COMPARING REACTION TIMES AND ACCURACY RATES IN A *SAME / DIFFERENT* TASK WITH FACIAL EXPRESSIONS

Galina V. Paramei[1], David L. Bimler[2], Slawomir J. Skwarek[3]
[1]*Liverpool Hope University, Hope Park, L16 9JD Liverpool, United Kingdom*
[2]*Massey University, Private Bag 11-222, Palmerston North 4442, New Zealand*
[3]*University of St. Gallen, Dufourstrasse 40a, 9000 St. Gallen, Switzerland*
parameg@hope.ac.uk; d.bimler@massey.ac.nz; slawomir.skwarek@unisg.ch

Abstract

The relationship was examined between two performance measures of a same / different *task*: different *reaction times (RTs) and percentage of* different *judgements. A set of facial expressions (FEs) of emotion (N=15) included seven emotion prototypes and eight intermediate morphs. The FEs were presented pairwise for 500 ms, 100 times each pair. The percentage of* different *judgements behaved as an index of perceptual dissimilarity. Contrary to expectation, the relationship between the two measures was non-monotonic: As a FE pair became more similar,* different *RTs increased but only to a maximum at an intermediate level of similarity where* different *and* same *responses were equally likely. Below that level, for even more similar FE pairs,* different *RTs became progressively shorter. We explain the phenomenon by a dual-process model: incompletely developed FE percept implies a decision based predominantly on global (shared) features and, thus, feeds in a fast* same-*process followed by premature responses.*

The *same / different* task has been widely used for measuring the similarities among stimuli, such as polygons (Cooper, 1976); letters (Podgorny & Garner, 1979; Courrieu et al., 2004); schematic facial expressions (Takane & Sergent, 1983), and colours (Paramei & Cavonius, 1999).

For *N* stimuli, the *N* (*N*-1)/2 pairs of different stimuli are presented some number of times (*M*), in random order, interspersed with repetitions of the *N* identical-stimulus pairs. To prevent the task from becoming trivial, often its difficulty is increased by conditions such as a brief stimulus exposure. Inter-stimulus dissimilarity is generally above the threshold of discrimination, and different pairs are recognized as such in the majority of trials.

One index of dissimilarity is discriminative RT required for the *same/different* decision, averaged over all observations of a stimulus pair (cf. Podgorny & Garner, 1979). A widely-accepted generalisation is that in the case of accurate descriptions of a pair of stimuli as *different*, RT declines steadily as their dissimilarity increases. The dissimilarity measures considered were either direct subjective ratings (Podgorny & Garner, 1979; Paramei & Cavonius, 1999) or confusability estimated as the percentage of *different* responses under both *same* and *different* response options (Takane & Sergent, 1983) or only *different* option (*go/no-go* procedure; Courrieu et al., 2004).

In these studies, the discriminative RT function was found to slope down steeply when the dissimilarity is subtle and the decision is difficult, levelling out and approaching a floor value as the difference between the stimuli becomes immediately apparent. Following Shepard (1987), a negative exponential function often fits the function well.

To derive the RT function, dual-process decision models have been suggested. The diffusion model, in particular, postulates a pair of competing ‘evidence accumulators’,

one receptive to any points of difference between the stimuli and the other of sameness (e.g. Ratcliff, 1985). The accumulators function gradually and in parallel until one or other reaches a threshold. The relative rate of accumulation of evidence between *same* and *different* comparisons can be manipulated (Ratcliff, 1985) but in general the *same* process tends to be faster than the *different* one (e.g. Eviatar et al., 1994). In addition, the *different* and *same* processes may have different thresholds to attain, affected, for instance, by manipulating the proportion of trials where the stimuli are identical (Krueger & Shapiro, 1981).

In the present study we used the *same/different* procedure and employed a set of FEs, i.e. visual stimuli more elaborate than those in the above mentioned studies. As we argue elsewhere (Bimler et al., 2009), the percentage of correct *different* judgements for a pair of FE stimuli is a robust index of their dissimilarity. Here we examine the relationship between the rate of *different* responses and *different* RTs while asking whether the two measures demonstrate monotonicity and RTs can be considered a sensitive measure of FE dissimilarity.

Method

Photographs of FEs were selected from Ekman and Friesen (1976) that represented the prototype emotions: *Happiness*, *Surprise*, *Anger*, *Sadness*, *Fear*, *Disgust*, *Neutral*. Seven images featured a female poser MO while the other seven a male poser WF. The "MO series" and "WF series" were both extended by eight morphs produced by image interpolation midway along the continua between two 'parent' prototypes (Figure 1).

The digitalized stimuli were presented on a monitor; each image occupied 12.8 x 8.7 cm (subtending 10° x 6.7° at a viewing distance of 74 cm). On each trial, two FEs were presented symmetrically side-by-side with a 3.8 cm gap between them (subtending 3°). The exposure duration was 500 ms. After this, the screen went blank until the subject responded via a two-button keyboard. Subjects were instructed to judge whether the *emotion* expressed in the two stimuli was *same* or *different* as quickly and correctly as possible.

RTs were measured from the appearance of the FE pair to the response. A MS-DOS program controlled presentations and recorded the response (*same* or *different*) as well as RT (to the nearest 20 ms). Each response was followed by an inter-stimulus interval of 300-400 ms, during which a small red fixation cross was displayed on the monitor.

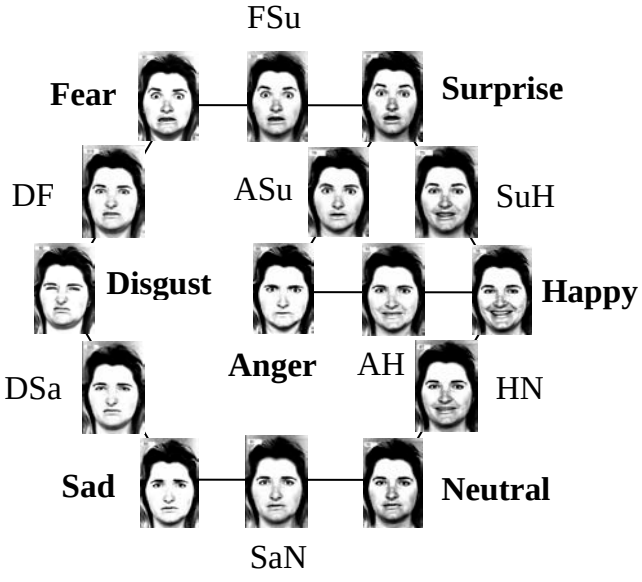

Fig. 1. FE stimuli illustrated by the "MO series" arranged in a distorted circumplex. The prototype FEs are accompanied by their emotion names and the morphs by abbreviations.

In a single run, all possible 15 x 15 = 225 pairings of FEs were presented in randomised order. All stimulus pairs were repeated equally often; thus identical FE pairs were shown on about 7% of trials. Each of the FE pairs was presented 100 times in the course of 30 sessions, spread out over four months. This high number of trials per stimulus pair was intended to reduce the noise and thus provide robust data (cf. Eviatar et al., 1995).

Four undergraduate psychology students, aged 21 to 25 years old, participated as experimental subjects for pay. Subject gender and poser gender were counterbalanced: stimuli from the "MO series" were presented to one female (DK) and one male (HK) observer; likewise the "WF series" was presented to one female (SB) and one male (BF) observer.

Results and Discussion

For each FE pair, median RTs and accuracy rate were calculated. For the present analysis, median RTs for the correct *different* responses and percentage of *different* judgements were considered. Noteworthy the complementary *same* responses occurred on about 25% of trials (compared to 7% of trials with objectively identical FEs), giving an indication that, under the condition of rather short exposure to complex visual stimuli (500 ms), the subjects tended to overlook points of difference or paid more attention to points of similarity. In particular, the percept develops in such a way that global features are identified before local features, which can distinguish non-identical stimuli from each other (cf. Eviatar et al., 1994). Both identical and non-identical stimulus pairs produce early priming for a *same* response, which competes with the *different* response when stimuli are non-identical. Because of the "exposure stress", the FE percept seemingly has developed to the stage of extracting global features (which similar FEs share) but not for fine discriminations to be made. The global-feature similarity should have promoted the *fast* process of accumulating *sameness*. However, in a low percent of trials the responses are accomplished as *different* – perhaps because some difference between local features 'pop out' before the decision.

Figure 2 shows median RTs plotted against percentage of *different* responses. Data for two subjects are presented, HK and DK, though the diagrams are representative of those for the other two subjects. The size of each dot represents the number of decisions on which that median is based, as an indication of its reliability. The unexpected feature here is that RTs follow an inverted-U function rather than an exponential function or any other monotonic decline. As expected, RTs are short for easy *different*-judgement FE pairs and lengthen

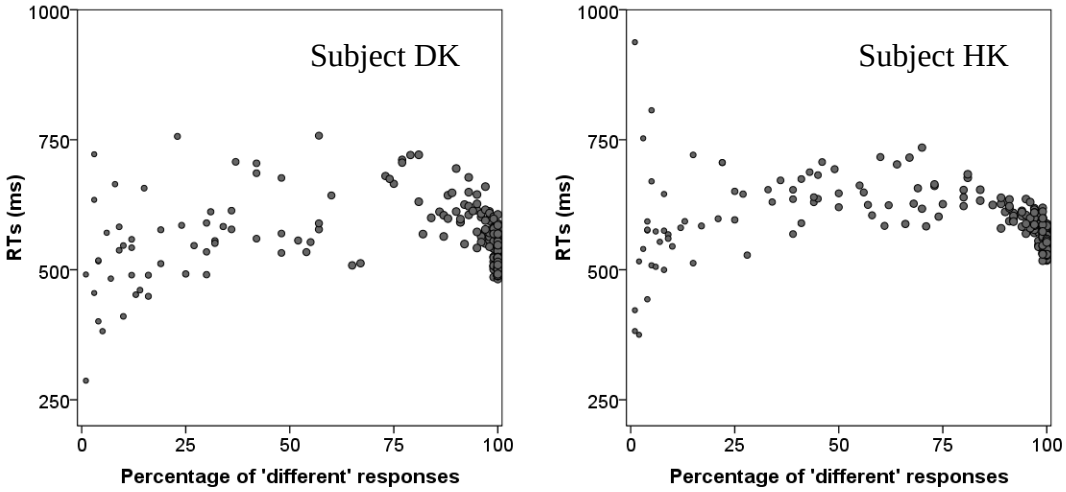

Fig. 2. Relationship between two performance measures in the *same* / *different* task with facial expressions: medians of *different* RTs vs. percentage of *different* responses.

with increasing FE similarity, but this holds only down to about 50% of *different*-judgement rate. For each subject, *different* RTs reach their peak at the dissimilarity where *different* and *same* responses are equally common. Below this value, for subjectively similar FEs, the *different* RTs are shorter. In addition, for FE pairs with very little dissimilarity, i.e. difficult *different*-decisions accompanied by a predomination of erroneous *same* responses, significant variation among RTs is apparent.

In explaining the present results, i.e. shortening of *different* RTs for highly similar FEs, the diffusion model provides the helpful insight that the production of *same* RTs, as a function of subjective dissimilarity, tends to mirror the *different*-RT function (Takane & Sergent, 1983, Fig. 4; Ratcliff, 1985). That is, the accumulation of *sameness* is more rapid for more-similar but non-identical pairs. And crucially, when the *same*-response option is used frequently for a given pair, it places an upper limit on the RTs of *different* responses. The entire distribution of *different* RTs (which might well exhibit an increasing median as dissimilarity decreases) is not observed – only a truncated tail of that distribution, of unusual cases where the *difference*-accumulation process has outraced the accumulation of *sameness*.

To put it another way, the non-monotonicity is in fact an inherent feature of the *same / different* task. Usually concealed, it became evident in the present results because conditions curtailed the development of visual analysis, and shifted the balance between the 'evidence accumulators' in the dual-process decision, facilitating erroneous *same* responses.

According to Ratcliff (1985), the low frequency of identical FE pairs in the present design should have promoted the accumulation rate of *different* comparisons. Apparently, this facilitation could not override the *same*-accumulation process – driven by extraction of shared global features due to the brief exposure to facial expressions, visually complex stimuli.

References

Bimler, D.L., Skwarek, S.J., & Paramei, G.V. (2009). Processing facial expressions of emotion: Upright vs. inverted images (under revision).

Cooper, L.A. (1976). Individual differences in visual comparison processes. *Perception & Psychophysics*, *19*, 433-444.

Courrieu, P. Farioli, F., & Grainger, J. (2004). Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition, 11*, 901–919.

Ekman, P. & Friesen, W.V. (1976). *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

Eviatar, Z., Zaidel, R., & Wickens, T. (1995). Nominal and physical decision criteria in *same-different* judgments. *Perception & Psychophysics, 56*, 62-72.

Krueger, L.E. & Shapiro, R.G. (1981). A reformulation of Proctor's unified theory for matching-task phenomena. *Psychological Review, 88*, 573-581.

Paramei, G.V. & Cavonius, C.R. (1999). Color spaces of color-normal and color-abnormal observers reconstructed from response times and dissimilarity ratings. *Perception & Psychophysics, 61*, 1662-1674.

Podgorny, P. & Garner, W.R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics, 26*, 37-52.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92*, 212-225.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Takane, Y. & Sergent, J. (1983). Multidimensional models for reaction times and same-different judgments. *Psychometrika, 48*, 393-423.

THE OVERCONFIDENCE PARADOX IN PERCEPTUAL TASKS

Dan Zakay [1] and Dida Fleisig [2]
1 New School of Psychology, IDC, Herzeliya, Israel.
2 Department of Behavioral Science, Peres Academic Center, Rehovot, Israel.

Abstract

A paradoxical state regarding people's confidence in the correctness of their retrieved knowledge has been demonstrated in numerous studies of general-knowledge tests: when comparing the average of specific confidence judgments that are measured for each question in a test (i.e., local confidence), with global confidence judgments that are measured for the total number of questions in a test, local confidence is higher than global confidence. The present study investigated this overconfidence paradox in a perceptual task. Twenty subjects performed a difficult version and another 20 performed an easy version of the task, and were asked to make specific and global confidence judgments, and to provide verbal explanations to the way they judged their global confidence. The overconfidence paradox was found in both versions. We propose that global confidence judgments are heuristically based on the recalled frequencies of specific confidence judgments' categories.

People's confidence in the correctness of knowledge retrieved from memory can be assessed in several ways. Participants may be asked to judge their confidence in the correctness of each answer in a test ("specific confidence" judgments). The average of all specific judgments is then considered to reflect their confidence for the entire test. This measure is called the "local confidence". Most studies document local overconfidence such that local confidence is higher than actual performance in the test (percentage correct) (e.g., Ronis & Yates, 1987). In a different way of judgment which is termed “global confidence” participants are asked, at the end of a test, to judge the number of questions they answered correctly out of the entire number of the test's questions. It has been shown that local confidence is often higher than global confidence (Gigerenzer, Hoffrage, & Kleinbolting, 1991; Griffin & Tversky, 1992). These studies find the same participants overconfident in one way of measurement and calibrated or under-confident in the other. This phenomenon has been called the Confidence-Frequency Effect (Gigerenzer et al., 1991) and we refer to it as the Overconfidence Paradox. The paradox has been demonstrated in several studies that used general-knowledge tasks (e.g., Gigerenzer et al., 1991; Keren, 1991; Sniezek, Paese, & Switzer, 1990; Schneider, 1995).

Several explanations have been suggested for the overconfidence paradox. Schneider (1995) as well as Sniezek et al. (1990) argued that each of the confidence judgments rely on information concerning distinctly different issues, therefore these judgments should not be expected to yield similar results. According to Gigerenzer et al.'s (1991) Probabilistic Mental Model (PMM), participants base their confidence judgments on a probabilistic mental model that they construct in order to perform an inductive inferential process. Within this model, participants construct a reference class that determines what cues can function as probability cues. As long as the cues are valid and relevant to the judgment at hand, judgments turn out to be accurate. Specific confidence cues and global confidence cues are extracted from a different reference classes. In non-ecological tests the reference class for specific confidence cues yields non-valid cues, resulting in local overconfidence. Yet, global confidence relies on

a different reference class resulting in accurate judgments. Another type of explanation provided by Keren (1991) argues that specific confidence is restricted to a 50-100 scale (In double-choice questions) while the global confidence scale ranges between 0-100. As there is no meaning to below-chance judgments, when those are excluded from the analyses, no difference between local and global confidence is found. In addition, Keren (1991) suggests that when participants estimate below-chance accuracy, they might not understand the concept of their judgment accurately.

As none of the above explanations has been sufficiently validated, and there is still disagreement in the literature that needs to be sorted out, we suggest that the global confidence judgment is an outcome of a heuristic process that is based on the specific confidence judgments. When participants judge their specific confidence ratings, they spontaneously encode the frequency of those confidence judgments, grouped into categories. For example, a category may consist of X answers with specific confidence ratings of 100, Y answers with specific confidence ratings of 50, Z answers with intermediate specific confidence ratings, etc. The global confidence is then based on a weighted average of those frequencies, e.g., if a participant recalled the frequencies of his/her specific confidence ratings like in the example given above, his/her global confidence might be composed of X plus Y/2 plus Z/n (n is a weight assigned to the intermediate level of specific confidence). This heuristic, in most cases, will result in a lower global confidence as compared to local confidence (the Overconfidence Paradox) since not all the actual correct answers are taken into account.

The Overconfidence Paradox was demonstrated mainly for general knowledge tasks, and was not explored in the context of perceptual tasks. We claim that due to the universal nature of the global confidence heuristic, the Overconfidence Paradox should be obtained in perceptual tasks.

The present study has two targets: 1. to demonstrate the existence of the Overconfidence Paradox in perceptual tasks and, 2. to support the global confidence heuristic hypothesis.

Method

Participants**:** Forty students (8 male 32 female) participated in the experiment, in partial fulfillment of course requirements. Mean age was 24.7 years (range 18-30; sd=4.02). Half of the sample was given a difficult version of the test, and the other half received an easy version. All participants had normal or corrected-to-normal visual acuity, and all were naïve as to the purpose of the experiment.

Apparatus**:** A perceptual discrimination task, based on the tasks used by Baranski and Petrusic (1994) as well as Olsson and Winman (1996), was administered. The experiment was conducted on a PC computer, which presented the stimuli and recorded participants' responses. Stimuli were presented on a PC monitor. Participants were asked to press the X key to indicate "left" and the M key to indicate "right".

Stimuli: In each trial, a 10mm long vertical line appeared at the center of the monitor, along with two 5mm long lines, one on each side. The long central line served as a reference point for the comparison, and as a division of the monitor to left and right halves. All lines were black and 1mm wide. Notations "x" and "y" are used for marking the distance of a pair of stimuli in pixels (3.13 pixels = 1 mm), left of the central line (x) and right of the central line (y). Six pairs were presented (x, y): (30 ,28; 28, 30) ;(50 ,48; 48, 50); (70 ,69; 69, 70); (260, 243; 243, 260); (280 ,269; 269, 280); (300 ,296; 296, 300). The degree of difficulty is expressed by the ratio between the long and the short distances: 1.07; 1.04; 1.01; 1.07; 1.04; 1.01 of the pairs 1-6, respectively. A ratio of 1.07 expresses the easiest comparison, a ratio of

1.04 expresses intermediate difficulty, and a ratio of 1.01 expresses the hardest comparison. Participants in the "easy version" were presented with pairs of the ratio 1.07 (12 trials) and 1.04 (8 trials) randomly ordered; participants in the "difficult version" were presented with pairs of the ratio 1.01 (12 trials) and 1.04 (8 trials) randomly ordered as well. Viewing distance was about 60 cm; the closest and the farthest stimuli formed visual angles of 2^0 and 19^0, respectively

Procedure: The procedure and instructions were similar for both versions of the experiment. Each participant performed the experiment individually with the presence of an experimenter. The participant was instructed to read the instructions carefully and to follow them accurately. The following instructions appeared on the computer monitor (original instructions were given in Hebrew): "*In each step a vertical line will appear on the screen, along with two additional lines, one on each side. Before the appearance of the lines, you will see the title "closer" or "farther" on the upper part of the monitor. You have to decide which of the two lines is closer (or farther) to the central line. To indicate that the left-hand line is closer to the central line, press the X key. To indicate that the right-hand line is closer to the central line, press the M key"*. Participants were then instructed that after each choice they would be asked to rate the degree of their confidence in the accuracy of their answer, on a scale that ranged between 50 (expressing "guess") to 100 ("sure"). It was explained that they might use all scores between 50 and 100, with higher scores expressing higher confidence. Participants then performed four practice trials, and started the test. The test included 20 randomly ordered trials. Each trial started with the caption "*closer*" or "*farther*" appearing on the upper part of the monitor. After 1.5 seconds the stimuli set appeared, presenting a central line and two additional lines, one on each side. Both captions and stimuli stayed on the monitor until the participant pressed either the X or the M keys. Participants were then asked to judge their specific confidence, facing a clear monitor. The next trial began two seconds after the specific confidence judgment was made. After completing the test, participants were presented with the following question: "*How many questions out of the 20 questions presented to you do you believe you answered correctly?"* The participant had to type any number between 0-20. Participants were then asked to provide a detailed and extensive written explanation to the way they performed their judgment.

Results

Six measures were calculated for each participant, including (a) Actual performance (AP - percentage of correct answers in the test); (b) Local confidence (LC - average of specific confidence judgments on the 50-100 scale); (c) Global confidence (GC - estimate of the number of correct answers, represented as percentage out of 20); (d) Local overconfidence (L-OC - local confidence minus actual performance); (e) Global overconfidence (G-OC - global confidence minus actual performance); and (f) Overconfidence paradox (OCP - local confidence minus global confidence). Means and standard deviations (in brackets) of all six variables, in the difficult (D) and the easy (E) versions of the task, are presented in Table 1.

Table 1.

version	AP	LC	GC	L-OC	G-OC	OCP
D	39.75	73.49	59.50	33.74	19.75	13.99
N=20	(6.17)	(11.53)	(19.45)	(10.78)	(18.78)	(12.94)
E	79.75	72.95	56.25	-6.79[(1)]	-23.50[(1)]	16.71
N=20	(10.32)	(8.92)	(15.88)	(11.02)	(14.06)	(12.79)

[(1)] A minus sign indicates under-confidence

Difficult version: A repeated measure analysis of variance (ANOVA) was conducted with actual performance and local confidence as within-subject variables. This analysis found local confidence to be significantly higher than actual performance, $F(1, 19) = 194.80$, $p < .000$, thus showing local overconfidence. A repeated measure ANOVA with local and global confidence as within-subject variables, found local confidence to be significantly higher than global confidence, $F(1, 19) = 23.41$, $p < .0001$, thus revealing the overconfidence paradox. A repeated measure ANOVA with actual performance and global confidence as within-subject variables demonstrated global overconfidence, as global confidence was significantly higher than actual performance, $F(1, 19) = 22.05$, $p < .0001$.

Easy version: A repeated measure ANOVA was conducted with actual performance and local confidence as within-subject variables. This analysis found local confidence to be significantly lower than actual performance, $F(1, 19) = 7.60$, $p < .012$, thereby demonstrating local under-confidence. A repeated measure ANOVA with local and global confidence as within-subject variables revealed a significantly higher local confidence relative to global confidence, $F(1, 19) = 34.12$, $p < .000$, demonstrating the overconfidence paradox. A repeated measure ANOVA with actual performance and global confidence as within-subject variables revealed global under-confidence, as global confidence was significantly lower than actual performance, $F(1, 19) = 55.89$, $p < .000$.

Testing the "global confidence heuristic": Specific confidence judgments ranged between 50 and 100. Each participant's specific confidence judgments were grouped into 4 categories: 50 (representing guess); 51-69 (standing for not so sure); 70-89 (quite sure); and 90-100 (sure). The frequency of judgments in each category was calculated, for each participant. A stepwise multiple regression analysis was performed for each difficulty group. The predicted variable was the global confidence, and the predicting variables were actual performance and local confidence. In both groups, when local confidence and actual performance were entered as potential predictors of global confidence, only local confidence was found to be a significant predictor. (Difficult version: $R^2=0.59$, $F(1, 18) = 25.66$, $p<.0005$; Beta coefficient = 0.77, $p<.0005$; Easy version: $R^2=0.44$, $F(2, 17) = 6.79$, $p<.05$; Beta coefficient = 0.48, $p<.005$).

Additional support to the suggestion that global confidence is based on local confidence rather than on actual performance was derived by comparing confidence ratings across the two difficulty versions of the task. Whereas a one-way ANOVA that compared actual performance across the two difficulty versions revealed a significant difference, $F(1, 38) = 221.29$, $p < .0000$, a one-way ANOVA that compared local confidence across the two difficulty versions found no significant difference, $F(1, 38) = .03$, n.s. This was also the case for global confidence, $F(1, 38) = .33$, n.s. Consequently, the difference between these confidence judgments (i.e., the overconfidence paradox) was the same across the two difficulty versions, $F(1, 38) = .44$, n.s. That is, in spite of the very different levels of performance, the overconfidence paradox was similar in both tasks (see Graph 1), such that the actual accuracy of performance had little or no influence on both confidence judgments and on the overconfidence paradox.

In order to analyze participants' verbal explanations to the way they judged their global confidence, the following three main explanation-categories were defined: (a) relying on their specific feelings of confidence that arose during the test; (b) relying on the frequencies of specific confidence categories; (c) relying on the difficulty or the content of the task. For each difficulty group the proportion of each explanation-category was calculated. Seventy six percent of the participants in the difficult task, and 85% of the participants in the easy task reported that they relied on their specific feelings of confidence;

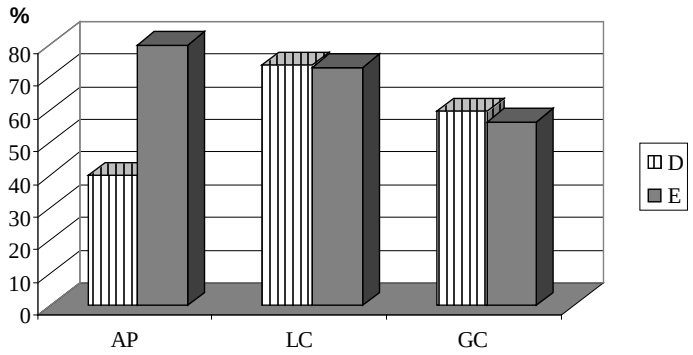

Fig. 1. Actual performance (AP); Local confidence (LC); and Global confidence (GC) in the difficult (D) and the easy (E) versions of the task

53% of the participants in the difficult task, and 75% of the participants in the easy task mentioned relying on frequencies of specific confidence categories; 18% of the participants in the difficult task and 30% of the participants in the easy task reported relying on the difficulty or the content of the task.

Discussion

The present study had two purposes. First, we wanted to display the overconfidence paradox in a perceptual task. Second, we wanted to examine whether global confidence judgments rely on specific confidence judgments by means of a "global confidence heuristic" rather than on performance's accuracy. Two identical perceptual discrimination tasks differing only in their difficulty level were conducted. The difficult task led to low accuracy of performance and to local overconfidence, whereas the easy task led to high accuracy and to local under-confidence. Nevertheless, each task displayed the overconfidence paradox, as local confidence was higher than global confidence in both tasks. Importantly, despite the considerable difference in accuracy of performance between the tasks, they differed in neither local confidence nor global confidence. Consequently, the overconfidence paradox was similar in both tasks. This similarity in both confidence judgments may suggest that there was only minimal (if any) influence of actual performance on both local and global confidence judgments, and that global confidence relies on local confidence rather than on actual performance. Further support to this argument was received from multiple regression analyses, and from the analysis of verbal explanations. These analyses showed that global confidence relies on specific confidence judgments rather than on actual performance.

Some authors have criticized the ability to draw conclusions regarding mental processes from multiple regression analyses (e.g., Slovic, Fischoff, & Lichtenstein, 1977), as well as the ability to learn about mental processes from verbal reports (Ericsson & Simon, 1993). Yet, Sniezek et al. (1990) suggested that asking participants to produce written verbal explanations of their judgments could provide a deeper understanding of the global confidence judgments process. Indeed, Griffin and Buehler (1999) followed this procedure, while other authors recommended combining both types of research in an integrative way (e.g., Svenson, 1985)

The findings obtained in the present study support the hypothesis that global judgments rely on a heuristic process which is based on the recalled frequencies of specific confidence judgments' categories. It could be the case that participants underestimate the probability of specific non-sure answers to be eventually correct. Thus, when they judge global confidence while relying on the specific confidence judgments they do not consider

these non-sure answers as potentially correct. Namely, local confidence is the average of *all* specific judgments, but when evaluating in a heuristic way the global confidence, not *all* answers are taken into consideration, thus leading to a lower estimation of global confidence. Note that this explanation has not been tested directly in the present study and will require further research.

We thus displayed the existence of the overconfidence paradox in a perceptual task, when participants were overconfident as well as when they were under-confident. We presented a partial explanation to the process of judging global confidence, Finding the overconfidence-paradox in a perceptual task, as in general-knowledge tasks, adds support to the hypothesis that the process underlying confidence judgments are the same regardless of the task at hand.

References

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics, 55*, 412-428.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Griffin, D., & Buehler, R. (1999). Frequency, probability and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology, 38*, 48-78.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411-435.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273.

Olsson, N., & Winman, A. (1996). Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Perception and Psychophysics, 58*, 374-382.

Ronis, D. I., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of the participant matter and assessment method. *Organizational Behavior and Human Decision Processes, 40*, 193-218.

Schneider, S. L. (1995). Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes, 61*, 148-167.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28*, 1-39.

Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46*, 264-282.

Svenson, O. (1985). Cognitive strategies in a complex judgment task: Analyses of concurrent verbal reports and judgments of cumulated risk over different exposure times. *Organizational Behavior and Human Decision Processes, 36*, 1-15.