report additional results from CAC plots for scale range for the 4-level scale.

Scale range

Table 1 shows number of observations for various types of identification responses in each confidence bin separately for TP and TA lineups and different scale ranges. We combined the lowest two ratings of the 4-level scale to compare them to the lowest rating of the 2-level scale (low confidence) and the highest two ratings of the 4-level scale to compare them to the highest rating of the 2-level scale (high confidence). First, for TP lineups subjects gave more high confidence responses for correct identifications and fewer high confidence responses for incorrect identifications, suggesting that subjects' confidence level was an indicator of their accuracy. For correct identifications in TP lineups and for correct rejections in TA lineups, subjects using the 4-level scale gave more low confidence responses and fewer high confidence responses compared to subjects using the 2-level scale. This indicates that subjects used the low confidence options on a 4-level scale more ("Not sure at all", "Somewhat sure") than the low confidence option on a 2-level scale ("Not sure at all").

To compare the scale ranges of 2-level and 4-level statistically, two chi-square analyses were conducted. For TP lineups, a 3 (identification response) x 2 (range) analysis revealed a marginal effect of scale range, χ^2 (2, N = 813) = 5.31, p = .070, BF_{10} = .15; however, when the material sets were analyzed separately, the effect was not significant, χ^2 (2, N = 407) = 3.36, p = .187, BF_{10} = .07, and χ^2 (2, N = 406) = 4.53, p = .104, BF_{10} = .22, for set A and B, respectively.

This marginal effect was probably driven by more incorrect identification responses for the 2-level scale (see Table 2). For TA lineups, a 2 (identification response) x 2 (range) analysis revealed no effect of scale range, χ^2 (1, *N* = 815) = 1.86, *p* = .173*, BF$_{10}$* = .24.

We further examined the effect of scale range at different confidence levels. For low confidence TP lineups, there was a significant effect of scale range, χ^2 (2, *N* = 413) = 6.61, *p* = .037, *BF$_{10}$* = .58, indicating that for low confidence responses, the 4-level scale led to more suspect identifications. When the same analyses were conducted for the material sets separately, however, the results did not reach significance, χ^2 (2, *N* = 189) = 3.65, *p* = .161, *BF$_{10}$* = .20, χ^2 (2, *N* = 224) = 4.00, *p* = .135, *BF$_{10}$* = .29, for Set A and Set B, respectively. For high confidence TP lineups, there was no effect of scale range, χ^2 (2, *N* = 400) = 4.14, *p* = .126*, BF$_{10}$* = .15. Responses of 2 on the 2-level scale were about the same as 3 and 4 on the 4-level scale.

The results from TA lineups followed the same pattern. For low confidence responses, there was a significant effect of scale range, χ^2 (1, *N* = 439) = 5.08, *p* = .024, *BF$_{10}$* = 1.88, indicating that the 4-level scale led to more correct rejections (see Table 1). Nonetheless, separate chi-square analyses of the material sets did not reach significance, χ^2 (1, *N* = 235) = 1.72, *p* = .190, *BF$_{10}$* = .45, χ^2 (1, *N* = 204) = 2.95, *p* = .086, *BF$_{10}$* = .97, for Set A and Set B, respectively. For high confidence TA lineups, there was no effect of scale range, χ^2 (1, *N* = 376) = 1.27, *p* = .259*, BF$_{10}$* = .26.

The results indicated that the scale range did not affect overall identification performance, nor did it affect high confidence identification performance. Nevertheless, for low confidence responses the 4-level scale led to more suspect identifications. This was driven by higher accuracy of subjects when they chose “Somewhat sure” on the 4-level scale compared to when they chose “Not sure at all” on the 2-level scale. This difference might stem from a shift in subjects' confidence responses from the 2-level scale to the 4-level scale. For instance, on the 2-level scale subjects who were above a certain level of confidence was forced to choose "Absolutely sure." However, subjects could choose from three different options on the 4-level scale and this would increase accuracy at those options (e.g., they could select "Somewhat sure" on the 4-level scale compared to "Not sure at all" on the 2-level scale). In fact, when provided with less extreme options of "Somewhat sure" and "Very sure" on a 4-level scale, subjects in the current experiment were more likely to select those options expressing less confidence. In short, scale range did not affect correct identifications or correct rejections for high confidence responses, but the 4-level scale seem to cause subjects to select more responses expressing less confidence.

Scale type

Table 2 shows number of observations for identification responses for each confidence bin separately for TP and TA lineups for verbal and verbal+numeric scales. For both TP and TA lineups, the verbal scale and the verbal+numeric scale led to almost identical number of observations.

To analyze the data in Table 2, we compared two scale types (verbal vs verbal+numeric) for overall identification performance and then for performance in each confidence bin. For both TP and TA lineups, chi-square analyses revealed no effect of scale type, χ^2 (2, *N* = 813) = .51, *p* = .774, BF_{10} = .01, and χ^2 (1, *N* = 815) = .14, *p* = .706*,* BF_{10} = .10, respectively. Further, we examined whether the scale type had an effect for low and high confidence bins separately, and chi-squares found no significant differences, *p*s > .05. The results show that adding numeric labels to scales did not affect identification performance.

CAC analyses

In the main text, for CAC analyses we binned the lowest two ratings of the 4-point scale and the highest two ratings of the 4-point scale to compare them to the 2-point scale and reported our results as binned. Here we further compared the 4-point scale and the 2-point scale without binning confidence ratings. Even though the results with binning revealed no accuracy differences between the two scales, it is possible that such binning can create a disadvantage for the 4-point scale. That is, binning 3-point and 4-point confidence ratings together might have lowered accuracy level that would otherwise be obtained from only 4-point. To address this issue, we compared accuracy levels of the highest points on both scales without binning (2-point and 4-point).

Figure 1 shows CAC plot for *only* the highest confidence ratings (i.e., 2-point from the 2-point scale and *only* 4-point from the 4-point scale). For these confidence ratings, accuracy (A) was again computed using the formula, A = # correct suspect IDs/(# correct suspect IDs + # incorrect suspect IDs/6). Accuracy levels were .91 and .94 for 2-point and 4-point, respectively. This difference, however, was not significant: The standard error bars overlapped even when only highest confidence ratings of the scales were compared.

To further demonstrate this point, we plotted a CAC plot only for the 4-point scale in Figure 2. (The number of observations for each of the 4 levels is shown in Table 2 in the main article). Figure 2 shows that even though accuracy steeply increased from 1-point confidence to 3-point confidence, the function plateaued at 3. Accuracy was .93 and .94 for 3-point and 4-point confidence values, respectively. Therefore, we conclude that binning these two confidence levels did not lead to a disadvantage for the 4-point scale and the highest confidence ratings of the 2-point and 4-point scales did not differ in their accuracy levels.

Table 1. *Frequencies of Identification Responses for Scale Ranges in Each Confidence Bin for Target-Present and Target-Absent Lineups.*

Identification response: Target-present lineup

	Suspect ID			Filler ID			Non-ID		
Confidence	total	low	high	total	low	high	total	low	high
2-level	195	73	122	104	70	34	105	37	68
4-level	211	110	101	78	63	15	120	60	60

Identification response: Target-absent lineup

	Non-ID			Filler ID		
Confidence	total	low	high	total	low	high
2-level	210	68	142	196	123	73
4-level	232	116	116	177	132	45

Table 2. *Frequencies of Identification Responses for Scale Types in Each Confidence Bin for Target-Present and Target-Absent Lineups.*

Identification response: Target-present lineup

	Suspect ID			Filler ID			Non-ID		
Confidence	total	low	high	total	low	high	total	low	high
Verbal	199	87	112	95	63	32	113	50	63
Verbal+numeric	207	96	111	87	70	17	112	47	65

Identification response: Target-absent lineup

	Non-ID			Filler ID		
Confidence	total	low	high	total	low	high
Verbal	225	87	138	184	123	61
Verbal+numeric	217	97	120	189	132	57

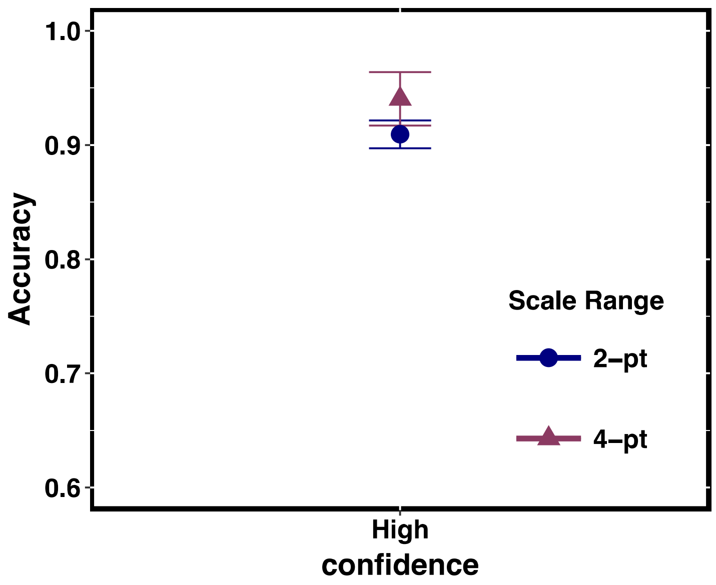

Figure 1. Comparison of the highest confidence points of the 2-point scale and the 4-point scale. Error bars indicate standard errors.

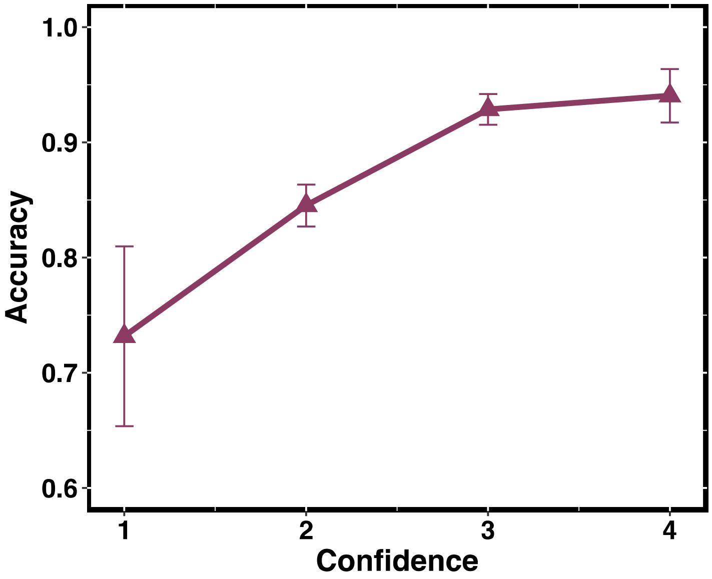

Figure 2. CAC plot of the 4-point scale without binning. Error bars indicate standard errors.