Psychology Science, Volume 47, 2005 (2), p. 230 - 245

Accuracy, latency, and confidence in abstract reasoning: the influence of fear of failure and gender

FRANZIS PRECKEL[1] & PHILIPP ALEXANDER FREUND[2]

Abstract

Aims: For many cognitive tasks, participants take longer to make mistakes than to answer correctly. Known as the false>correct (i.e., F>C)-phenomenon, effects have been replicated for both adaptive and non-adaptive tests. Support for a choice-accuracy heuristic comes from the unrelated observation that latency appears more strongly related to confidence than to accuracy. Bridging various fields of research in the present study, it was predicted that latencies for answers with high confidence are shorter than latencies for low confidence responses. *Method:* Students (N=103) were tested with a non-adaptive computer-assisted figural matrices test. Participants gave confidence ratings on the correctness of each response. *Results:* The F>C-phenomenon was replicated, though there were no differential effects of ability level. In addition, confident responses had shorter latencies than responses given with low confidence. Of note, confidence explained a small amount of variance in response latencies when accuracy was controlled, although gender and fear of failure both explained variance in confidence ratings (independent of latency, score, or motivational variables). *Conclusion:* The results support the conceptualization of confidence as a personality trait that is influenced by answer accuracy, gender, and fear of failure.

Key words: response latencies, false>correct-phenomenon, confidence ratings, abstract reasoning tasks, gender differences

[1] Dr. Franzis Preckel, Department of Psychology, Ludwig-Maximilians-Universität München, Germany; email address: preckel@psy.uni-muenchen.de

[2] Dipl.-Psych. Philipp Alexander Freund, Psychologisches Institut IV, Westfälische Wilhelms-Universität Münster, Germany

Correspondence concerning this article should be addressed to F. Preckel at Department of Psychology, Ludwig-Maximilians-Universität, Leopoldstraße 13, D-80802 München, Germany

Accuracy, latency, and confidence in abstract reasoning: the influence of fear of failure and gender

Throughout a variety of non-speeded psychometric tasks, response latencies are longer for false than for correct responses. This finding has been referred to as the false>correct-phenomenon (henceforth referred to as the F>C-phenomenon; Beckmann, 2000). Response latencies are defined as the time from stimulus onset to answer execution by a person completing a psychometric test. The F>C-phenomenon has been replicated using both adaptive (Beckmann, Guthke, & Vahle, 1997; Hornke, 1997, 2000; Rammsayer, 1999; Rammsayer & Brandler, 2003) and non-adaptive (Beckmann, 2000; Ebel, 1953) test forms. Item types under investigation have been disparate, including processes associated with verbal, figural, and numerical reasoning (Beckmann, 2000; Beckmann, Guthke, & Vahle, 1997; Hornke, 1997, 2000), knowledge (Ebel, 1953; Zakay & Tuvia, 1998), and perceptual discrimination (Rammsayer, 1999; Rammsayer & Brandler, 2003; Ratcliff & Rouder, 1998) tests. Thus, the F>C-phenomenon appears to be rather general for complex and basal tasks, suggesting that reasoning-specific explanations are not particularly compelling (Beckmann, 2000) and that different explanatory models are needed for tasks of varying complexity. Even so, until now the F>C-phenomenon has solely been investigated in studies where the participants had been instructed to maximize accuracy under power conditions. Thus, instructions (and test settings) have not required participants to trade-off speed and accuracy, which influences the relationship between mean response time and probability of a correct response[3].

So far, a sufficient explanatory model for error response times (Luce, 1986) and for the F>C-phenomenon has not been forthcoming. With the diffusion-model, Ratcliff (1978, 1981, 1988) offers a statistical model for choice reaction times that models latencies of errors and correct responses for tasks on which response time is under a second (e.g., Grosjean, Rosenbaum, & Elsinger, 2001; Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999). However, given the complex and multiple decisions required in reasoning tasks of the preceding type and the resulting longer response latencies this model appears unsuitable. According to Hornke (1997), for more complex tasks the F>C-phenomenon might be explained by processes of testing a large amount of hypotheses preceding errors, though no model has yet been provided to test this assumption.

Potential moderator variables for the relation between answer accuracy and latency are the ability of the participant or the difficulty of a task, respectively. Beckmann (2000) formulated the hypothesis that low ability participants may give wrong answers faster than their counterparts (due to motivational factors) and correct responses slower (because of a relatively higher cognitive load that the task puts on them). However, research on the universality of the F>C-phenomenon has revealed heterogeneous results for different ability groups. Using adaptive, computer-aided learning ability tests, Beckmann, Guthke, and Vahle (1997) reported the F>C-phenomenon only for high scoring participants. For non-adaptive reasoning tests Beckmann (2000) replicated the F>C-phenomenon for all ability groups but it also turned out that the differences in latencies for errors and correct responses increased with ability level. In adaptive test settings, Hornke (1997; figural matrices) and Rammsayer (1999; sensory discrimination tasks) found no differential effects in 'false' or 'correct' laten-

[3] In the present study the term "accuracy" will be used to describe the correctness of a response to a single task. The term "score" will be used to refer to the percentage of items answered correctly.

cies for different performance groups. Among other explanations, the differences in findings might be due to differences in test administration (e.g., adaptive vs. non-adaptive). As another potential moderator variable for the relationship between answer accuracy and latency the certainty with which answers are given was investigated in the present study.

Confidence ratings

The level of certainty (or uncertainty) with which answers are given can be assessed by confidence ratings (Stankov & Crawford, 1997). In the confidence rating methodology, participants are asked to judge the degree of accuracy of their own performance for every single item in the course of working through the test (which is different from post test confidence ratings for the test as a whole or from prospective confidence ratings which are given after seeing a task but before making a choice). Usually participants give their confidence ratings on a scale in percentages. Various studies have revealed the existence of a confidence factor that can be interpreted as an independent metacognitive trait which mediates the accuracy of self-assessment (see e.g., Pallier et al., 2002; Schraw, 1994, 1997; Stankov & Crawford, 1997). Even so, the validity of this confidence factor is still uncertain. Thus, Stankov and his co-workers found no correlations between confidence scores and extraversion, neuroticism, psychoticism (Crawford & Stankov, 1996, Stankov & Crawford, 1996, 1997), or between confidence scores and introversion, self-monitoring, and task specific measures of academic self-concept (when controlling for the variance of the confidence ratings attributable to accuracy; Pallier et al., 2002; Stankov & Crawford, 1997). Research has shown that participants are quite successful in estimating their accuracy by confidence ratings when working on figural matrices tasks (May, 1986; Pallier et al., 2002; Stankov & Crawford, 1996, 1997). Mean correlations over various tasks between expressed confidence in the correctness of an answer and performance are positive and around $r = .50$ (Stankov & Crawford, 1997).

However, Zakay and Tuvia (1998) reported only weak correlations between accuracy and confidence when response latency had been partialled out ($r = .13$ and $r = .14$, respectively). Nevertheless, they found moderate correlations between confidence and latency which were independent of response accuracy ($r = -.36$ and $r = -.32$). Latencies were shorter for responses given with high confidence. These results were stable over tasks of fluid (i.e., visual perceptual) as well as crystallized intelligence (verbal knowledge) tasks. Zakay and Tuvia (1998) concluded that response latencies are more closely related to confidence than to answer accuracy. They assumed that "confidence is partly due to factors which have high impact on feelings of confidence, but are not related to choice processes which determine choice accuracy." (p. 104). As one factor influencing confidence but not accuracy they discussed perceived mental effort in the context of a choice latency heuristic. The faster a decision is made, the more confident people feel about the correctness of the decision. This is reasonable because latencies are positively related to mental effort and perceived mental effort is negatively related to confidence (Wickens, 1992; Zakay & Tzal, 1993).

To shed more light on the relationships between confidence, latency, and answer accuracy, we re-analyzed some of the data that Stankov and Crawford (1997) collected on these three variables with the Ravens Progressive Matrices (RPM; Raven, 1936, 1962). As a result of this re-analysis we found a pattern that seems to be the opposite of what Zakay and Tuvia

(1998) found: The partial correlation between confidence and score, when controlling for latency was r = .42 (vs. r = .51 non-partialled), between confidence and latency when controlling for accuracy was r = .09 (vs. r = .32), and between score and latency when controlling for confidence was r = .40 (vs. r = .49). Thus, in the data of Stankov and Crawford (1997) the relationship between score and latency was relatively independent of confidence while there was hardly any relationship between confidence and latency when partialing out accuracy.

But it has to be taken into account that in the study of Stankov and Crawford (1997) other than in the study of Zakay and Tuvia (1998) there was no F>C-phenomenon: In the study of Stankov and Crawford (1997) people with higher scores in the RPM had longer average response times (r = .49). Previous research documented hardly any relationship between score and time spent on the RPM or comparable tests (see e.g., Jensen, 1998; Knorr & Neubauer, 1996; Preckel & Thiemann, 2003). Hence, the positive relationship between average response time and score which is documented in the study of Stankov and Crawford (1997) might not be representative.

Gender differences in confidence ratings. Recent research has indicated that gender is another factor with high impact on feelings of confidence but no relation to choice accuracy (e.g., Beyer, 1998; Milto, Rogers, & Portsmore, 2002; Pallier, 2003). Unrelated to accuracy, Pallier (2003) reported higher confidence ratings for men than for women in figural matrices tasks, which require abstract reasoning ability (Carpenter, Just, & Shell, 1990; Preckel, 2003). There are no gender differences in the aptitude to solve matrix completion tasks (e.g., Lynn, Backhoff, & Contreras-Nino, 2004; Mills & Tissot, 1995; Preckel, 2003), or in abstract reasoning ability (e.g., Halpern & LaMay, 2000).

Recent research has documented that women tend to estimate their abstract reasoning ability lower than men do (Bennett, 2000; Furnham, Clark, & Bailey, 1999; Furnham, Fong, & Martin, 1999; Rammstedt & Rammsayer, 2000, 2001, 2002). This finding bears no relation to their true level of ability and to the accuracy of their ability estimate (Holling & Preckel, 2005; Pallier, 2003). Similar results emerged when subjects had to estimate the ability of others (e.g., family members): Men regularly were judged more able than women in the domain of abstract reasoning (Rammstedt & Rammsayer, 2000). Thus, confidence ratings and estimates of one's own intellectual ability seem to be influenced by differences in self-perceptions that are caused by gender stereotypes (Beloff, 1992; Beyer, 1998; Pallier, 2003).

Aims of this study

1. The F>C-phenomenon appears to be a rather general phenomenon. Up to now it has not been studied explicitly with a non-adaptive figural matrices test. However, Stankov and Crawford (1997), applying the RPM, report a rather strong tendency for people with higher scores to have longer average response times. Therefore, one aim of the present study was to show that the F>C-phenomenon could be replicated in a non-adaptive test setting using a figural matrices test.
2. Previous studies revealed heterogenous results on the universality of the F>C-phenomenon for different ability groups. Hence, the F>C-phenomenon was investigated in different ability groups.

3. In the study of Zakay and Tuvia (1998) response latencies were more closely related to confidence than to accuracy. Response latencies for answers given with high confidence were shorter than response latencies for answers given with low confidence, independent of the correctness of response. Therefore, the contributions of confidence and answer accuracy to the explanation of latencies were to be analyzed.
4. Stankov and Crawford (1997) conceptualized confidence as a personality trait associated with motivational variables. Moreover, gender acts as a moderating variable in confidence ratings, with women giving lower ratings than men. In this study it was examined if motivational variables and gender added significantly to the explanation of confidence after controlling for answer accuracy and latency.

Method

Participants. A total of 103 participants were included in the analysis, ranging from 15 to 48 years (M = 22.74, SD = 6.38). Sixty-nine percent of the sample were female. Participants were undergraduate students from the University of Münster (72%), high-school students (17%) in tenth (n = 7), eleventh (n = 8), and twelfth grade[4] (n = 4), as well as students in second-chance education[5] (11%). They were recruited by poster advertisements and received no financial rewards. Feedback of results was offered if desired.

Measures. Test material comprised two tests for the assessment of abstract reasoning ability, one test for the assessment of concentration ability, one questionnaire for the assessment of current motivation, and one scale for the assessment of conscientiousness.

Abstract reasoning ability. For the assessment of abstract reasoning ability two figural tests were applied. The first test was a newly developed computer-based figural matrices test (Münsteraner Matrizentest, MMT-1; for details see Freund, 2003), which consisted of 34 items. Item construction was based on the findings of numerous studies and included five different construction rules as well as various drawing features (e.g., Carpenter, Just, & Shell, 1990; Embretson, 1998; Preckel, 2003; Vodegel Matzen, van der Molen, & Dudnik, 1994). Each item had nine answer options: The correct solution, seven distractors, and the option ‘no correct alternative’ (which was implemented in order to prevent exclusion strategies; Gittler, 1989). The distractors were built systematically and contained one or more rule omissions (1, ... a - 1 with a = number of rules).

As a second measure for abstract reasoning ability a German adaptation (Weiß, 1998) of the Culture Fair Intelligence Test (Cattell, 1960) – the CFT-20 – was applied. Actuality of norm data for this test was last controlled in 1996. The CFT-20 contains four types of figural tasks which are series, classifications, matrices, and topologies. Tasks are answered in a multiple-choice format. The test was presented as a paper-and-pencil version. Testing took place under speeded power conditions with generous time limits.

Concentration ability. Concentration ability was assessed with the test d2 (Brickenkamp, 2002). The d2 consists of a standardized sheet in a landscape layout of 14 test lines with 47 characters in each line. Each character is either the letter 'd' or 'p' marked with one, two,

[4] Students came from a German Gymnasium that requires up to thirteen years of schooling.

[5] These are students who accomplish their Abitur (which is the final exam of the German Gymnasium) outside of the Gymnasium at special schools.

three or four small dashes. The respondent's task is to scan the lines and cross out all 'd's with two dashes while ignoring all other characters. Scoring keys for different aspects of concentration ability are provided. The test was last normed in 2002. Testing took place under speed conditions.

Current motivation. The German version of the Questionnaire on Current Motivation (QCM; Rheinberg, Vollmeyer, & Burns, 2001) uses 18 items to measure four motivational factors in achievement situations: challenge (4 items), interest (5 items), fear of failure (5 items), and probability of success (4 items). The questionnaire was applied in a paper-and-pencil format and answers were given on 7-point-Likert scales.

Conscientiousness. Conscientiousness was assessed by using the conscientiousness scale (12 items) from the German version of the NEO-Five Factor Inventory (NEO-FFI; Costa & McCrae, 1989, 1992; German translation: Borkenau & Ostendorf, 1993). The scale was applied in a paper-and-pencil format. Answers were given on 5-point-Likert scales.

Procedure. Participants were tested in groups of up to 14 persons on one occasion. Testing took about 200 minutes with rest pauses included. Due to organizational conditions, the tests were applied in two different sequences (sequence 1: MMT-1, 10 min. break, NEO-FFI-scale, CFT-20, d2, n = 60; sequence 2: NEO-FFI-scale, CFT-20, d2, 10 min. break, MMT-1, n = 43). Participants were randomly assigned to conditions. Two experimenters were present throughout the entire testing session. Instructions for the CFT-20, the test d2, and the NEO-FFI-scale were given verbally by one experimenter and were also written on the test material.

Participants were instructed in a standardized way how to handle the computer-based test. Experimenters ensured that all questions regarding the test and the test taking procedure were answered. Instructions for the MMT-1 were presented on the screen. Before starting the 34 test-items, participants worked for a maximum of 30 minutes on eight practice items which explained all rules in detail and provided exhaustive training. After completion of the practice items and immediately before starting the MMT-1 participants answered the questionnaire on current motivation (QCM). Instructions for the QCM were given in a written format with the questionnaire. There were no time constraints for answering the questionnaire. The test items of the MMT-1 were arranged in two test parts with 17 items each and a break of 15 minutes in between. For each part, a maximum time of 50 minutes was assigned. After each item, participants were required to indicate how confident they were that their answer was correct. In accordance to methods in recent research on confidence-ratings (e.g., Stankov & Crawford, 1996, 1997; Zakay & Tuvia, 1998), this was expressed in terms of percentages with a minimum confidence rating of 11% as an indicator for pure guessing (because of nine answer options) and a maximum confidence rating of 100%. For the MMT-1 and also the CFT-20 and the test d2 participants were instructed to work as accurately and also as quickly as possible.

Data analyses. At first, comparability of participants in the different test taking procedures and gender differences were tested with respect to abstract reasoning ability (CFT-20), concentration ability, and raw score in the MMT-1. Reliabilities (internal consistencies) of the applied tests and questionnaires were assessed. Next, descriptive statistics (means and standard deviations) of the sample were documented for the applied tests and questionnaires (MMT-1, CFT-20, d2, QCM, NEO-FFI-scale). For the replication of the F>C-phenomenon mean latencies for correct responses and errors in the MMT-1 were compared by t-tests. The effect of ability level on response latencies for correct responses and errors was investigated

by an analysis of variance. Partial correlations were computed in order to investigate the relationships between confidence, latencies, and accuracy. Finally, the contribution of motivational variables and gender to the explanation of confidence was analyzed by hierarchical regression analysis.

Results

Participants in the different test taking procedures did not differ significantly in their abstract reasoning ability as assessed by the CFT-20 ($t(101) = -1.31$, $p = .19$)[6], their concentration ability as assessed by the test d2 ($t(101) = .71$, $p = .48$), or in their MMT-1-score ($t(101) = -.54$, $p = .59$). Thus, the data sets from both sequences could be analyzed together. There were no gender differences in abstract reasoning ability or concentration ability (MMT-1-score: $t(101) = .63$, $p = .53$; CFT-20: $t(101) = 1.31$, $p = .19$; d2: $t(101) = -.87$, $p = .39$). In addition, there was no significant correlation between age and MMT-1-score ($r = .04$, $p = .67$).

Sample reliability (internal consistency) was $\zeta = .88$ for the MMT-1 (with sufficient item-total correlations of $r_{it} > .30$ for 85 % of the items), $\zeta = .69$ for the CFT-20, $\zeta = .85$ for the NEO-FFI, and $\zeta = .97$ for the test d2. Three of the four motivational scales of the QCM showed sufficient internal consistencies (challenge: $\zeta = .84$; interest: $\zeta = .63$; fear of failure: $\zeta = .85$). However, the internal consistency of probability of success was unsatisfactory with $\zeta = .32$. Mean internal consistency of this factor in previous studies was $\zeta = .74$ (Rheinberg, Vollmeyer, & Burns, 2001).

Table 1 shows the sample statistics for the various tests that were applied in this study.

Table 1:
Sample Descriptives for the Applied Tests and Questionnaires.

	M	*SD*	Min	Max	*N*
Abstract reasoning ability (CFT-20)[a]	126.89	10.02	100	147	103
Concentration ability (test d2)[b]	110.36	8.89	82	130	103
Challenge (QCM-scale)[c]	5.22	.91	1.75	7.00	102
Interest (QCM-scale)[c]	4.82	1.27	1.00	6.80	102
Fear of failure (QCM-scale)[c]	3.08	1.35	1.00	6.80	102
Probability of success (QCM-scale)[c]	4.28	.98	2.25	6.25	102
Conscientiousness (NEO-FFI-scale)[d]	2.68	.57	1.42	3.84	103
MMT-1 raw score	22.22	6.07	5	31	103

Note. [a]*Standard IQ-Scale with M = 100, SD = 15.* [b]*Standard Z-Scale with M = 100, SD = 10.* [c]*on a scale from 1 to 7 with 7 = "I totally agree" and 1 = "I totally disagree".* [d]*on a scale from 1 to 5 with 5 = "I totally agree" and 1 = "I totally disagree".*

[6] If not otherwise mentioned, tests were two-tailed.

As can be seen in Table 1, with a mean IQ of 127 participants were clearly above-average in their abstract reasoning ability. Compared to the non-restricted standard deviation of 15 IQ-points of the standard IQ-scale variability of test scores was restricted. On average, participants were one standard deviation above the mean in their concentration ability. Comparing the four factors of current motivation with each other, participants were more motivated to gain success than to avoid failure ($t(101) = -7.31$, $p < .01$). However, this finding is questionable because of the insufficient reliability of the scale "probability of success". With reference to the norm data of the German version of the NEO-FFI participants showed an average level of conscientiousness.

Relationships between accuracy, latency, and confidence. For the investigation of relationships between accuracy, latency, and confidence we used the data collected with the MMT-1. Ninety-two percent of the participants worked on all items of the MMT-1, indicating that test taking took place under power conditions. On average, participants worked on the test for 72.33 minutes ($SD = 13.94$). There was no significant correlation between total time working on the MMT-1 and raw score ($r = -.10$, $p = .32$).

Replication of the F>C-phenomenon. Latencies for errors and correct responses differed significantly with the former being about 30% longer than the latter (errors: $M = 146.97$ sec., $SD = 41.81$ sec.; correct responses: $M = 112.67$ sec., $SD = 27.42$ sec.; $t(101) = -9.39$, $p < .01$). Thus, the F>C-phenomenon could be replicated in the present study. The difference in percentages between both latencies was comparable to those found in other studies (e.g., Hornke, 1997; Rammsayer & Brandler, 2003). The correlation between latencies for errors and correct responses was $r = .50$ ($p < .01$), revealing an individual tendency to respond either fast or slow.

Effects of ability level on response latencies for errors and correct responses. The question if the F>C-phenomenon could be shown for all participants independent of ability was investigated by an analysis of variance (within factors: mean error-latency, mean latency for correct responses; between factor: MMT-1 quartiles). Mean error latencies and mean latencies for correct responses were normally distributed (errors: *Kolmogorov-Smirnov-Z* = .04, $p = .20$; correct responses: *Kolmogorov-Smirnov-Z* = .05, $p = .20$; $df = 103$). The F>C-phenomenon could be replicated for all ability groups ($F(1, 99) = 93.33$, $p < .01$). There was no significant main effect for ability group ($F(3, 99) = .08$, $p = .97$) and no interaction between ability group and latency for errors or correct responses ($F(3, 65) = 2.04$, $p = .11$). These results support the universality of the F>C-phenomenon. However, it has to be taken into account that the sample of the present study was above average in intellectual ability.

Relationships between confidence, accuracy, and latencies. For every participant, data were aggregated over all items to receive mean confidence scores and mean latency scores. Responses given with a high level of confidence (answers given with more than 90% confidence) had shorter latencies ($M = 182.77$ sec., $SD = 61.52$ sec.) than responses given with a low level of confidence (answers given with less than 90% confidence; $M = 345.25$ sec.; $SD = 103.97$ sec.; $t(66) = -12.80$, $p < .01$). This finding was independent of answer accuracy: For answers given with high confidence mean latency for correct responses was 92.33 sec. ($SD = 57.84$) and for errors 95.68 sec. ($SD = 56.55$). For answers given with low confidence mean latency for correct responses was 138.54 sec. ($SD = 86.50$) and for errors 152.60 sec. ($SD = 80.66$).

Over all participants the partial correlation between score and latency when controlling for confidence was $r = .01$ ('raw' $r = -.15$), between confidence and latency when controlling

for score it was r = -.19 ('raw' r = -.24), and between confidence and score when controlling for latency it was r = .65 ('raw' r = .66). Thus, there was a stronger relationship between confidence and latency than between accuracy and latency. The correlation between confidence and latency was hardly affected by accuracy while the correlation between accuracy and latency could be totally explained by confidence.

Results were stable when analyzing individual data (not aggregated over participants). Multiple regressions of latency on confidence and accuracy were calculated. In 92% of all cases there was no increase in the amount of explained variance by including accuracy after confidence (ΔR^2 did not reach statistical significance; when including confidence after accuracy ΔR^2 did reach statistical significance in 75% of all cases).

Investigation of the relationship between latency and confidence separately for correct responses and mistakes did not change the results: For both answer types participants were more confident when giving quicker answers (correct responses: r = -.27, p < .01; errors: r = -.13, p = .09). Correlations did not differ significantly (Fisher-Z = .14, p = .23). Confidence ratings for correct responses and mistakes correlated positively with each other (r = .76, p < .01).

Summing up these findings, confidence was negatively related to latency, independent of answer accuracy. The less time participants needed to give an answer, the more confident they were in it. At the same time confidence was positively related to answer accuracy, and this was independent of latency. This correlation was even larger than the one between confidence and latency. Thus, participants were more confident with correct answers than with errors. There was no relationship between accuracy and latency when controlling for confidence.

Contributions of gender and motivational variables to the explanation of confidence. The inspection of correlations (see table 2) between confidence, motivational variables, and gender showed that women expressed less confidence than men, independent of their current motivation or intellectual ability (women were coded '1', men by '2'). Also, women took longer to answer the tasks. To investigate the influence of gender and motivational variables on confidence we used hierarchical regression analysis. Confidence as dependent variable was explained by three models including the following variables successively: Score and latency, gender, and the four motivational variables (order of inclusion as listed in table 2). All models explained confidence significantly (model 1: $F(2,99) = 40.75$, $p < .01$; model 2: $F(3,98) = 34.54$, $p < .01$; model 3: $F(7,94) = 18.49$, $p < .01$). Each model significantly enhanced the explanation of variability of confidence scores. Table 3 contains the model summaries, with the detailed results of the regression analysis for model 3 shown in table 4.

Men (82%) gave higher confidence ratings than women (69%), unaffected by score and latency. Gender alone explained about 12% of the variance in confidence ratings. With respect to motivational variables, only fear of failure contributed significantly to the explanation of confidence. People higher in fear of failure, as assessed before test taking, were less confident in their answers, even after becoming acquainted with the kind of task demanded from them and independent of score, latency, or gender. Fear of failure explained about 6% of additional variance in confidence ratings. Results were stable when changing the order in which the variables were included in the regression equation (e.g., including latency before score, including the motivational variables before gender, or varying the order of the four motivational variables).

Table 2:
Correlations between mean confidence ratings, score, mean latency, gender, and four variables of current motivation (n = 102).

	1	2	3	4	5	6	7	8
1. Confidence	1.00							
2. Score	.66**	1.00						
3. Latency	-.24*	-.15	1.00					
4. Gender	-.32**	-.06	.23*	1.00				
5. Challenge	.01	.06	.01	.03	1.00			
6. Interest	.28**	.20*	-.07	.06	.49**	1.00		
7. Fear of failure	-.23*	.04	.14	.14	.39**	-.10	1.00	
8. Prob. success	.06	.06	-.03	-.03	-.20*	.03	.00	1.00

*Note: ** = $p < .01$; * = $p < .05$.*

Table 3:
Model summaries for the hierarchical regression analysis to explain mean confidence ratings (N = 102).

Model	R	R^2	Adjusted R^2	*SEE*	Change in F	*df 1*	*df 2*	*p*
1[a]	.67	.45	.44	14.45	40.75	2	99	< .01
2[b]	.72	.51	.50	13.67	12.59	1	98	< .01
3[c]	.76	.58	.55	12.98	3.65	4	94	.01

Note: [a]model included score and latency as independent variables. [b]model included score, latency, and gender as independent variables. [c]model included score, latency, gender, challenge, interest, fear of failure, and probability of success as independent variables.

Table 4:
Regression coefficients of model 3 for the explanation of mean confidence ratings by score, latency, gender, and variables of current motivation (N = 102).

	B	*SE*	*β*	*t*	*p*	Partial *r*	Tolerance
(Constant)	49.20	13.79		3.57	< .01		
Score	.62	.07	.61	8.81	< .01	.67	.92
Latency	.00	.00	-.05	-.74	.46	-.08	.92
Gender	-10.20	2.91	-.24	-3.50	< .01	-.34	.92
Challenge	-.17	2.00	-.01	-.09	.93	-.01	.51
Interest	2.26	1.32	.15	1.72	.09	.18	.60
Fear of failure	-2.82	1.17	-.20	-2.42	.02	-.24	.68
Prob. success	.33	1.38	.02	.24	.81	.02	.91

Discussion

The present study investigated the universality of the F>C-phenomenon for a non-adaptive figural matrices test with respect to different ability groups. The relationships between response latencies, answer accuracy, and confidence with which answers were given were analyzed. The variance in confidence ratings that was independent of latency and accuracy was explained by gender and motivational variables.

The F>C-phenomenon could be replicated for figural matrices tasks applied in a non-adaptive test setting. In accordance with former studies, error latencies were about 30% longer than latencies for correct responses and both latencies were positively correlated (e.g., Hornke, 1997; Rammsayer, 1999; Rammsayer & Brandler, 2003). A more detailed analysis revealed the universality of the F>C-phenomenon for different ability groups. While Beckmann, Guthke, and Vahle (1997) reported the F>C-phenomenon only for high-scoring participants and Beckmann (2000) found the difference in latencies for errors and correct responses to be positively related to ability level, no interaction between ability level and response latency for errors and correct responses was found in the present study. But the sample studied here was above average in intellectual ability. Therefore, the results confirm the F>C-phenomenon for this group of participants.

However, although the total sample was above average in intellectual ability, there still were interindividual differences in ability. Thus, the MMT-1 items were of different subjective difficulty for participants with varying levels of ability. Indications that the F>C-phenomenon is also present in testing situations with tasks of comparable subjective difficulty levels have come from studies that investigated error and correct response latencies with adaptive test forms (e.g., Beckmann, Guthke, & Vahle, 1997; Hornke, 1997, 2000; Rammsayer, 1999; Rammsayer & Brandler, 2003). In the present study it was not possible to unconfound the effects of task difficulty and ability. However, this was not an aim of the present study because we were interested in the diagnostic value of latencies in typical test taking situations. Thus, our results refer to data collected with tasks that were of different subjective difficulty for participants of varying levels of ability – a situation that can be considered typical in most real life settings.

One aim of this study was to add to the understanding of the relationship between accuracy, latency, and confidence. Zakay and Tuvia (1998) documented that response latencies are more closely related to confidence than to answer accuracy: Response latencies for answers given with high confidence were shorter than response latencies for answers given with low confidence, independent of the correctness of response. Therefore, confidence was deemed to be a relevant concept for the explanation of latencies, more so than accuracy. Zakay and Tuvia (1998) discuss a choice latency heuristic to explain why answers with shorter latencies are rated with higher confidence. This heuristic states that short latencies are associated with low perceived mental effort which in turn is associated with higher confidence in the correctness of the response. Also, in the present study responses given with a high level of confidence had shorter latencies than responses given with a low level of confidence. This finding was independent of answer accuracy. When controlling for confidence there was hardly any relation between latency and accuracy while there was still a small but significant correlation between latency and confidence when controlling for accuracy. Thus, the results of the present study seem to support the existence of a choice latency heuristic. But several other aspects have to be taken into account. These include:

(1) Confidence ratings for correct answers and mistakes were positively correlated, which indicated an individual tendency to rate one's own confidence either high or low. This finding is inconsistent with a choice latency heuristic because of the F>C-phenomenon present in the data. Rather, the positive correlation of confidence ratings for answers of varying accuracy can be interpreted with respect to a confidence factor that has been documented in the research of the confidence paradigm (Pallier et al., 2002; Schraw, 1994, 1997; Stankov & Crawford, 1997).

(2) Confidence was more strongly related to answer accuracy than to latency. Participants were quite successful in expressing the correctness of their responses in their confidence ratings. In addition, the relation between confidence and accuracy was not affected by latency. Results of the regression analysis showed that accuracy contributed significantly to the explanation of confidence ratings while latency made no significant contribution.

(3) It has been shown that gender acts as a moderator variable in both self estimates of intellectual ability and confidence ratings: Women rated their confidence lower than men in a variety of tasks, independent of level of achievement and accuracy of self-estimates (e.g., Holling & Preckel, 2005; Pallier, 2003). This finding can be explained by the influence of gender stereotypes (Beloff, 1992; Beyer, 1998; Pallier, 2003). In the present study a substantial amount of the variability of confidence ratings was solely explained by gender (12%). Women gave significantly lower confidence ratings than men. Differences in confidence ratings could not be explained as a reflection of minor but existent differences in intellectual ability, which are then exaggerated in self-estimates (Furnham & Rawles, 1995). Moreover, there were no gender differences in the four aspects of current motivation to work on the figural matrices tasks (challenge: $t(100) = -.26$, $p = .80$; interest: $t(100) = -.57$, $p = .57$; fear of failure: $t(100) = -1.57$, $p = .12$; probability of success: $t(100) = .23$, $p = .82$). It is more reasonable to assume that the gender related differences in confidence ratings were a reflection of gender stereotypes (Beloff, 1992). As Ackerman, Beier, and Bowen (2002) have pointed out, further research is needed here to combine objective measures and measures of self concept.

(4) Confidence ratings are positively correlated with the level of prospective confidence before making a choice (Beyer, 1998; Zakay & Tuvia, 1998). This finding does not support the existence of a choice-latency-heuristic. It has been interpreted in terms of a need for self-consistency which causes posttask evaluations to be influenced by pretask expectancies (Beyer, 1990, 1998; Beyer & Bowden, 1997). Individual differences in pretask expectancies were again interpreted in terms of gender stereotypes (Beyer, 1998).

(5) It has been hypothesized that motivational variables act as another moderator variable for confidence ratings (Stankov & Crawford, 1997). In our study fear of failure (assessed before working on the test items but after working on eight practice items) explained six percent of the variability of confidence ratings, independent of all other variables. Challenge, interest, or probability of success did not contribute to the explanation of confidence. However, the reliability of the scale "probability of success" was not sufficient. Fear of failure has been conceptualized not only as a motive to avoid failure but also as a need, or as an affective tendency (Conroy, 2003). The fear of failure construct involves future-oriented apprehension about social evaluation, the threat of appearing incompetent, and the resulting consequences. Fear of failure is positively correlated to test anxiety (Elliot & McGregor, 1999) but refers to a broader context than test anxiety, which is mainly related to school settings. As factors influencing fear of failure socialization, early childhood experiences,

learning experiences, biological constitution, as well as subjective and contextual factors have been investigated (Zeidner, 1998).

To conclude, in the present study latency was more strongly related to confidence than to answer accuracy. Therefore, the investigation of confidence appears to be a fruitful avenue for the study of response latencies. Yet, further investigation of the factors influencing confidence is necessary. The results of this study demonstrated that confidence cannot sufficiently be explained by a choice latency heuristic. Instead, in the present study confidence was a multiple determined construct. About 60% of its variability could be explained by task related achievement, gender, and fear of failure.

Besides answer accuracy, response latencies provide an interesting additional source of psychometric information. The question of their diagnostic value, however, cannot be answered easily. This study revealed that for figural matrices tasks latencies are more strongly related to subjective confidence which partly appeared to be influenced by non-cognitive variables like gender and fear of failure than to answer accuracy as an indicator of cognitive ability. Therefore, taking response latencies as an additional source of information about cognitive ability would mix up cognitive and non-cognitive factors. A practical implication of this would be the disadvantaging of women over a broad range of tasks. This is not a desired consequence of any diagnostic test.

References

1. Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. Personality and Individual Differences, 33, 587-605.
2. Beckmann, J. F. (2000). Differentielle Latenzzeiteffekte [Differential effects on latencies in solving reasoning items]. Diagnostica, 46, 124-129.
3. Beckmann, J. F., Guthke, J., & Vahle, H. (1997). Analysen zum Zeitverhalten bei computergestützten adaptiven Intelligenz-Lerntests [Analysis of item response latencies in computer-aided adaptive intelligence learning ability tests]. Diagnostica, 43, 40-62.
4. Beloff, H. (1992). Mother, father, and me: our intelligence. The Psychologist, 5, 309-311.
5. Bennett, M. (2000). Correlations between self-estimated and psychometrically measured IQ. Journal of Social Psychology, 139, 405-410.
6. Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. Journal of Personality and Social Psychology, 59, 960-970.
7. Beyer, S. (1998). Gender differences in self-perception and negative recall biases. Sex Roles, 38, 103-133.
8. Beyer, S., & Bowden, E. M. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. Personality and Social Psychology Bulletin, 23, 157-172.
9. Borkenau, P., & Ostendorf, F. (1993). NEO-Fünf-Faktoren Inventar (NEO-FFI) [NEO-Five-Factor Inventory]. Göttingen: Hogrefe.
10. Brickenkamp, R. (2002). Test d2 – Aufmerksamkeits-Belastungs-Test (9., überarbeitete und neu normierte Aufl.) [Test d2 – attention-stress-test (9th ed.)]. Göttingen: Hogrefe.
11. Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. Psychological Review, 97, 404-431.

12. Cattell, R. B. (1960). The individual or group culture fair intelligence test. Champaign, Illinois: IPAT.
13. Conroy, D. E. (2003). Representational models associated with fear of failure in adolescents and young adults. Journal of Personality, 71, 757-783.
14. Costa, P. T. Jr., & McCrae, R. R. (1989). The NEO PI/FFI manual supplement. Odessa, FL: Psychological Assessment Resources.
15. Costa, P. T. Jr., & McCrae, R. R. (1992). NEO-PI-R professional manual. Odessa, FL: Psychological Assessment Resources.
16. Crawford, J., & Stankov, L. (1996). Age differences in the realism of confidence judgments: A calibration study using tests of fluid and crystallized intelligence. Learning and Individual Differences, 6, 84-103.
17. Ebel, R. (1953). The use of item response time measurements in the construction of educational achievement tests. Educational and Psychological Measurement, 13, 391-401.
18. Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierachical model of approach and avoidance achievement motivation. Journal of Personality and Social Psychology, 76, 628-644.
19. Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3, 380-396.
20. Freund, P. A. (2003). Beiträge unterschiedlicher Konstruktionsregeln zur Schwierigkeit von figuralen Matrizenaufgaben [Contribution of different construction rules to item difficulty in figural matrices tasks]. Unpublished diploma thesis, University of Münster, Germany.
21. Furnham, A., Clark, K., & Bailey, K. (1999). Sex differences in estimates of multiple intelligences. European Journal of Personality, 13, 247-259.
22. Furnham, A., Fong, G., & Martin, N. (1999). Sex and cross-cultural differences in the estimated multi-faceted intelligence quotient score for self, parents, and siblings. Personality and Individual Differences, 26, 1025-1034.
23. Furnham, A. & Rawles, R. (1995). Sex differences in the estimate of intelligence. Journal of Social Behaviour and Personality, 10, 741-745.
24. Gittler, G. (1989). Dreidimensionaler Würfeltest. Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. Testmanual [Three-dimensional cube test. A Rasch-scaled test for measuring spatial power of imagination. Testmanual]. Weinheim, Germany: Beltz Test.
25. Grosjean, M., Rosenbaum, D. A., & Elsinger, C. (2001). Timing and reaction time. Journal of Experimental Psychology: General, 130, 256-272.
26. Halpern, D. F., & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. Educational Psychological Review, 12, 229-246.
27. Holling, H. & Preckel, F. (2005). Self-Estimation of Intelligence: Methodological approaches and gender differences. Personality and Individual Differences, 38, 503-517.
28. Hornke, L. F. (1997). Untersuchungen von Itembearbeitungszeiten beim computergestützten adaptiven Testen [Investigating item response times in computerized adaptive testing]. Diagnostica, 43, 27-39.
29. Hornke, L. F. (2000). Item response times in computerized adaptive testing. Psychologia – Revista de Metodologia y Psycologia Experimental, 21, 175-189.
30. Jensen, A. R. (1998). The *g* factor: The science of mental ability. Westport: Praeger Publishers.
31. Knorr, E., & Neubauer, A. C. (1996). Speed of information-processing in an inductive reasoning task and its relationship to psychometric intelligence. Personality and Individual Differences, 20, 653-660.

32. Luce, R. D. (1986). Response times. New York: Oxford University Press.
33. Lynn, R., Backhoff, E., & Contreras-Nino, L.A. (2004). Sex differences on *g*, reasoning and visualisation tested by the progressive matrices among 7-10 year olds: some normative data for Mexico. Personality and Individual Differences, 36, 779-787.
34. May, R. S. (1986). Current issues in West German decision research. In R. W. Scholz (Hrsg.), Psychologie des Entscheidungsverhaltens und des Konfliktes, Band 4 (S. 13-30). Frankfurt: Lang.
35. Mills, C. J., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Ravens Progressive Matrices a good idea? Gifted Child Quarterly, 39, 209-217.
36. Milto, E., Rogers, Ch., & Portsmore, M. (2002). Gender differences in confidence levels, group interactions, and feelings about competition in an introductory robotics course. Frontiers in Education Conference: Boston.
37. Pallier, G. (2003). Gender differences in self-assessment of accuracy on cognitive tasks. Sex Roles, 48, 265-276.
38. Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. The Journal of General Psychology, 129, 257-299.
39. Preckel, F. (2003). Diagnostik intellektueller Hochbegabung. Testentwicklung zur Erfassung der fluiden Intelligenz [Assessment of intellectual giftedness: Test development for the assessment of fluid intelligence]. Göttingen: Hogrefe.
40. Preckel, F., & Thiemann, H. (2003). Online- versus paper-pencil-version of a high potential intelligence test. Swiss Journal of Psychology, 62, 131-138.
41. Rammsayer, T. (1999). Zum Zeitverhalten beim computergestützten Testen: Antwortlatenzen bei richtigen und falschen Lösungen [Timing behavior in computerized adaptive testing: Response times as a function of correct and incorrect answers]. Diagnostica, 45, 178-183.
42. Rammsayer, T., & Brandler, S. (2003). Zum Zeitverhalten beim computergestützten adaptiven Testen: Antwortlatenzen bei richtigen und falschen Lösungen sind intelligenzunabhängig [Timing behavior in computerized adaptive testing: Response times for correct and incorrect answers are not related to general fluid intelligence]. Zeitschrift für Differentielle und Diagnostische Psychologie, 24, 57-63.
43. Rammstedt, B. & Rammsayer, T. H. (2000). Sex differences in self-estimates of different aspects of intelligence. Personality and Individual Differences, 29, 869-880.
44. Rammstedt, B. & Rammsayer, T. H. (2001). Geschlechtsunterschiede bei der Einschätzung der eigenen Intelligenz im Kindes- und Jugendalter [Gender differences in self-estimated intelligence in infancy and adolesence]. Zeitschrift für Pädagogische Psychologie, 15, 207-217.
45. Rammstedt, B. & Rammsayer, T. H. (2002). Self-estimated intelligence: Gender differences, relationship to psychometric intelligence and moderating effects of level of education. European Psychologist, 7, 275-284.
46. Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85, 59-108.
47. Ratcliff, R. (1981). A theory of order relations in perceptual matching. Psychological Review, 88, 552-572.
48. Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling the accumulation of partial information. Psychological Review, 95, 238-255.
49. Ratcliff, R., & Rouder, J. F. (1998). Modeling response times for two-choice decisions. Psychological Science, 9, 347-356.

50. Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and Diffusion Models of Reaction Time. Psychological Review, 106, 261-300.
51. Raven, J. C. (1936). Standard Progressive Matrices, Sets A, B, C, D, E II. London: Lewis.
52. Raven, J. C. (1962). Advanced Progressive Matrices Set II. London: Lewis.
53. Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [QCM: A questionnaire for the assessment of current motivation in learning situations]. Diagnostica, 47, 57-66.
54. Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring.
55. Contemporary Educational Psychology, 19, 143-154.
56. Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. Journal of Experimental Education, 65, 135-146
57. Stankov, L., & Crawford, J. (1996). Confidence judgments in studies of individual differences. Personality and Individual Differences, 21, 971-986.
58. Stankov, L., & Crawford, J. (1997). Self-confidence and performance on tests of cognitive abilities. Intelligence, 25, 93-109.
59. Vodegel Matzen, L. B. L., van der Molen, M. W., & Dudnik, C. M. (1994). Error analysis of Raven test performance. Personality and Individual Differences, 16, 433-445.
60. Weiß, R. H. (1998). Grundintelligenztest Skala 2 (CFT-20) mit Wortschatztest (WS) und Zahlenfolgentest (ZF). Handanweisung (4., überarbeitete Auflage) [CFT-20, culture fair intelligence test with additional tests vocabulary and number series]. Göttingen: Hogrefe.
61. Wickens, C. D. (1992). Engineering Psychology and Human Performance (2nd ed.). New York: Harper Collins.
62. Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. Acta Psychologica, 98, 103-115.
63. Zakay, D., & Tsal, Y. (1993). The impact of using forced decision-making strategies on post-decisional confidence. Journal of Behavioural Decision-Making, 6, 53-68.
64. Zeidner, M. (1998). Test anxiety: The state of the art. New York: Plenum.