Psychology Science, Volume 47, 2005 (2), p. 246 - 260

The "false > correct-phenomenon" and subjective confidence: two distinct phenomena influencing response latencies in psychological testing

STEFAN TROCHE[1] & THOMAS RAMMSAYER

Abstract

The "false > correct-phenomenon" refers to the finding that, over a large range of various tasks, response times for correct answers are reliably faster than for false ones. The major goal of the present study was to investigate whether the "false > correct-phenomenon" depends on subjective confidence rather than on objective correctness of a given answer. For this purpose, 87 participants performed two visual discrimination tasks. The participants' task was to decide which of two lines was longer. Task 1 was a two-alternative forced-choice task, whereas Task 2 was a four-level confidence-judgment task. With Task 1, but not with Task 2, correct responses were reliably faster than false ones. However, latencies of responses given with high confidence (Task 2) were significantly faster than those given with low confidence. The overall pattern of results suggests correctness of response and subjective confidence in a response as two distinct factors influencing response latency. Furthermore, evidence has been provided for the restricted universality of the "false > correct-phenomenon".

Key words: Response times, "false > correct-phenomenon", confidence judgment, line-length discrimination

[1] Dr. Stefan Troche, and Prof. Dr. Thomas Rammsayer, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen, Germany; email: stroche@uni-goettingen.de
Correspondence concerning this article should be addressed to Stefan Troche, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen, Goßlerstr. 14, 37073 Göttingen, Germany

Introduction

Commonly accepted advantages of computer-assisted psychological testing and assessment refer to the objectivity, reliability, accuracy, and efficiency these procedures can offer (Butcher, 1987; Fowler, 1985). Another important argument for the use of computer-assisted assessment represents the recording of response times that can also be used for diagnostic purposes (cf., Hornke, 2000). Great potential has been assumed in investigating response times for each single item instead of the time for working on the whole test or a subtest only (e.g., Wildgrube, 1990; Jäger & Krieger, 1994). The diagnostic meaning and significance of response times, however, is still unclear. Test performance and time on task as indicated by response times, obviously do not seem to reflect the same underlying processes. Only very few studies reported a positive correlational relationship between test performance and response time. For example, Hornke (1997) found a correlation of $r_{xy} = .60$ in a computerized adaptive test based on figural matrix type items. On the other hand, Beckmann, Guthke, and Vahle (1997) yielded correlational coefficients between test performance and response time of $r_{xy} = -.11$, $r_{xy} = -.36$, and $r_{xy} = -.52$ for three subtests of an adaptive computer-assisted intelligence learning test battery. Several other studies also failed to show a substantial correlational relationship between performance and response time (Nährer, 1982; Wildgrube, 1990; Rammsayer, 1999; Rammsayer & Brandler, 2003). Furthermore, Rammsayer and Brandler (2003) provided convincing evidence for the notion that response latencies in perceptual and cognitive temporal discrimination tasks are independent of fluid intelligence. Thus, it seems that response time does not represent an additional measure of achievement or task performance but may rather indicate some other personality-related individual differences (cf., Beckmann, 2000; Rammsayer, 1999).

More detailed analyses of response latencies revealed that, across a wide range of different tasks, response times for correct responses or correct answers were reliably faster than for false ones (Beckmann et al., 1997; Beckmann, 2000; Hornke, 1997, 2000; Rammsayer, 1999; Rammsayer & Brandler, 2003). This consistent and highly stable effect has been referred to as "false > correct-phenomenon" (Beckmann, 2000).

Hornke (1997) showed that false answers require about 26% more processing time than correct answers in a computerized adaptive test based on 273 matrix items. He proposed that test takers spend more time on the process of building and evaluating hypotheses about the solution before they give a false answer. Furthermore, a strong correlational relationship of $r_{xy} = .67$ between response latencies of correct and false answers pointed out an individual disposition to respond either fast or slowly (Hornke, 1997). Similarly, Beckmann et al. (1997) analyzed the timing behavior of a large student sample by means of three subtests of an adaptive computer-assisted intelligence learning test battery. With these adaptive tests, correctly answered items were followed by items with higher complexity, whereas incorrectly answered items were followed by specific training items. Thus, test takers were assisted to learn and practice skills they needed to perform successfully on more complex items. Beckmann et al. (1997), however, found a differential effect on response latencies: only high performers showed the "false > correct-phenomenon", while low performers spent the same amount of time on false responses as they did on correct ones. In a subsequent study using three non-adaptive reasoning tests (figure series, number series, and verbal analogies), Beckmann (2000) could confirm the generality of the "false > correct-phenomenon"; i.e., the "false > correct-phenomenon" was observed for all three types of reasoning tasks. Again, however, the more unsatisfactory the test performance the smaller was the difference between response latencies for correct and false answers; a finding that suggests restricted universality of the "false > correct-phenomenon". After ruling out cognitive styles, such as "reflexivity-impulsivity", as a possible explanation for the "false > correct-phenomenon", he put forward the idea that there may be an individual time frame which is limited by the participant's mental capacity. If the duration of mental problem-solving exceeds this time frame, problem solving processes will be discontinued with the consequence of an increased likelihood of false answers.

In perceptual and cognitive discrimination tasks, Rammsayer (1999; Rammsayer & Brandler, 2003) also confirmed the "false > correct-phenomenon". With this type of task, participants were presented with two consecutive auditory intervals within a trial, a constant standard interval and a variable comparison interval. The duration of the comparison interval varied, depending on the participant's previous response, according to an adaptive rule to estimate the individual 75% difference threshold as a psychophysical indicator of performance. Preceding studies showed that the timing mechanism underlying duration discrimination of intervals in the range of milliseconds is highly sensory in nature, while temporal processing of intervals in the range of seconds appears to be mediated by higher cognitive processes (Rammsayer, 1999; Rammsayer & Lima, 1991). With both tasks, latencies for false responses were approximately 25% longer than latencies for correct responses. An additional analysis of response latencies in relation to task difficulty did not reveal any association between both of these aspects of timing behavior. Therefore, longer response latencies for false answers cannot be explained by higher task difficulty. Furthermore, there was no indication of a correlational relationship between response latencies and the personality traits of extraversion, neuroticism, psychoticism, anxiety, and need for achievement (Rammsayer, 1999).

Thus, the available data suggest that, across a wide range of different tasks, such as reasoning and discrimination tasks, the "false > correct-phenomenon" could be demonstrated as a general and robust effect. At the same time, however, one has to state that the origin of this phenomenon is still unclear. Possible explanations for the "false > correct-phenomenon" have been introduced in previous research but, unfortunately, did not hold. For example Beckmann's (2000) findings are in conflict with the assumption that cognitive style may be an influencing factor. Similarly, the alternative proposal of differential time frames with limited capacity (Beckmann, 2000) failed to explain the appearance of the "false > correct-phenomenon" in studies on temporal discrimination (Rammsayer, 1999; Rammsayer & Brandler, 2003). Eventually, it could be shown that the "false > correct-phenomenon" does neither depend on task difficulty (Rammsayer, 1999) nor on the individual level of fluid intelligence (Rammsayer & Brandler, 2003).

An alternative explanation for the "false > correct-phenomenon" may represent participants' subjective confidence with regard to correctness of their answers. In a most recent study on decision-making, Diederich (2003) showed that decision time becomes longer on increasing strength of a response-related conflict. An important reason for a stronger conflict is uncertainty with regard to the outcomes of the decision. Thus, with increasing uncertainty, decision time will increase too. This line of reasoning may also account for longer response latencies for false than for correct answers. False answers may be associated with longer response times since the participant is less confident that his/her answer is correct. If a participant recognizes that he/she has no or no adequate strategy to solve successfully a given test item, a time sacrificing search is being initiated for a possible method how to cope with this situation. At the same time subjective confidence in the answer to be given will decrease, especially if the participant's final answer resulted from guessing. On the other hand, a faster and correct response with high subjective confidence could be expected if the participant had an adequate strategy at his or her disposal for solving successfully the test item.

This line of argumentation implies that, actually, the "false > correct-phenomenon" does not depend on a participant's task-specific objective performance but rather reflects his or her subjective confidence in the given answer. This means, whenever a participant suffers from a lack of confidence in his or her answer, response latencies are expected to increase. Subsequently, rather than the objective criterion of having given a "correct" or "wrong" answer, the "false > correct-phenomenon" is expected to be a function of the experienced feeling of high or low confidence associated with the given answer irrespective of its actual correctness.

Therefore, the major goal of the present study was to test the hypothesis that the "false > correct-phenomenon" depends on subjective confidence rather than on the objective quality (correct vs. wrong) of a given answer. If this hypothesis would be correct, answers given with low confidence should take markedly longer than answers given with high confidence. Furthermore, correct and wrong answers would be expected not to differ, if given with the same level of subjective confidence. In order to test these predictions, visual comparison of the length of two horizontal lines, subsequently presented on a monitor screen, was used as non-adaptive experimental task.

Another purpose of the present study was to further examine the universality of the "false > correct-phenomenon". Most studies confirming the "false > correct-phenomenon" (Beckmann et al., 1997; Hornke, 1997, 2000; Rammsayer, 1999; Rammsayer & Brandler, 2003) applied an adaptive testing strategy. 'Adaptive' means that the difficulty of a test item on any given trial is determined by a preceding set of items and responses. Correct responding results in an increased level of item difficulty for the next item, while incorrect responding makes the next task easier. To our knowledge, the study by Beckmann (2000) was the only study using non-adaptive reasoning tests. In the latter study, only high-, but not low-performers showed reliably faster response latencies for correct than for false answers. This finding has been interpreted as an indication of the restricted universality of the "false > correct-phenomenon" (Beckmann, 2000). By using a non-adaptive experimental task, we may provide additional evidence for the notion that the generality and universality of the "false > correct-phenomenon" is not limited to adaptive testing strategies.

Method

Participants. Participants were 15 male and 72 female undergraduate students ranging in age from 19 to 43 years (mean ∂ standard deviation of age: 22.9 ∂ 4.6 years). The participants were enrolled in introductory psychology courses at the University of Göttingen. Their participation served as partial fulfillment of course requirements. All participants had normal or corrected-to-normal vision.

Apparatus and Stimuli. Stimuli consisted of white lines presented on a black background. The presentation of the lines and the recording of the participants' responses were computer controlled. There was one standard line with a length of 120 pixel (= 58 mm on a 17'' monitor screen) and six comparison lines. The lengths of the comparison lines were 108 pixels (52 mm), 112 pixels (54 mm), 116 pixels (56 mm), 124 pixels (60 mm), 128 pixels (62 mm), 132 pixels (64 mm). The height of the stimulus lines was 1 pixel (= 1 mm on a 17’’ monitor screen). Lines were presented on a 17’’ monitor screen (Acer 7176i).

Procedure. The participants’ task was to decide whether the first or the second horizontal line was longer. There were two types of response format employed in the present study. One type was a two-alternative forced-choice response (“first line longer” and “second line longer”). The other type was a four-level confidence judgment (“first line certainly longer”, “first line likely to be longer”, “second line likely to be longer”, “second line certainly longer”). All participants performed on both types of task. To investigate whether the “false > correct-phenomenon” is or is not related to objective task features, there was an additional experimental condition with both lines within one trial being of the same length. This means, there was no objectively correct answer to these trials. It should be noted, however, that this latter experimental condition was analyzed for the four-level confidence judgment task only. If the “false > correct-phenomenon” is mediated by subjective judgment confidence, high-confidence responses (“first line certainly longer”, “second line certainly longer”) should be faster than low-confidence responses (“first line likely to be longer”, “second line likely to be longer”).

An experimental session consisted of 160 trials; 40 trials for each level of task difficulty. There was a low level of task difficulty with 20 comparison lines being 12 pixels shorter and 20 comparison lines being 12 pixels longer than the standard line. Similarly, comparison lines were 8 and 4 pixels shorter/longer than the standard line for the intermediate and high levels of task difficulty, respectively. At the fourth level of task difficulty, standard and comparison lines were of identical length. Since the participants had to decide which of the two lines was longer, with this latter type of trials, correct responding was not possible, neither with the two-alternative forced-choice task nor with the four-level confidence judgment task.

Each trial consisted of two stimuli, one 120 pixel standard line and one comparison line that was either as long as the standard line or 4 pixels, 8 pixels or 12 pixels shorter or longer than the standard line. Vertically, the lines were presented exactly in the middle of the monitor screen. Horizontally, the standard line was displaced to the right 20 pixels (9.7 mm) from the central position, while the comparison line was displaced 20 pixels to the left. This horizontal displacement was chosen to prevent the participants from using cues that might have made the task easier. At the beginning of each trial, a fixation cross was presented for 400 msec in the middle of the monitor screen. Then the standard line was presented for 400 msec followed by an empty interstimulus interval of 400 msec and the presentation of the comparison line for also 400 msec. The next trial started after the participant indicated his/her response by pressing one of the designated response keys. Comparison lines from all three levels of difficulty were presented in random order.

The participant was seated at a table with a response panel and a computer monitor in front of him/her in a sound-attenuated room. The distance between the monitor screen and the participant was held constant at 60 cm. The participant’s task was to decide which of the two lines was longer and to indicate his/her decision by pressing one of the designated response keys.

With the two-alternative forced-choice task, which had to be performed first, participants were required to press one of two designated response keys (“first line longer”, “second line longer”). With the four-level confidence judgment task, they were required to press one of four designated response keys (“first line certainly longer”, “first line likely to be longer”, “second line likely to be longer”, “second line certainly longer”). For both tasks, experimental trials were preceded by eight practice trials.

As dependent variables, percentage of correct responses and response latencies for correct and false answers were determined. With the four-level confidence judgment task, high- and low-confidence responses were scored as correct answers as long as the participant identified correctly the longer line.

Results

Two-alternative forced-choice task. In a first analysis, we focused on the “false > correct-phenomenon”. In order to examine the “false > correct-phenomenon”, all trials on which the comparison line was equal to the standard line, were excluded from data analysis. The reason for this was that all answers on these trials had to be wrong. Thus, no valid conclusions with regard to the “false > correct-phenomenon” could be drawn on the basis of these trials.

Overall percentage of correct responses for the two-alternative forced-choice task was 78%. For three different levels of task difficulty, the percentages of correct responses were 69%, 79%, and 86% for differences of ∂ 4, ∂ 8, and ∂ 12 pixels, respectively, between the standard and the comparison line. A one-way analysis of variance with task difficulty on three levels of a repeated-measurement factor revealed a statistically significant main effect of task difficulty [$F(2,172) = 149.36$, $p < .001$, effect size $\eta^2 = .64$]. Post-hoc Scheffé’s tests indicated that all levels were significantly different to each other (see Figure 1). This confirms that the physical manipulation of task difficulty was reflected by error rate as an indicator of performance.

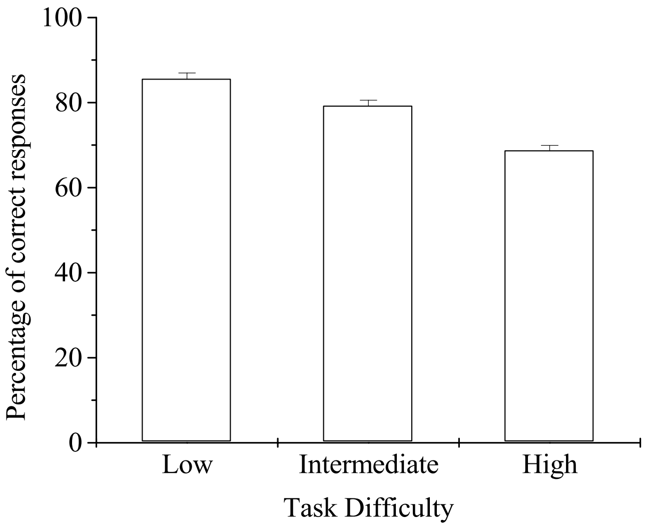

Figure 1:
Mean (± S. E. M.) percentage of correct responses for the two-alternative forced-choice task as a function of task difficulty. All three means differed significantly from each other ($p < .001$).

For statistical analysis of response latencies, a two-way analysis of variance was performed with Task Difficulty (three levels: low, intermediate, high) and Quality of Answer (two levels: correct, false) as two within-subject factors. A significant main effect of Quality of Answer [$F(1,80) = 8.85$, $p < .01$, $\eta^2 = .10$] showed that latencies for false responses were reliably longer than those for correct responses. Mean response latencies (∂ S.E.M.) were 972 ∂ 409 msec and 897 ∂ 213 msec for false and correct responses, respectively. Thus, response latencies for false responses were approximately 8.5% longer than those for correct responses. There was neither a main effect of Task Difficulty [$F(2,160) = .02$, $p = .98$, $\eta^2 = .00$] nor a significant interaction between both repeated-measurement factors [$F(2,160) = 1.44$, $p < .24$, $\eta^2 = .02$]. All simple-main effect means (∂ S.E.M.) are presented in Figure 2.

Also on those trials in which the length of the comparison line was equal to the length of the standard line, participants were required to decide which of the two lines was longer. Thus, under this experimental condition, participants were forced to guess. A one-way analysis of variance was performed with guessed, correct and false response latencies of all trials as three levels of a repeated-measurement factor in order to examine if response latencies of guessed trials were more similar to actually correct or to actually false response latencies. This analysis revealed a statistically significant main effect [$F(1,172) = 7.27$, $p < .01$; $\eta^2 = .12$]; mean latencies (∂ S.E.M.) were 894 ∂ 22 msec and 979 ∂ 33 msec for correct and false answers, respectively, and 941 ∂ 29 msec for guessed answers. Post-hoc Scheffé's tests indicated that there was a significant difference between latencies of false and correct answers ($p < .01$), while latencies of guessed answers differed neither from latencies of correct nor from false answers, respectively.

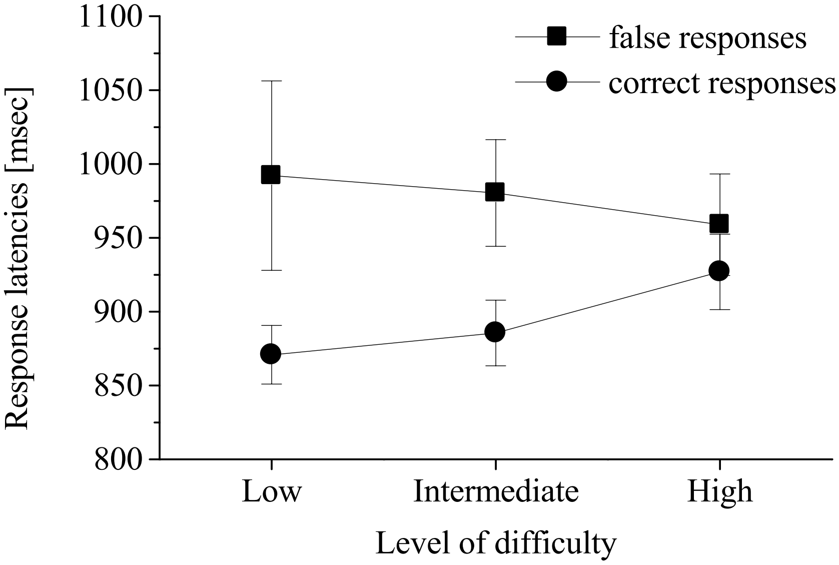

Figure 2:
Means (± S.E.M.) of latencies for correct and false responses in the two-alternative forced-choiced task as a function of task difficulty.

Correlational analysis of the functional relationship between response latencies of correct and false answers yielded an overall coefficient of $r_{xy} = .65$ ($p < .01$). Additional correlational analyses for each level of difficulty resulted in correlational coefficients of $r_{xy} = .22$ ($p = .051$), $r_{xy} = .61$ ($p < .01$), and $r_{xy} = .86$ ($p < .01$) for the low, intermediate, and high level of task difficulty, respectively. Transformation of the correlational coefficients into Fisher's z and subsequent *z* tests (Steiger, 1980) revealed that all three coefficients differed significantly from each other (*z* values ranging from 3.31 to 7.32; all *p* values < .001). This finding clearly indicates a stronger functional relationship with increasing task difficulty.

In Beckmann's (2000) study, only high-, but not low-performers showed a reliable "false > correct-phenomenon". In order to examine whether such a differential effect of individual performance level on latencies for correct and false responses also holds for perceptual tasks, we performed an additional statistical analysis. All participants were divided into two groups according to below and above median scores for correct responses. Mean latencies of correct and false responses were 928 msec and 1,020 msec for high performers, and 864 msec and 943 msec for low performers. Two-way analysis of variance with latencies of correct and false answers as two levels of a repeated-measurement factor (Quality of Answer) and high and low performers as two levels of a between-subjects factor (Level of Performance) yielded a significant main effect of Quality of Answer [$F(1,82) = 10.82$; $p < .01$; $\eta^2 = .12$]. However, there was neither a statistically significant difference between high and low performers [$F(1,82) = 1.89$; $p = .17$; $\eta^2 = .02$] nor a significant interaction between both factors [$F(1,82) = .07$; $p = .80$; $\eta^2 = .00$]. This failure to demonstrate a significant interaction argues against the assumption that the "false > correct-phenomenon" is more pronounced in high compared to low performers.

Four-level confidence judgment task. Overall task difficulty as indicated by percentage of correct responses was 76%. As with the two-alternative forced-choice task, experimental manipulation of three levels of difficulty proved to be successful; mean percentages of correct answers were 66%, 79%, and 84% for the ∂ 4-, ∂ 8-, and ∂ 12-pixel condition, respectively. Also one-way analysis of variance yielded a statistically significant effect [$F(2,172) = 143.33$; $p < .001$, effect size $\eta^2 = .63$]. Scheffé's tests revealed that all three levels were significantly different from each other (see Figure 3).

A three-way analysis of variance with Task Difficulty (three levels: low, intermediate, high), Quality of Answer (two levels: correct, false), and Confidence Judgment (two levels: low, high) as three within-subject factors revealed a significant main effect of Confidence Judgment [$F(1,45) = 15.46$; $p < .001$, $\eta^2 = .26$]. Responses given with high confidence were reliably faster than low-confidence responses; mean latencies were 1292 ∂ 419 msec and 1476 ∂ 577 msec for high- and low-confidence responses, respectively. There were neither significant main effects of Task Difficulty [$F(2,90) = 2.03$; $p = .14$, $\eta^2 = .04$] and Quality of Answer [$F(1,45) = 1.02$; $p = .32$, $\eta^2 = .02$] nor a significant interaction between Task Difficulty and Quality of Answer [$F(2,90) = .20$; $p = .82$, $\eta^2 = .004$; see Figure 4], Task Difficulty and Confidence Judgment [$F(2,90) = 1.1$; $p = .34$, $\eta^2 = .02$; see Figure 5], or Quality of Answer and Confidence Judgment [$F(2,45) = 1.1$; $p = .30$, $\eta^2 = .02$; see Figure 6]. Similarly, the three-way interaction of all three within-subject factors combined failed to reach statistical significance [$F(2,90) = .93$; $p = .40$, $\eta^2 = .02$]. To sum up, no reliable "false > correct-phenomenon" could be identified with the four-level confidence judgment task. Instead, latencies of responses given with low confidence were 14% longer than those of responses given with high confidence no matter whether the responses were right or wrong.

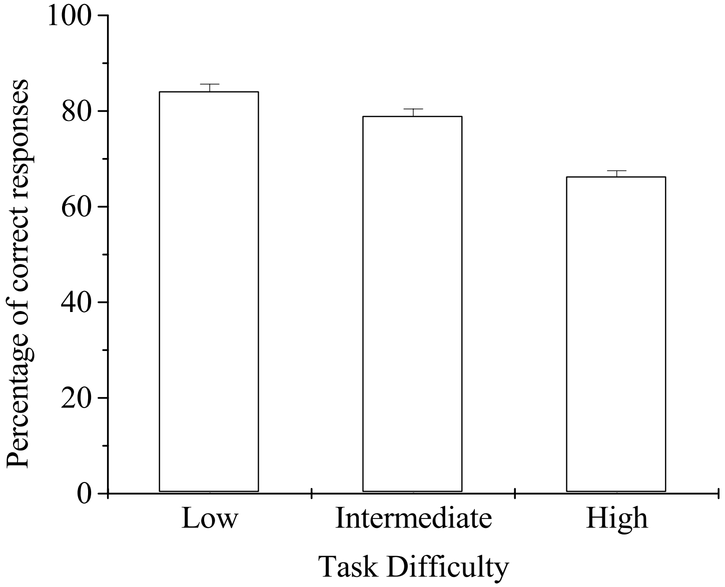

Figure 3:
Mean (± S.E.M.) percentage of correct responses for the four-level confidence judgment task as a function of task difficulty. All three means differed significantly from each other (p < .001).

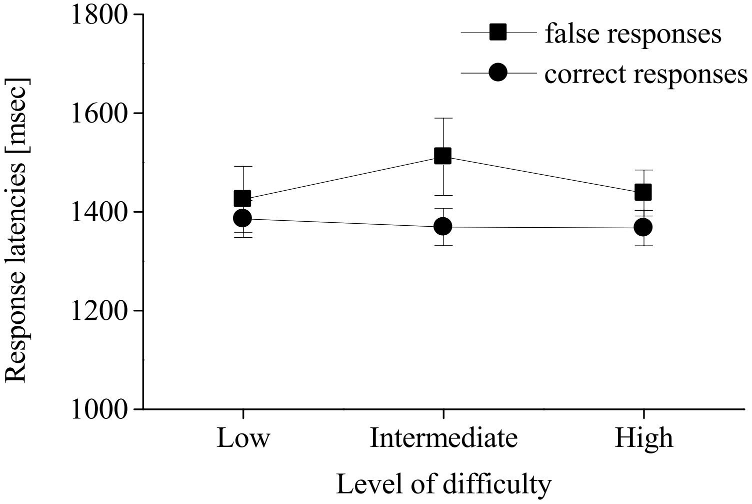

Figure 4:
Means (± S.E.M.) of latencies for correct and false responses as a function of task difficulty in the four-level confidence judgment task.

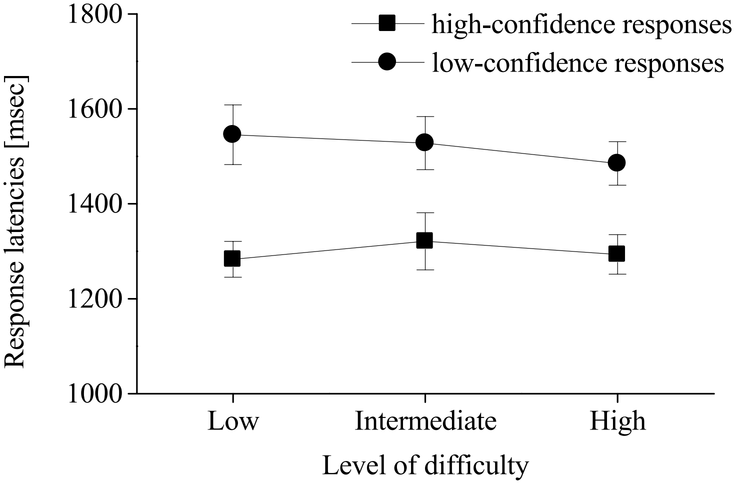

Figure 5:
Means (± S.E.M.) of latencies for responses with high and low confidence as a function of task difficulty.

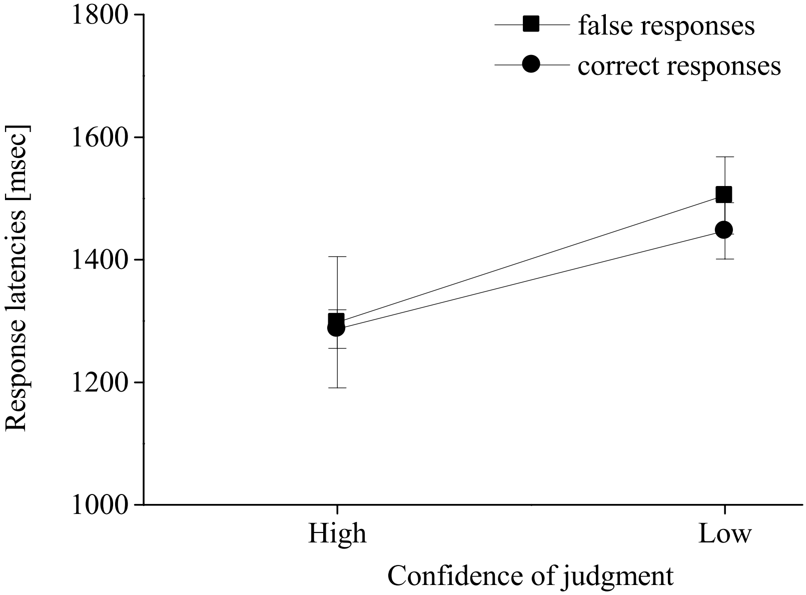

Figure 6:
Means (± S.E.M.) of latencies for correct and false responses as a function of subjective confidence.

Correlational analysis yielded an overall coefficient of $r_{xy} = .67$ ($p < .01$) between latencies for correct and false responses. Additional analyses revealed that for responses given with high subjective confidence, the correlation between correct and false responses was $r_{xy} = .33$ ($p < .01$), while for low-confidence responses the resulting correlation coefficient was $r_{xy} = .85$ ($p < .001$). The difference between both these coefficients was shown to be highly significant ($z = 6.28$, $p < .001$). A similar differential correlational relationship was observed for the association between high- and low-confidence responses as a function of quality of answer. The correlation coefficients between latencies for high- and low-confidence responses were $r_{xy} = .69$ ($p < .01$) and $r_{xy} = .29$ ($p < .01$) for correct and false responses, respectively. Also these coefficients differed significantly from each other ($z = 3.94$, $p < .001$).

For the four-level confidence judgment task, also trials were analyzed in which the length of the comparison line was equal to the length of the standard line. Because participants were required to perform a longer/shorter judgment on each trial and because, at the same time, both lines were virtually of identical length, no correct responses could be given under this experimental condition. Therefore, task difficulty as indicated by percentage of correct responses was not considered to represent an appropriate measure. For this reason, only latencies for high- and low-confidence responses were submitted to data analysis. Mean latencies were 1321 ∂ 382 msec and 1479 ∂ 464 msec for high- and low-confidence responses, respectively. A *t*-test revealed that low-confidence responses were performed significantly slower than high-confidence responses [$t(83) = 4.05$, $p < .001$, $d = 0.29$].

As for Task 1, a further statistical analysis was performed to compare the magnitude of the "false > correct-phenomenon" in high- and low-performers. For high-performers, mean latencies of correct and false high-confidence responses were 1,258 msec and 1,497 msec, respectively, and 1,567 msec and 1,781 msec, respectively, for low-confidence responses. For low performers, on the other hand, mean latencies of correct and false high-confidence responses were 1,288 msec and 1,283 msec, respectively, and 1,401 msec and 1,356 msec, respectively, for low-confidence responses. A three-way analysis of variance was performed with Quality of Answer (correct and false) and Confidence Judgment (high and low) as two repeated-measurement factors, and Level of Performance (high and low) as a between-subjects factor. Main effects of both Level of Performance [$F(1,77) = 3.84$; $p = .054$; $\eta^2 = .05$] and Quality of Answer [$F(1,77) = 3.43$; $p = .07$; $\eta^2 = .04$] just failed to reach the 5% level of statistical significance. However, a significant interaction of both these factors could be shown [$F(1,77) = 5.33$; $p < .05$; $\eta^2 = .07$]. Post-hoc Scheffé's tests revealed reliably shorter latencies for correct (1,412 msec) than for false (1,639 msec) responses in high performers ($p < .05$). No such difference was found in the low performing group; mean response latencies were 1,344 msec and 1,319 msec for correct and false answers, respectively ($p = .99$). Also the main effect of Confidence Judgment became significant [$F(1,77) = 10.48$; $p < .01$; $\eta^2 = .12$]; mean latencies were 1,331 msec for high-confidence judgments and 1,526 msec for low-confidence judgments. There was no significant interaction, neither between Confidence Judgment and Level of Performance [$F(1,77) = 2.87$; $p = .09$; $\eta^2 = .04$] nor between Quality of Answer and Confidence Judgment [$F(1,77) = .85$; $p = .77$; $\eta^2 = .00$] nor among all three factors combined [$F(1,77) = .01$; $p = .94$; $\eta^2 = .00$].

In an additional analysis, latencies of guessed answers were compared with mean latencies of correct and false answers given with low and high confidence, respectively. A two-way analysis of variance with Confidence Judgment (two levels: low, high) and Quality of Answer (three levels: correct, false, guessed) as two within-subject factors revealed a statistically significant main effect of Confident Judgment [$F(1,80) = 17.97$, $p < .01$, $\eta^2 = .18$]. Responses given with high confidence were reliably faster than responses given with low confidence; mean response times were 1,276 msec and 1,479 msec for high- and low-confident guessed answers, respectively. There was neither a significant main effect of Quality of Answer [$F(2,160) = 2.95$, $p = .06$, $\eta^2 = .04$] nor a significant interaction between both factors [$F(2,160) = .18$, $p = .84$, $\eta^2 = .00$]. The lack of a statistically significant main effect of Quality of Answer indicates that latencies of guessed answers were not reliably different from correct or false answers.

Discussion

The general validity of the "false > correct-phenomenon" has been confirmed in Task 1. With the two-alternative forced-choice task, correct judgments of line length were reliably faster than false ones. Obviously, length comparison of visually presented lines represents a mainly perceptual task. Therefore, the outcome of Task 1 also provided additional experimental evidence for the notion that the "false > correct-phenomenon" is not limited to more complex cognitive tasks. Such a conclusion is consistent with the outcome of previous studies establishing a "false > correct-phenomenon" for perceptual auditory temporal discrimination tasks (Rammsayer, 1999; Rammsayer & Brandler, 2003).

Unlike previous studies (e.g., Beckmann et al., 1997; Hornke, 1997, 2000; Rammsayer, 1999; Rammsayer & Brandler, 2003), a non-adaptive experimental procedure was applied in Task 1. Thus, the finding of longer response latencies for false than for correct responses points to the conclusion that occurrence of the "false > correct-phenomenon" is independent of the assessment procedure applied. Beckmann (2000), who used non-adaptive reasoning tests, also reported longer response latencies for false than for correct answers. In a more detailed analysis, however, Beckmann (2000) revealed that only high-, but not low-performers showed a reliable "false > correct-phenomenon". In Task 1 of the present study, however, there was no indication of such differential effect. Rather, a reliable main effect of level of performance revealed longer response latencies for high- than for low-performers. This difference in response time held for both correct and false responses. The finding of longer response latencies for high- compared to low-performers suggests that the latter group devoted less time to the line discrimination task which, in turn, could have resulted in poorer discrimination performance. High-performers, on the other hand, spent more time on task and, thus, achieved better discrimination performance.

With Task 2, using a four-level confidence judgment rather than a two-alternative forced-choice mode, there was no indication of a "false > correct-phenomenon". Latencies for both false and correct responses did not differ significantly. In both tasks, the stimuli to be compared and the experimental procedure were absolutely identical, except that in Task 2 the participants were required to give a confidence judgment rather than a two-alternative forced-choice response. Therefore, the reason for the failure to replicate the "false > correct-phenomenon" with Task 2 should be attributed to the different response formats in both tasks. Furthermore, since previous studies (Beckmann, 2000; Beckmann et al., 1997; Hornke, 1997), using forced-choice tasks with more than two response alternatives did show a reliable "false > correct-phenomenon". Therefore, differences in response alternatives between Task 1 and Task 2 cannot account for the absence of the "false > correct-phenomenon” in the latter task. Rather than differences in the number of response alternatives, qualitative differences between both response modes appear to be responsible for the disappearing of the “false > correct-phenomenon” in Task 2. The involvement of at least partly different cognitive mechanisms in both types of tasks may also be indicated by an increase in overall response latencies of more than 50% with the confidence judgment task compared to the two-alternative forced-choice task. With the two-alternative forced-choice task (Task 1), a participant had only to decide which of the two presented lines was longer. With the four-level confidence judgment task (Task 2), however, in addition to the decision on line length, the participant was also required to judge his or her subjective confidence in the line length decision. This latter component provided a source of additional variance that may have masked genuine task-specific variance. Therefore, if the “false > correct-phenomenon“ is associated with task-specific variance, this may account for the failure to demonstrate the “false > correct-phenomenon“ with the confidence judgment task.

Most interestingly, however, a different phenomenon also related to response latencies could be identified in Task 2. Latencies of responses given with high confidence were reliably faster than those of responses given with low confidence judgment. This effect was shown to be independent of the level of task difficulty and of the response correctness. Even more surprisingly, if there was no possibility to give an objectively correct answer because both lines were virtually identical in length, high-confidence responses were also faster than those of low-confidence. Obviously, these longer latencies for low-confidence responses cannot be explained in terms of physical stimulus characteristics or task difficulty. Thus, the origin of this “low > high-confidence phenomenon” seems to be located somewhere within the decision process. This implies that the “low > high-confidence phenomenon” is unrelated to the perceptual processing of the stimuli to be compared and, more likely, represents an effect of higher cognitive processes associated with decision-making (c.f., Juslin & Olsson, 1997; Vickers & Pietsch, 2001).

A major finding of the present study represents the emergence of two different phenomena, a “false > correct-phenomenon” and a “low > high-confidence phenomenon”. A mere change of the response format from a two-alternative forced-choice mode to a four-level confidence judgment eliminated the “false > correct-phenomenon” and generated the “low > high-confidence phenomenon” instead.

The relationship between both these phenomena still remains unclear. A tentative explanation suggests that the “false > correct-phenomenon” may reflect a specific aspect of the more general “low > high-confidence phenomenon”. This idea proceeds from the assumption that false answers are given with less subjective confidence than correct answers. Consequently, the longer response latencies associated with false answers merely reflect lower subjective confidence. To date, no studies appear to exist using a hybrid procedure to obtain independent measures of both response formats. Nevertheless, this latter hypothesis can be tested indirectly by comparing latencies for correct and false responses obtained by means of an adaptive procedure. Let us assume that subjective confidence decreases with increasing task difficulty. Then, false answers are also associated with the lowest level of subjective confidence, whereas a reduction of task difficulty should result in increased subjective confidence. If longer latencies for false answers merely reflect lower subjective confidence, in an analysis of individual *adaptive-testing* sequences the following picture should emerge: 1) Tasks not correctly answered, because they were too difficult for a given participant, should be associated with the lowest subjective confidence and slowest response latency. 2) Task difficulty close to a participant’s individual performance limit, but still within the participant’s competence range, should be answered correctly and should be accompanied by higher subjective confidence and faster response latencies as compared to tasks he or she failed to perform successfully. 3) Very easy tasks should produce the highest level of subjective confidence and the fastest response latencies.

In a previous study applying an adaptive psychophysical procedure, Rammsayer (1999) compared latencies for three levels of task difficulty in a sample of 120 subjects on two temporal discrimination tasks. Analysis of both tasks revealed slowest response latencies for highly difficult tasks incorrectly answered by the participants. On the other hand, both extremely difficult as well as extremely easy tasks which were correctly answered, resulted in almost identical response latencies. This finding clearly argues against the notion that the “low > high-confidence phenomenon” represents a generalized form of the “false > correct-phenomenon”. If this notion is correct, differences in response latencies should proportionally vary as a function of task difficulty which was not the case in the present study. Additional evidence against this notion can be derived from the results of our four-level confidence judgment task. Although 71% of all incorrect answers were given with low confidence but only 48.3% of correct answers, an analogous, reliable difference was not found for response latencies. Furthermore, statistical analyses did not support the notion that highest level of subjective confidence and shortest response latencies should be observed with the easiest task.

Eventually, an additional three-way analysis of variance was conducted to assess whether the observed “low > high-confidence phenomenon” was moderated by performance level. Another aim of this analysis was to compare the magnitude of the “false > correct-phenomenon” in high and low performers. No evidence was found for a moderating effect of performance level on the “low > high-confidence phenomenon”. As opposed to Task 1 of the present study, in Task 2, level of performance effectively modulated the “false > correct-phenomenon”; high performers showed a reliable “false > correct-phenomenon”, which was virtually absent in the low performing group. The identified “false > correct-phenomenon” in high- but not low-performers provides converging evidence for Beckmann’s (2000; Beckmann et al., 1997) conclusion that performance level modulates effectively the “false > correct-phenomenon”. Furthermore, this latter finding points to the restricted universality of the “false > correct-phenomenon” since low performers appear to show no or only a marginal “false > correct-phenomenon”. Nevertheless, it remains unclear why such a differential effect of performance level on the “false > correct-phenomenon” was not observed with Task 1. While the “false > correct-phenomenon” observed with Task 2 was much more pronounced in high than in low performers, the “low > high-confidence phenomenon” could be observed in both groups. This lack of a performance-dependent differential effect on the “low > high-confidence phenomenon” suggests that both phenomena do not reflect the same underlying cognitive processes.

References

1. Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. Journal of Educational Psychology, 32, 285-296.

2. Beckmann, J. F. (2000). Differentielle Latenzzeiteffekte bei der Bearbeitung von Reasoning-Items. Diagnostica, 46, 124-129.
3. Beckmann, J. F., Guthke, J. & Vahle, H. (1997). Analysen zum Zeitverhalten bei computergestützten adaptiven Intelligenz-Lerntests. Diagnostica, 43, 40-62.
4. Butcher, J. N. (1987). The use of computers in psychological assessment: An overview of practices and issues. In J. N. Butcher (Ed.), Computerized psychological assessment: A practioner's guide (pp. 3-14). New York: Basic Books.
5. Diederich, A. (2003). Decision-making under conflict: Decision time as a measure of conflict strength. Psychonomic Bulletin & Review, 10, 167-176.
6. Fowler, R. D. (1985). Landmarks in computer-assisted psychological assessment. Journal of Consulting and Clinical Psychology, 53, 748-759.
7. Hornke, L. F. (1997). Untersuchung von Itembearbeitungszeiten beim computergestützten adaptiven Testen. Diagnostica, 43, 27-39.
8. Hornke, L. F. (2000). Item response times in computerized adaptive testing. Psychologia – Revista de Metodologia y Psycologia Experimental, 21, 175-189.
9. Jäger, R. S. & Krieger, W. (1994). Zukunftsperspektiven der computerunterstützten Diagnostik, dargestellt am Beispiel der treatmentorientierten Diagnostik. Diagnostica, 40, 217-243.
10. Juslin, P. & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. Psychological Review, 104, 344–366.
11. Nährer, W. (1982). Zur Beziehung zwischen Bearbeitungsstrategie und Zeitbedarf bei Denkaufgaben. Zeitschrift für experimentelle und angewandte Psychologie, 24, 147-159.
12. Rammsayer, T. (1999). Zum Zeitverhalten beim computergestützten adaptiven Testen: Antwortlatenzen bei richtigen und falschen Lösungen. Diagnostica, 45, 178-183.
13. Rammsayer, T. & Brandler, S. (2003). Zum Zeitverhalten beim computergestützten adaptiven Testen: Antwortlatenzen bei richtigen und falschen Lösungen sind intelligenzunabhängig. Zeitschrift für Differentielle und Diagnostische Psychologie, 24, 57-63.
14. Rammsayer, T. H. & Lima, S. D. (1991). Duration discrimination of filled and empty auditory intervals: Cognitive and perceptual factors. Perception & Psychophysics, 50, 565-574.
15. Vickers, D. & Pietsch, A. (2001). Decision-making and memory: A critique of Juslin and Olsson's (1997) sampling model of sensory discrimination. Psychological Review, 108, 789-804.
16. Wildgrube, W. (1990). Computergestützte Diagnostik in einer Großorganisation. Diagnostica, 36, 127-147.