ANNETTE VERHAEG

To what extent do the various measures of confidence affect the accuracy-confidence relationship in earwitnesses? A review of research on the earwitness testimony

Review

Despite inconsistent findings regarding a relation between accuracy and confidence, the use of confidence as a correlate of accuracy is a common subject in forensic psychology research. Therefore, in this review the existence of an accuracy-confidence relationship (A-C relationship) in earwitnesses is examined by considering several variables influencing earwitness memory. First, two different methods of assessing confidence are discussed, the discrimination method and the absolute accuracy method, indicating no differences between these two kinds of methods. Subsequently, confidence is discussed in relation to incorrect/correct decisions and target present/absent conditions. It appears that the mere behavioural act of making an identification (vs. rejecting identification) increases the individual's perceived confidence, regardless whether this identification is correct or not. Finally, future directions for research based on the findings of this review are highlighted.
Keywords: literature study, earwitnesses, accuracy-confidence relationship, line-ups

Annette Verhaeg; Master student Neuropsychology
Maastricht University, Maastricht, the Netherlands

a.verhaeg@student.maastrichtuniversity.nl

INTRODUCTION

When a crime occurs, it is common that an eyewitness is asked to attempt to identify the perpetrator of the crime in a line-up. However, sometimes there are no eyewitnesses who have witnessed the crime, but only earwitnesses. In these situations, the only information available about a crime is the memory of hearing one or more voices. Although it might be thought that people will never forget the voice of a perpetrator, evidence shows that the recognition of a voice can be very difficult (Read & Craik, 1995; Van Wallendael, Surace, Parsons, & Brown, 1994) and that earwitnesses are less accurate in identifications than eyewitnesses (McAllister, Dale, & Keay, 1993). Why is voice identification that difficult?

Voice identification is difficult because of the many factors that influence our memory of a voice. These factors can be divided into three different variable categories, speaker variables, procedural and situational variables, and listener variables (Yarmey, 1995). The influence of these variables is investigated in a lot of studies, but it remains difficult to study only one variable in particular because one variable is almost never independent from other variables. To get an idea of the different variables within the three categories, some examples will be discussed with regard to the categories.

The first category, speaker variables, consists of differences in the voice of the perpetrator between crime and line-up. This category incorporates variables like voice distinctiveness (Mullennix et al., 2011; Saslove & Yarmey, 1980), voice familiarity (Read & Craik, 1995; Yarmey, 1995), accent (Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2006) and voice disguise by whispering (Orchard & Yarmey, 1995), which are all found to influence voice recognition.

The second category, procedural and situational variables, includes variables that differ between interrogation situations. One of these variables is the length of the interval between the witnessing itself and voice identification. Many studies compared the effects of different retention intervals, but evidence in this topic is not unambiguous (Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2004; Kerstholt, et al., 2006; Saslove & Yarmey, 1980; Van Wallendael, et al., 1994). Therefore, it is very difficult to define the optimal retention interval. Other variables within the procedural category are about the effects of being both an earwitness and an eyewitness (McAllister, Dale, Bregman, McCabe, & Cotton, 1993; Stevenage, Howland, & Tippelt, 2011), and about the effects of the line-up, in which topics like one-person versus many-person line-ups (Yarmey, 1995), number of voices in the line-up (Bull & Clifford, 1984) and the effects of feedback (Quinlivan et al., 2009) are emphasized.

The third category, listener variables, is about the witness who may feel different in two situations. This could for instance refer to effects of stress and arousal on voice recognition evoked by the presence of a weapon (Yarmey, 1995). However, a disadvantage in studies investigating these effects is the ecological validity. For ethical reasons participants in such studies cannot be a direct witness of the crime, but just watch a crime on a TV screen, which is not comparable to being witness of a real crime. This could influence the degree of stress and arousal. The listener variable category furthermore includes the effect of confidence (or certainty). "It

seems intuitively plausible that a person is more likely to be correct when he or she is certain of being correct" (Bull & Clifford, 1984, p.104). This assumption is, however, not invariably confirmed, as Bull & Clifford also indicate. There are studies which have found a relationship between the accuracy of an identification (correct or incorrect) and the confidence of a person on his or her identification (Saslove & Yarmey, 1980; Van Wallendael, et al., 1994). However, on the other hand, some studies are published in which no or only a weak relationship is found (Kerstholt, et al., 2004, 2006; Yarmey, 1986).

Despite inconsistent findings regarding a relation between accuracy and confidence, and a warning to be cautious to infer accuracy from confidence, the use of confidence as a correlate of accuracy is a common subject in forensic psychology research (Boydell & Read, 2011). This emphasizes why it is still important to evaluate the existence of an accuracy-confidence (A-C) relationship. Many factors, such as the variables described before, could have an influence on the A-C relationship, but also the measurement of confidence itself could affect this relation (Van Wallendael, et al., 1994). Therefore, this article will discuss some findings about the relationship between accuracy and confidence in detail. The central research question in this article will therefore be: To what extent do the various measures of confidence affect the accuracy-confidence relationship in earwitnesses? Different measurements of confidence will be discussed with respect to the following issues:

1. Does the method of measurement influence results?
2. Correctness and line-up: Do these variables explain differences in confidence rates? In this section, two main issues are discussed:
   a) Does a difference in confidence exist between participants who make an accurate and participants who make an inaccurate identification?
   b) Does a difference in confidence exist between participants who are in a target present condition and participants who are in a target absent condition (line-ups in which the suspect is respectively presented or not)?

The answers to these questions will help in giving a more concrete insight to the strength of the A-C relationship in earwitness identifications. This could also help jurors to make better decisions about a witness with a higher or lower degree of certainty.

DOES THE METHOD OF MEASUREMENT INFLUENCE RESULTS?

Several ways of measuring confidence are used in the studies testing the A-C relationship in earwitnesses. In some experiments, participants first have to listen to a line-up consisting of several voices. Then they have to decide if a target voice (which was heard during an earwitness situation) is present in this line-up. After this decision participants have to rate how confident they are about their choice (e.g. Orchard & Yarmey, 1995). However, in other experiments subjects are told that they will listen to a line-up in which they will hear one voice at a time. In this case, subjects have to record their judgment after each voice as to whether the person is the perpetrator or not, and how confident they are of their decision (e.g. Saslove

& Yarmey, 1980). Van Wallendael and his colleagues (1994) explain that the two methods lead to two different kinds of information about confidence. When the first method is used, the absolute accuracy is considered. "This is defined as the subject's final choice of which line-up voice was the target" (Van Wallendael, et al., 1994, p. 665). In this way, the participant chooses only one voice after hearing the entire line-up and only rates his/her confidence about this voice. This confidence score will only give information about the voice which is, at that time, mostly experienced as the target voice, and gives no information about how other voices are perceived.

When participants have to judge about each voice in particular, a second method is used: discrimination. Discrimination is a ratio of the rated confidence in all the voices. Thus, someone who gives the target a confidence rate of '5', and all other voices '0' will obtain a perfect 1.00 discrimination score. Given that a discrimination score less than 1.00 might indicate that the subject is not entirely sure about his/her decision, the experimenter might receive more information about the memory and cognitive processes of the participant (Van Wallendael, et al., 1994). Moreover, the data obtained for each voice component (target/ distractor) allows further analyses of any accuracy changes (Van Wallendael, et al., 1994). Therefore the use of this method could be of more theoretical interest.

To examine whether these two methods lead to differences in results regarding the A-C relationship, several studies will be discussed in which these different methods are used. Obviously, only the studies which reported results about an A-C relationship are included in this discussion. With regard to the studies that used the absolute accuracy method, only five studies were found (Kerstholt, et al., 2004, 2006; Orchard & Yarmey, 1995; Philippon, Cherryman, Bull, & Vrij, 2007; Read & Craik, 1995). Only four studies were found which used the discrimination method (Saslove & Yarmey, 1980; Stevenage, et al., 2011; Van Wallendael, et al., 1994; Yarmey, 1986). Differences and similarities between studies are discussed.

The absolute accuracy method

In an experiment by Orchard and Yarmey (1995), people had to listen to six voices in a voice line-up. Afterwards, they had to decide whether or not the perpetrator was in the line-up by selecting the number of the voice, by saying that he was not in it or by saying that they did not know. The participants who chose a voice had to indicate their confidence about their choice on a 5-point scale. The relationship between accuracy and confidence was measured as follows: They used a 10-point accuracy-confidence index. In this method, correct identifications (hits) in target present line-ups and correct rejections in target absent line-ups with a confidence rating of 1, 2, 3, 4, or 5 are scored as 6, 7, 8, 9, or 10, respectively. Incorrect identifications with a confidence rating of 1, 2, 3, 4, or 5 are scored as 5, 4, 3, 2, or 1, respectively. This means that the higher the score, the better the accuracy and confidence in the voice-selection. Altogether, results of all measurements revealed significant A-C relationships (point biserial correlations) for both target present line-ups ($r = .25$, $p < .001$) and target absent line-ups ($r = .36$, $p < .001$). This is, however, the only study which used the absolute accuracy method and which found significant A-C relationships.

Four other studies found no significant A-C relationship (Kerstholt, et al.,

2004, 2006; Philippon, et al., 2007; Read & Craik, 1995). In the research conducted by Read and Craik (1995), participants listened to a 6-person line-up, and had to indicate their certainty on a 4-point scale. Three experiments were provided, but overall, no significant A-C relationship was found. They only found a significant relationship when the voice in the line-up was identical (exact same recording) to the voice in the witness situation ($r = .25$, $p < .05$ and $r = .40$, $p < .001$). When the line-up voice was rerecorded there was only a slight tendency for an A-C relationship ($r = .17$, no p-value reported). However, in reality the voice line-up will never be an exact copy of the voice during the crime. Moreover, this tendency for a relationship was only due to an increase in certainty for correct choices, but not for incorrect choices. Therefore they concluded that the A-C relationship was not strong enough to trust on it.

Two very similar experiments by Kerstholt et al. (2004) and by Kerstholt et al. (2006) again found no significant A-C relationship. In both studies people listened to 6 voices and indicated their confidence ratings on a 7-point scale after hearing all the voices. In order to investigate whether the accuracy of the judgements could be predicted by the confidence judgement of the participant, a logistic regression analysis was carried out. In both experiments no relationship was found. However, it is notable that in both articles no r-values and p-values are reported. Finally, a research conducted by Philippon et al. (2007), found no significant relationship ($r = 0.094$, $p = .475$). In this study point biserial correlations were used to measure the A-C relationship. Participants listened to a 6-person line-up, and rated their confidence on a 5-point scale.

In conclusion, most studies using the absolute accuracy method found no strong A-C relationships and overall this relationship was found by assessing the point biserial correlation. However, it is very difficult to compare the results of all these studies, because very different scales are used to assess the person's confidence, ranging from 4- to 7-point scales. This could lead to differences between results, as people differ in their answers depending on the scale length (Forzano & Gravetter, 2009). Moreover, it is remarkable that only the study of Orchard and Yarmey (1995) transformed the raw confidence rates in ten new scores by using an accuracy-confidence index. In this way, the fact of making an accurate or inaccurate identification was taken into account when calculating the A-C relationship. The biserial correlation calculated by using these scores is different from using the raw scores. With regard to the four other studies, it is clear that they also used a biserial correlation, but it is not clear if they used such index as well. This could produce differences in the final A-C correlation. These issues make it difficult to draw a definite conclusion about whether the use of the absolute accuracy method leads to a significant or insignificant A-C relationship.

The discrimination method

In a research conducted by Saslove and Yarmey (1980) participants listened to a 5-person line-up and judged whether each voice was old (perpetrator) or new (innocent person). In addition, subjects had to indicate their confidence level in their old/new decisions on a 3-point scale (possible, probable, or certain). For each participant the line-up consisted of one target voice and four distractor voices.

Recognition memory was analysed in terms of hit - miss scores at one scale and false alarms - correct rejection scores at a second scale. The hit - miss scale consisted of six possible scores ranging from certain miss (score 1) to certain hit (score 6). Therefore, the maximum score on this scale was six. On the other hand, the false alarm - correct rejection scale also consisted of six possible scores ranging from certain false alarm (score 1) to certain correct rejection (score 6). This was evaluated for each of four distractor voices. In this way the maximum score on this scale was 24 (4 times a certain rejection score of 6). The higher the score on each scale, the more accurate is the response to the voices. A point biserial correlation was used to assess the relationship between certainty (total score of a participant) and accuracy (correct identification or not). A small but significant correlation was found ($r = .26$, $p < .01$).

Another study which used the discrimination method was the experiment by Yarmey (1986). Similar to the study by Saslove and Yarmey (1980), this experiment used a five-person line-up. However, participants were not told how many voices they would hear. Again the participants had to indicate whether the voice they heard belonged to the perpetrator or not. Additionally, subjects had to indicate their certainty for each decision on a 4-point scale. Although this research used the discrimination method as well, the point biserial correlation between confidence of response on correct identifications and confidence of response on incorrect identifications was not significant ($r = -.003$, p-value not reported).

It is notable, however, that Yarmey (1986) did not completely describe which scores of confidence he used in the correlation measurement. It is not clear, whether he also used a total certainty score as a measurement for confidence, as used in Saslove and Yarmey (1980), or whether they used another scoring method. Furthermore, in this experiment the participants did not only hear the perpetrator during the observation of the assault, but also saw the perpetrator. This makes a comparison difficult, since visual information can interfere with auditory information (McAllister, Dale, Bregman, et al., 1993).

In the study by Stevenage, Howland, & Tippelt (2011), participants were randomly exposed to a dual-input (audio and visual) or to a single-input condition (visual only/ audio only). In this way, it was possible to investigate the influence of interference on the A-C relationship, as described above. The results of McAllister, Dale, Bregman, et al. (1993) were replicated, although this did not lead to differences in significance of the A-C relationship. The results showed neither a significant A-C relationship for the dual-input ($r = .15$) nor for the single-input condition ($r = .13$) at the 0.05 significance level. These findings make a comparison between Saslove and Yarmey (1980) and Yarmey (1986) more meaningful.

Three striking points, with regard to the study by Stevenage, Howland, & Tippelt (2011) are noticed. First, in this research they use mean confidence scores to examine the A-C relationships. However, as in the experiment by Yarmey (1986), it is not completely clear how they calculated these mean confidence rates. Secondly, they used a 7-point scale to assess the confidence of participants, which is very different from the 3-point scale used by Saslove & Yarmey (1980) or the 4-point scale used by Yarmey (1986). The remaining point to notice is about the voices that had to be remembered. In this study participants had to remember eight studied voices instead of one perpetrator's voice. As indicated in the study, this could also

influence the results, because the task of studying and later identifying eight targets it is very different from identifying a voice which is incidentally learned.

Finally, Van Wallendael et al., (1994) also discussed the discrimination method. In this study the A-C relationship of earwitnesses was measured in three ways. Before listening to the line-up, subjects had to indicate their confidence (on a 7-point scale) in their ability to recognise the voice. They then listened to the line-up, during which they rated each voice on a scale from 0 (sure, this is not the voice) to 6 (sure this is the voice). After listening to all the voices, participants were asked to choose the voice that they believed to be of the perpetrator. In addition, they had to indicate if they would swear to this identification in a court of law. The results revealed no significant relationship between subjects' pre-line-up confidence and the actual performance in the recognition task. This was measured by comparing the average pre-line-up confidence ratings of accurate and inaccurate participants. The relationship between accuracy and post-line-up confidence in the chosen voice was significant ($p < .01$), in that accurate subjects showed higher confidence ratings than inaccurate subjects. Furthermore, a greater accuracy was found for subjects who were willing to swear to their identification ($p < .001$ and $p < .05$ for respectively target present and target absent line-ups). Unfortunately, it is not completely clear how the mean confidence scores were estimated. The researchers asked people to indicate their confidence level on each voice, but it seems that they only used the confidence score of the chosen voice in their assessment of the A-C relationship. If this is the case, they did not use the discrimination method, but the absolute accuracy method.

In conclusion, two of four discussed studies seemed to find correlations between accuracy and confidence by using the discrimination method. An explanation for these ambiguous findings might be sought in the way in which these studies use the discrimination method. Not every experiment exactly clarified how they computed the A-C relationship. It is often not clear how they used all confidence scores (on each voice) in their assessment of an A-C relationship. Many studies report a 'mean' confidence (Yarmey, 1986), but do not explain how this mean confidence rate is calculated. Furthermore, remarkable differences are found in the length of the confidence scales used to measure the certainty of participants, the number of voices in the line-up, and the use of one voice versus more voices that have to be remembered. Due to these differences, no definite conclusion can be made about whether the use of the discrimination method mostly leads to a significant or insignificant A-C relationship.

Conclusion absolute accuracy and discrimination methods

In summary, it has become clear that there are considerable discrepancies between the discussed studies within each method. This makes it very difficult to conclude whether the discrimination method and absolute accuracy method really differ in results about an A-C relationship. To investigate whether these methods really differ from one another, future research has to compare more similar experiments. Furthermore, calculations, with regard to the confidence rates used to assess the A-C relationship, must be more clearly defined. If it turns out to be that those two

methods do not differ, it might be better to use the discrimination method, since this method contains rates of confidence for every voice separately. This allows further analyses and could give us more insight into the memory processes of the participants (Van Wallendael, et al., 1994).

CORRECTNESS AND LINE-UP: DO THESE VARIABLES EXPLAIN DIFFERENCES IN CONFIDENCE RATES?

Most studies and experiments measuring the A-C relationship use mean confidence rates of all participants in their calculation (e.g. Saslove & Yarmey, 1980). However, sometimes it is better not to merge all participants in measuring relationships, because this could give a distorted view of real situations. Confounders could cause relationships to be overestimated or underestimated. An overestimation occurs when a correlation seems to exist when all participants of an experiment are generalised, but disappears when some variables (confounders) are controlled for. An underestimation occurs when no correlation seems to exists, but by controlling for some variables, actually some relationship appears. In calculating the A-C relationship in earwitnesses such over- or underestimating could be present, because in most studies all participants (incorrect, correct, participants in target-present and in target absent conditions) are merged in the analysis.

Read and Craik (1995) found that controlling for a variable did have an influence on results by producing different conclusions. In their study people first had to listen to a target voice and then listened to either recordings of conversations (not the same as initially heard), an identical line-up (exactly same recording as initially heard target voice) or a rerecorded line-up (rerecording of the initially heard target voice). After making a voice identification, participants had to indicate their confidence in their decision on a 4-point scale with four representing high confidence. Afterwards, mean confidence rates (C) were calculated in each condition separately for correct and incorrect decisions. The A-C relationship was assessed by calculating whether C of correct and incorrect responses significantly differed from each other. A significant difference between correct and incorrect C rates was found in the identical line-up condition ($p < 0.001$). However, it was remarkable that this A-C relationship only appeared to be significant, because participants had a higher confidence rate for correct identifications in the identical line-up ($C = 2.63$) than when they had to listen to recordings of conversations ($C = 2.11$). The appearance of the significant A-C relationship was not due to lower confidence rates in incorrect identifications in the identical line-up ($C = 1.88$), because those were very comparable to the rates when they had to listen to recordings of conversations ($C = 1.87$). A second experiment, which was conducted to replicate these results, showed the same trend of results (Read & Craik, 1995).

The overall conclusion in the research conducted by Read and Craik (1995) was that the A-C relationship was too small to conclude that it could be used as a reliable instrument to rely on an earwitness or not. However, it can be concluded from of the study by Read and Craik (1995) that a small relationship seems to exist when

controlling for the variable, '*correctness*' (incorrect or correct identification). This illustrates why it could be important to take a closer look at the differences within groups, because mean confidence rates do not always show these small differences. Therefore, more studies reporting numbers of confidence rates for correct and incorrect participants separately are discussed in this section.

Furthermore, some studies contain another variable, 'line-up', in which confidence rates of target present (TP) line-ups and rates of target absent (TA) line-ups are compared. To investigate if these two kinds of line-ups differ in confidence rates, these results are discussed as well.

Only five studies were found (Pickel, French, & Betts, 2003; Read & Craik, 1995; Stevenage, et al., 2011; Van Wallendael, et al., 1994; Yarmey, 1986), in which numbers about confidence rates were represented separately for 'correctness' (participants who identified correctly or incorrectly), 'line-up' (participants who were in a TP line-up, or in a TA line-up), or both. In one of these experiments the participants were both an earwitness and an eyewitness (Pickel, et al., 2003). This study used both a TP line-up and a TA line-up, but reported numbers only for correct (witnesses who correctly identified the target's voice, or who correctly rejected all voices in target absent line-ups) and incorrect participants (witnesses who incorrectly rejected the line-up, or who incorrectly identified a distractor voice). A 7-point scale was used to measure confidence (C) with higher ratings representing more confidence. In the experiment, a marginally significant difference ($p = .062$) was found between correct (mean $C = 4$) and incorrect decisions (mean $C = 3.54$). However, because the participants in this experiment were both eyewitness and earwitness, results could be distorted (McAllister, Dale, Bregman, et al., 1993).

Another study, in which numbers about confidence rates were reported, was a study by Stevenage et al. (2011). In this research, no TA line-up was used, so the results only contained information about the '*correctness*' of the participants in a TP line-up condition. As in the experiment of Pickel, et al. (2003), they used a 7-point scale to measure confidence. Participants' decisions were only rated as correct when the target voice was indicated as being 'old'(previously heard). Otherwise, their decisions were rated as incorrect. Results revealed a significant difference between correct and incorrect decisions, in which correct decisions showed higher confidence rates than incorrect decisions (respectively, $C = 4.23$ and 3.99, $p = < .001$). This significant difference is, however, not comparable to the results that were found in the study of Pickel, et al. (2003). This may be caused by the fact that the confidence rates as reported by Pickel are averages of confidence rates in TP and TA line-ups (confidence rates of hits & correct rejections are merged, as are misses and false alarms), although Stevenage's study only reports information about the TP line-up condition.

A comparison of the two studies (Pickel, et al., 2003; Stevenage, et al., 2011) is impossible until it is proven that confidence rates between TA and TP line-ups do not differ. If confidence rates between these two kinds of line-ups differ, the rates as reported by Pickel, et al. (2003), are poor representations of the variable 'correctness', because these rates are averages of TA and TP line-ups. To investigate if these line-ups differ in confidence rates, two studies will be discussed that reported information about confidence rates of the two variables '*correctness*' and '*line-up*'

(Van Wallendael, et al., 1994; Yarmey, 1986). In the experiment by van Wallendael et al. (1994), participants could indicate their confidence on a 0 to 6 scale, instead of the 1 to 7 scale used in Pickel, et al. (2003) and Stevenage, et al. (2011). The difference between correct and incorrect decisions in this study was significant ($P < .01$). Participants who made correct decisions rated their confidence higher ($C = 5.456$) than participants who made incorrect decisions ($C = 4.891$). However, the difference in confidence rates between TP line-ups and TA line-ups was significant as well ($p < .001$): Participants in TP line-ups indicated higher confidence rates in their decisions than participants in TA line-ups ($C = 5.364$ and 4.784, respectively).

In line with Van Wallendael, et al. (1994), the experiment by Yarmey (1986) also showed that confidence rates of participants in the TP line-up were higher ($C = 3.05$) than the rates in TA line-ups ($C = 2.7$). Yarmey (1986) used, however, a 4-point confidence scale, instead of a 7-point scale. Furthermore, differences in confidence rates were found between participants that were correct ($C = 2.65$) and those who were incorrect ($C = 3.1$). However, these results were not complemented by p-values and were very much in contrast to the results found by the other studies (Pickel, et al., 2003; Stevenage, et al., 2011; Van Wallendael, et al., 1994), because in this experiment the confidence rates of correct decisions were lower (instead of higher) than the rates of incorrect decisions. It is not completely clear why these results differ so much from the other studies, but as Yarmey (1986) used a 4-point scale instead of a 7-point confidence scale, the difference could be due to the difference in scales. As noted before, it is important for future research to investigate the consequences of these differences in confidence scales used.

The fact that both Van Wallendael, et al. (1994) and Yarmey (1986) found differences in confidence rates between TA line-ups and TP line-ups, indicates that the mean confidence rates for the variable '*correctness*' could be distorted by the differences induced by the variable '*line-up*'. However, the confidence rates for the variable *'line-up*' are mean scores as well, because these scores are the mean of correct and incorrect decisions, separately for TA and TP conditions. Therefore, the results could also be distorted the other way round.

To get more information about this/these distortion(s), confidence rates of Yarmey's study will be re-evaluated in this review. This will be accomplished by first evaluating the value of mean confidence rates for incorrect and correct decisions, after which the value of mean confidence rates for TA and TP line-up will be evaluated. The section will be concluded with an evaluation of splitting up incorrect decisions in false alarms and misses. Table 1 schematically shows how in this article the different mean confidence rates will be calculated from the confidence rates of each particular condition - TA line-up and correct decisions, TP line-up and correct decisions, TA line-up and incorrect decisions, and TP line-up and incorrect decisions. Since the study conducted by Yarmey (1986) was the only study that reported all these confidence rates, only this study can and will be re-evaluated.

Table 1: A schematic representation of the calculations made to obtain mean confidence rates (CR).

	Kind of line-up		
Correctness of decision	TA	TP	Mean CR within correctness
Correct	A	B	(A+B)/2
Incorrect	C	D	(C+D)/2
Mean CR within line-up	(A+C)/2	(B+D)/2	

Note. TA = target absent line-up; TP = target present line-up; CR = rate of confidence; A = mean CR for participants who correctly rejected the line-up in the TA condition; B = mean CR for participants who correctly identified the target in the TP condition; C = mean CR for participants who incorrectly identified a distractor voice in the TA condition; D = mean CR for participants who incorrectly identified a distractor voice or missed the target voice in the TP condition; (A+B)/2 = mean CR for participants who made a correct decision; (C+D)/2 = mean CR for participants who made an incorrect decision; (A+C)/2 = mean CR for participants in the TA condition; (B+D)/2 = mean CR for participants in the TP condition.

The value of mean confidence rates for incorrect and correct decisions

To evaluate whether mean confidence rates of the variable '*correctness*' are informative, different confidence rates reported by Yarmey's study are discussed (see table 2). The first thing to notice is the difference of correct decisions between the TA line-up and the TP line-up. People who correctly chose a voice in the line-up, were far more confident about their decision (C = 3.65) than people who correctly rejected the voices in the line-up (C = 1.65). For this reason, the mean confidence rate of correct decisions (C = 2.65) is not a good measurement of overall correct decisions, as it underestimates confidence rates of participants who make correct decisions in the TP line-up condition, and overestimates correct decisions of participants in the TA condition.

Table 2: Mean confidence rates separate for participants in each condition (TA correct, TP correct, TA incorrect and TP incorrect).

	Kind of line-up		
Correctness of decision	TA	TP	Mean CR within correctness
Correct			2.65
Correct rejection	1.65		
Hit		3.65	
Incorrect			3.1[a]
False alarm	3.75	3.5	
Miss		1.4	
Mean CR within line-up	2.7	3.05[b]	

Note. TA = Target absent; TP = Target present; CR = confidence rate. Adapted from "Earwitness speaker identification", by A.D. Yarmey, 1986, *Psychology, Public Policy, and Law*, *1,* 792-816. [a]Calculated by first averaging the CR for 'incorrect decisions in TP condition' (mean of false alarm and misses in the TP condition = 2.45) after which the incorrect confidence rates for TA and TP condition are averaged [(2.45+3.75)/2]. [b]Calculated by first averaging the CR for 'incorrect decisions in TP condition' (mean of false alarm and misses in the TP condition = 2.45) after which the TP confidence rates for incorrect and correct decisions are averaged [(2.45+3.65)/2].

The same effect is found to be true in the evaluation of the mean confidence rate of incorrect decisions. Therefore, it can be concluded that it is important not to report mean confidence rates for incorrect and correct decisions, because these rates are both poor predictors of the evaluated rates of TA correct, TA incorrect, TP correct and TP incorrect. Future research needs to place greater value on the TA and TP confidence rates separately for correct and incorrect decisions, before using the difference between mean correct and mean incorrect confidence rates to draw conclusions about an A-C relationship (as it was calculated in e.g.Read & Craik, 1995).

The value of mean confidence rates for TA and TP line-up conditions

To evaluate whether mean confidence rates of the variable '*line-up*' are informative, again different confidence rates reported by Yarmey (1986) were re-evaluated (see table 2). The difference of TA line-up between correct and incorrect decisions is noteworthy. Participants who falsely identified a voice in the TA line-up, were far more certain about their decision (C = 3.75), than participants who correctly rejected a voice (C = 1.65). The report of mean confidence rates of TA line-up gives no information about these very different confidence levels, and therefore it is not of great value to use as a predictor of TA correct and TA incorrect confidence rates. The same conclusion holds for the mean confidence rate of the TP line-up, which is a poor predictor of TP correct and TP incorrect confidence rates. This results in a conclusion very similar to the one made in the previous section. Again, it can be concluded that future research needs to place greater value on the confidence rates of correct and incorrect confidence rates separately for TA and TP line-ups, before using the difference between mean TA and mean TP line-ups rates to draw conclusions about an A-C relationship.

The importance of splitting up incorrect decisions in false alarms and misses

A remarkable aspect of Yarmey's study is the fact that confidence rates for misses (incorrect rejection in TP line-up) are reported separately from false alarms. In the other studies (Pickel, et al., 2003; Read & Craik, 1995; Stevenage, et al., 2011; Van Wallendael, et al., 1994), confidence rates of these two groups were taken together, as incorrect decisions. However, as can be seen in table 2, the division of these two groups seems useful, because the confidence rates between the false alarm group and misses group (within TP line-up), are quite different (respectively, C = 3.5 and 1.4). By splitting up these two groups, more information is obtained about the confidence levels that people have in different kinds of decisions (choosing or rejecting). The mean confidence rate of incorrect decisions, when false alarms and misses are taken together, C = 3.1 (as it was calculated in Pickel, et al., 2003; Read & Craik, 1995; Van Wallendael, et al., 1994), does not give this information, because it underestimates confidence rates of false alarms (C = 3.63) and overestimates confidence rates of misses (C = 1.4). Furthermore, this mean confidence rate of 3.1 underestimates the confidence rate for participants in the TA condition (C = 3.75) and overestimates the confidence rate for participants in the TP condition (C = 2.45).

The same effect is found to be true in the evaluation of the mean confidence

rate of the TP line-up condition (see table 2). Therefore, it can be concluded that splitting up incorrect decisions to false alarms and misses would also have been better in the other studies (Pickel, et al., 2003; Read & Craik, 1995; Stevenage, et al., 2011; Van Wallendael, et al., 1994) and would give a more detailed view of participants' confidence in their decisions. Another reason why reporting misses and false alarms separately is important, is because in reality it is less serious when a perpetrator is accidently missed, than when an innocent person is falsely blamed. Moreover, since the confidence rates are much lower for misses than for false alarms, this emphasizes why it is worrisome to rely on these rates when using them to infer accuracy.

Conclusion correctness and line-up

Altogether, it can be concluded that the mean confidence rates of both '*correctness*' and '*line-up*' are poor predictors of mean confidence rates. Because these mean rates could give a distorted view of the situation, it is important not to use them in the assessment of a relationship between confidence and accuracy. Although the study of Yarmey (1986) was the only study that reported information about the specific confidence rates between conditions, these rates were so distinctive that this conclusion could be made in this review.

Furthermore, it is very remarkable and predictable, that participants who chose a voice (hit or false alarm) were very similar in their certainty level. They all indicated their certainty level very high (around 3.6). On the other hand, participants who rejected a line-up (correct rejection or miss) indicated much lower certainty levels (around 1.5). In general, this indicates that participants are not very certain about a rejection, but are more certain about an identification. This fact seems to fit a statement that was made by Loftus (1979) (as cited in Van Wallendael, et al., 1994) who believes that participants have a tendency to make an identification instead of rejecting a line-up. Loftus (1979) explained this by the participant's belief that no test would be conducted, unless there was a reason for it. As a consequence, this mindset causes participants to not feel very certain about a rejection of a line-up, because it conflicts with their own thoughts. Because of this finding the previous conclusion, about not using mean confidence rates of correct, incorrect, TA and TP to assess the existence of a relationship between accuracy and confidence, is emphasized.

SUMMARY AND FUTURE DIRECTIONS

This review discussed several issues that are related to the accuracy-confidence (A-C) relationship. It was investigated to what extent various measures of confidence could have an influence on the A-C relationship. This was evaluated by discussing different issues. First, two different measurements of confidence were evaluated, the discrimination method and the absolute accuracy method, which differed in the moment of measuring confidence of participants in their decisions. Several studies were discussed, to investigate if these two different measurements led to

different results in A-C relationships (Kerstholt, et al., 2004, 2006; McAllister, Dale, Bregman, et al., 1993; Orchard & Yarmey, 1995; Philippon, et al., 2007; Read & Craik, 1995; Saslove & Yarmey, 1980; Stevenage, et al., 2011; Van Wallendael, et al., 1994; Yarmey, 1986). However, no obvious differences were found.

Studies that used the discrimination method found very inconsistent results; some found an A-C relationship, whereas others did not. On the other hand, studies using the absolute accuracy methods, found no or weak relationships. It is remarkable that all discussed studies differed in many aspects, for instance in the number of voices in the line-up. Since this could influence the results, this made the comparison very difficult (Bull & Clifford, 1984). Another point that made a comparison very difficult was the difference in confidence scales used. Some studies used a 3-point scale (Saslove & Yarmey, 1980), whereas others used a 4- ,5- or even a 7- point scale (e.g. Orchard & Yarmey, 1995; Stevenage, et al., 2011; Yarmey, 1986). These different scales could also influence the results, since a scale from five to ten was found to be better than a shorter scale (Forzano & Gravetter, 2009).

Finally, the fact that most experiments do not clearly specify how they calculate the confidence rates that are used in the A-C calculation, especially when they use the discrimination method, made a comparison very difficult. Due to these differences between the discussed studies, I conclude that future research is needed to investigate if these two different confidence measurements lead to different results about the A-C relationship. However, until this difference is clearly investigated, it is advised to use the discrimination method, because this method gives more information about the memory processes used by people to make a decision.

In the second section, the potential risk of confounding in confidence rates (C) was discussed. Most studies measuring the accuracy-confidence relationship use mean confidence rates of all participants to assess this relation (mean C within the '*correctness*' of a decision and mean C within the different '*line-up´* conditions). However, after re-evaluating several mean confidence rates reported by the study of Yarmey (1986), it was concluded that these mean confidence rates are no reliable predictors of specific confidence rates for participants who accurately reject a line-up (correct rejection), participants who accurately identify a voice (hit), participants who inaccurately reject a line-up (miss), and participants who inaccurately identify a voice (false alarm). Mean confidence rates of correct decisions (hits, and correct rejections) were found to be poor predictors of these specific confidence rates. The same was found to be true for mean confidence rates of incorrect decisions, target present, and target absent line-ups. Only two confidence rates were found to be predictable, namely the mean confidence rate for people who (correctly or incorrectly) chose a voice, which was very high ($C \approx 3.6$), and the mean confidence rate for people who (correctly or incorrectly) rejected a line-up, which was very low ($C \approx 1.5$) (Yarmey, 1986). Therefore, it was concluded that mean rates should not be used in calculation of the A-C relationship, but that it is better to evaluate specific confidence rates to get information about people's decisions. In reality it remains, however, very difficult to implement all findings about A-C relationships, because in reality the real truth is never known. Therefore, it remains tricky to use confidence rates in the assessment of accuracy.

The conclusions of this review should be interpreted with caution, because of the low number of papers that could be included in the discussion. Earwitness testimony has been investigated less than eyewitness testimony, and so overall fewer papers could be found regarding this issue. Moreover, because issues were discussed and re-evaluated for which specific information was needed, only a few studies were useful to include in this review. However, despite the fact that this limitation causes the conclusions to be more difficult to generalise, it emphasizes that much more research is needed in this area. Not until more systematic research on the discussed issues is conducted, it is crucial to be careful in reporting whether a relationship between accuracy and confidence in earwitnesses exists or not in research on earwitness testimony. In this way, this review may serve as an eye opener for future research about specific issues related to earwitness testimony.

REFERENCES

Boydell, C. A., & Read, J. D. (2011). Accuracy of and confidence in mock jailhouse informants' recall of criminal accounts. *Applied Cognitive Psychology, 25*, 255-264.

Bull, R., & Clifford, B. R. (1984). Earwitness voice recognition accuracy. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 92 - 123). New York: Cambridge University Press.

Forzano, L. A. B., & Gravetter, F. J. (Eds.). (2009). *Research Methods for the Behavioral Sciences*. Belmont, CA: Wadsworth.

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology, 18*, 327-336.

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology, 20*, 187-197.

McAllister, H. A., Dale, R. H. I., Bregman, N. J., McCabe, A., & Cotton, C. R. (1993). When eyewitnesses are also earwitnesses: Effects on visual and voice identifications. *Basic and Applied Social Psychology, 14*, 161-170.

McAllister, H. A., Dale, R. H. I., & Keay, C. E. (1993). Effects of lineup modality on witness credibility. *The Journal of Social Psychology, 133*, 365-376.

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology, 25*, 29-34.

Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9*, 249-260.

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology, 21*, 539-550.

Pickel, K., French, T., & Betts, J. (2003). A cross-modal weapon focus effect: The influence of a weapon's presence on memory for auditory information. *Memory, 11*, 277-292.

Quinlivan, D. S., Neuschatz, J. S., Jimenez, A., Cling, A. D., Douglass, A. B., & Goodsell, C. A. (2009). Do prophylactics prevent inflation? Post-identification feedback and the effectiveness of procedures to protect against confidence-inflation in earwitnesses. *Law and Human Behavior, 33*, 111-121.

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1*, 6-18.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65*, 111-116.

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology, 25*, 112-118.

Van Wallendael, L. R., Surace, A., Parsons, D. H., & Brown, M. (1994). 'Earwitness' voice recognition: Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology, 8*, 661-677.

Yarmey, A. D. (1986). Verbal, visual, and voice identification of a rape suspect under different levels of illumination. *Journal of Applied Psychology, 71*, 363-370.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1*, 792-816.