PERGAMON

Personality and Individual Differences 34 (2003) 559–574

PERSONALITY AND INDIVIDUAL DIFFERENCES

www.elsevier.com/locate/paid

Stability and variability in the realism of confidence judgments over time, content domain, and gender

Anna-Carin Jonsson, Carl Martin Allwood*

[a]*Department of Psychology, Lund University, Box 213, 221 00 Lund, Sweden*

Received 15 June 2001; received in revised form 21 December 2001; accepted 21 January 2002

Abstract

This study investigates the influence on the realism of confidence judgments of four different factors, the individual, the knowledge domain (crystallized and fluid intelligence), gender and cognitive style (Need-for-Cognition, NfC). Seventy-nine high-school students answered questions on word knowledge (WORD) and logical/spatial ability (DTK); both tests were administered on three occasions with two weeks between each trial. After each test question, each individual gave a confidence rating of his or her answer. The results showed some, but not perfect, individual stability. Furthermore, within-subject differences were found between domains (WORD/DTK); the participants showed better calibration and less overconfidence for the WORD-test as compared to the DTK-test. No stable gender differences were found for any of the two tests. Finally, the results show that having high NfC is not associated with better realism in confidence judgments. These results suggest that the realism of confidence judgments is, at least on the distal level, influenced by many different factors.

Keywords: Metacognition; Confidence; Calibration; Individual stability; Gender; Need-for-Cognition; Knowledge domain; Cognitive style

1. Introduction

Poor realism in confidence judgments of the correctness of one's own decisions can have devastating consequences. For example, a physician who is totally confident about the level of an important reference value but who has got it wrong may risk his/her patient's life. In the present study we investigate the influence on the realism of confidence judgments of four different factors, the domain, the individual, gender and a cognitive style variable, Need-for-Cognition.

* Corresponding author. Fax +46-46-2224209.
E-mail address: carl_martin.allwood@psychology.lu.se (C. M. Allwood).

0191-8869/03/$ - see front matter
PII: S0191-8869(02)00028-4

1.1. Confidence judgments

A confidence judgment expresses how sure a person is about the correctness of his or her own performance, belief or knowledge state. Confidence judgments can be made both with respect to predictions, such as weather forecasting (e.g. Murphy & Winkler, 1971), and with respect to concurrent and retrospective tasks, such as general knowledge tasks (e.g. Lichtenstein, Fischhoff, & Phillips, 1982). By good realism we mean that answers assigned a certain confidence value of being correct (e.g. 60% sure) in the long run have the corresponding proportion of correct answers (i.e. 60% correct). Overconfidence means that the level of peoples' confidence judgments is higher than the level of their accuracy.

A quite robust phenomenon in previous research has been *the hard-easy effect* (Juslin, Winman, & Olsson, 2000; Lichtenstein & Fischhoff, 1977), that is, when accuracy is high, overconfidence tends to be low and when accuracy is low, overconfidence tends to be high. Juslin et al. (2000) pointed out that in most research the hard-easy effect has not been separated from statistical effects and scale-end effects (see also, Erev, Wallsten, & Budescu, 1994).

Much research on the realism in confidence judgments has in the last decades been concerned with whether or not people have a bias towards overconfidence (e.g. Gigerenzer, 1994; Griffin & Tversky, 1992; Juslin, 1993; Juslin et al., 2000; Keren, 1991,1997; Nelson, 1996). However, at the present time it may not be possible to answer this question since we do not have access to a generally accepted theory concerning which tasks may be ecologically representative for the study of confidence judgments. A similar, but maybe more fruitful question, concerns when and where we can expect overconfidence to occur (Allwood & Granhag, 1999; Bless & Strack, 1998; Bornstein & Zickafoose, 1999; McClelland & Bolger, 1994; Swann & Gill, 1998). For example, will a person who shows overconfidence on one occasion in a given domain be overconfident the next time in the same domain, or perhaps in a different domain? Or are specific cognitive styles associated with a tendency to be overconfident in many domains (Klayman, Soll, Gonzáles-Vallejo, & Barlas, 1999; Soll, 1996; Stankov, 1999; Stankov & Crawford, 1996)?

In the present study we analyzed the realism of the participants' confidence judgments by using measures from calibration research (see Lichtenstein et al., 1982). The measures used are calibration, overconfidence and resolution, and they are explained below.

1.2. Individual stability over time

We investigated individual stability in two ways. The first was to analyze stability over time. This was done by measuring each participant's result on different versions of the two tests used at three points in time. The second was to analyze individual stability by correlating the participants' results over the two domains investigated.

As far as we know, only Stankov and Crawford (1996) have measured the stability of the realism in individuals' confidence ratings on two time occasions.[1] In order to measure the test-retest

[1] After each choice of answer alternative and confidence rating, i.e. for each item, the participants made a choice whether they would "submit", or not, the item to a counting procedure computing their test score. The instruction was to maximize the final score of the counting procedure. It is not clear how this task may have affected the realism of the item-specific confidence judgments.

reliability of the tests used, Stankov and Crawford measured calibration, over/underconfidence, resolution and slope (the difference between the mean confidence for correct and incorrect items), for each individual on two occasions separated by one week.[2]

1.3. Individual stability over domains

Previous research on the realism of participants' confidence judgments as a function of domain has given somewhat mixed results, although most research has pointed to stability over domains (Bornstein & Zickafoose, 1999; Crawford & Stankov, 1996; Juslin & Olsson, 1997; Klayman et al., 1999; Kleitman & Stankov, 2001; Pallier, Danthiir, Kleitman, Knezevic, Stankov & Roberts, submitted; Soll, 1996; Stankov, 1999; Stankov & Crawford, 1996, 1997).

Crawford and Stankov (1996) and Stankov and Crawford (1996, 1997) used different tests from intelligence research drawing on the theory of fluid and crystallized intelligence such as tests on vocabulary, general knowledge, visualization and Raven's Matrices to investigate the realism in the participants' confidence judgments. Their results provided evidence for a broad overconfidence trait that is relatively independent from performance accuracy, "i.e. people tend to have the same relative position in their overconfidence on diverse types of tasks" (Stankov & Crawford, 1996, p. 980). Similarly, Kleitman and Stankov (2001), on the basis of their findings, argued that there are consistent individual differences with respect to the realism in confidence ratings (see also Pallier et al., submitted; Stankov, 1999). In a later study, Stankov and Crawford (1997) found less overconfidence on auditorily presented items compared with the same items presented/answered in writing.

In the present study, we used tests from the Swedish Scholastic Aptitude Test (Gustafsson, 1992). These tests measure crystallized (the WORD-test, in Swedish *ORD*) and fluid intelligence (the DTK-test). Each participant conducted one WORD-test and one DTK-test on three occasions, always separated by two weeks. We expected to find stable individual differences with respect to the realism of the individuals' confidence judgments over the three occasions.

1.4. Domain stability

Even though the realism of individuals may show the same rank-order over domains, domains may differ with respect to their mean realism. In several experiments, Klayman et al. (1999) showed that different kinds of knowledge domains of, for example, famous mountains and tourist cities, were associated with different levels of *calibration in the large*, that is, the mean proportion correct subtracted from the participants' mean confidence, for the same individuals.

We predicted that individuals would differ in their realism depending on whether a WORD-test or a DTK-test was performed. We speculate that it may not be the content per se of the domain that is the deciding factor for the outcome of at least some of this research. Instead it may be (1) the kind of information-search the person engages in when answering the questions and when giving the confidence ratings and/or (2) how the questions are constructed (Tversky & Koehler,

[2] Of calibration, over/underconfidence, resolution and slope only over/underconfidence was found to have a satisfactory reliability over the six different tests investigated.

1994) and/or (3) the rater's attitude towards the activated domain (Beyer & Bowden, 1997). This suggestion is further discussed below.

1.5. Information-search processes and the construction of questions

Koehler (1991, 1994, 2000) and Tversky and Koehler (1994) argued that the construction of the questions affects the rater's degree of realism. In brief, Tversky and Koehler (1994) argued and presented empirical evidence for the theory that unpacking the focal hypothesis leads to overconfidence but unpacking alternative hypotheses reduces overconfidence. Klayman et al. (1999) suggested that if more hypotheses are explicitly presented, then the realism of the confidence judgments will increase. But, as will be illustrated below, the extent to which this occurs may be a function of how the individual searches for the answer, and this in turn may depend on how the question is formulated.

In this context it is of relevance that Pallier et al. (submitted) found that open-ended questions gave rise to better realism than forced multiple-choice questions for all tests except for Raven's Matrices. The authors proposed that this was because people solving open-ended questions take more alternative hypotheses into consideration. However, Pallier et al. (submitted) could not explain why the results for Raven's Matrices did not differ between the open-ended and the multiple-choice questions. Our speculation is that when solving Raven's Matrices, one specific answer to the question is located through a reasoning process. This reasoning process consists of many steps, and in such situations more weight will be put on the focal hypothesis (i.e. the answer arrived at) no matter whether the answer format is open-ended or consists of multiple alternatives.

Moreover, a multiple step information search process will activate more information than a single step process. Previous research has shown that when people think they have a great deal of information of relevance to the question, overconfidence will, to some extent, increase independently of the correctness of this information (Bless & Strack, 1998; Gill, Silvera, & Swann, 1998; Klayman et al., 1999; Swann & Gill, 1998). On the basis of these considerations we predicted that questions, such as those in the DTK-test, solved by multiple step reasoning processes and leading to the activation of much information would be associated with high levels of confidence.

1.6. Gender differences

Folk-beliefs, at least in many Western cultures, seem to suggest that males exhibit higher levels of overconfidence than females. Unfortunately, results from previous research have been mixed and have so far not been able to resolve this issue (Beyer & Bowden, 1997; Pulford & Colman, 1997; Stankov, 1998, 1999; Swann & Gill, 1998). Stankov (1999) concluded that in general (over several studies) differences between males and females with respect to the realism of their confidence judgments have not been confirmed. This issue is further investigated in the present study.

1.7. Need-for-Cognition

Finally, the present study investigated whether or not a cognitive style variable, Need-for-Cognition (for a review see Cacioppo, Petty, Feinstein, & Jarvis, 1996), correlates with realism in

confidence. An individual who is high in Need-for-Cognition tends to see him/herself as an intellectual, enjoys hard intellectual thinking and thinks that he/she benefits from it. From the perspective of common sense, it seems likely that such an approach to thinking will be beneficial for metacognitive realism. However, Allwood, and Björhag (1990) using general knowledge questions did not find any relationship between Need-for-Cognition and realism in confidence judgments. In the present study we test whether this result replicates in a somewhat different context.

2. Method

2.1. Participants

Seventy-nine participants, 44 women and 35 men, all aged 18 years, completed all three tests. Initially, 120 high-school students in their third and last year at high school participated in the study. Each individual was required to participate three times and 41 participants dropped out. Fifteen of these participants refused to complete or did not show up on any of the three test occasions and 26 participants completed only one or two of the three WORD/DTK-tests.

2.2. Materials

2.2.1. The Swedish Scholastic Aptitude Test

This test is used as a screening-instrument for entrance to Swedish University studies and a new version is issued each half-year. In this study, only two out of the six components in the full test were included, word knowledge (WORD) and logical/spatial ability (DTK). Gustafsson (1992) showed that these two tests correlate well with crystallized (WORD) and fluid intelligence (DTK), respectively. The word knowledge test (WORD) included 30 five-alternative questions, and the logical/spatial-test (DTK) included 20 five-alternative questions. The tasks of the DTK-test demand that the test-taker use multi-step thinking processes. In contrast, the WORD-test only requires recognition of the correct synonym to the target word among the five answer alternatives.

The WORD- and DTK-tests from three slightly dated "scholastic aptitude tests" (from 04/1993, 10/1993 and 10/1994) were used in order to ensure that the participants had not seen the test before. The mean Cronbach's alpha for the three versions of the WORD-test used was 0.80 (range 0.79–0.81), and the Cronbach's alpha for the DTK-test was 0.76 (range 0.75–0.77).[3]

2.3. Need-for-Cognition

In order to keep the Need-for-Cognition test (NfC-test) short, we used a selection of the 34 original items in Cacioppo and Petty (1982). We selected the 25 items that, in a comparison

[3] These Cronbach alpha values were supplied in personal communication by Professor Christina Stage, Department of Educational Measurement, Umea University.

between academics and "workers", reported by Cacioppo and Petty (1982), discriminated best between individuals with high and low NfC. The criterion of selection was F-values >5.80. The selected items loaded high in the "NfC factor" in the two studies reported by Cacioppo and Petty (1982). Furthermore, sex differences and interaction effects were uncommon for these items. In addition, we deleted two of the 25 items that did not fit in with common Swedish cultural assumptions, ending up with 23 items in the presented test. Finally, after data-collection, one further item was removed because the translation of this item proved to be unclear to the participants. Each statement in the test was rated on a scale ranging from $+4$ (agree totally) to -4 (do not agree at all). The Cronbach's alpha for the NfC-test was 0.88, computed on the present sample of 22 items.

2.4. Procedure

Each participant was tested three times with two weeks between each occasion. The participants were tested as a class in their classrooms (in total 4 classes with on average 30 participants in each). After having received general instructions about the study, the participants completed a small training session with two questions, each of which also included confidence ratings. The experiment leader checked that all participants had understood what was meant by a confidence judgment. They were informed that they had 65 min to complete the test, 50 min for DTK and 15 min for WORD (these are the times used when the Scholastic Aptitude Test is used as an entrance test to the university). The participants were instructed to answer each question by choosing one of the five answer alternatives, whereof one was always correct. After each question the participant made a confidence rating of how sure they were that they had answered the question correctly on a scale that ranged from 20 to 100%. It was explained that 20% meant that he/she was guessing and 100% meant that he/she was absolutely sure that the answer was correct.

The tests were altered between participants in such a way that one-third of them completed tests from 04/93, one-third 10/93 and one-third 10/94 on each test occasion. After the third trial all participants had answered all six tests but on different occasions. The order of the two tests WORD and DTK was also altered between participants. Half of the participants started with WORD and the other half with DTK on each test occasion.

The procedure was the same on all three test-occasions with the exception that after finishing the other tests on the third trial they also filled in the NfC-test. After each occasion the participants were asked not to talk about the content of the experiment with anyone. The participants received free meals after the test-occasions but no payment.

2.5. Calibration measures

The following calibration measures were used to analyze the degree of realism in participants' confidence judgments. *Calibration* reflects the relation between the level of the confidence ratings and the accuracy. The formula for computing calibration is:

$$\text{Calibration} = 1/n \sum_{t=1}^{T} n_t (r_{tm} - c_t)^2$$

Here n is the total number of questions answered, T is the number of confidence classes used, c_t is the proportion of correct answers for all items in the confidence class r_t, n_t is the number of times the confidence class r_t was used and r_{tm} is the mean of the confidence ratings in confidence class r_t. Thus, calibration is computed by first dividing participants' confidence ratings into a number of confidence classes. Next, for each confidence class, the difference is taken between the mean confidence for the items and the proportion of correct items. Finally, the squared differences multiplied by the number of responses in the confidence class are summed over confidence classes and divided by the total number of items.

Over/underconfidence is computed in the same way, except that the differences are not squared. The measure indicates whether an individual is overconfident (positive value) or underconfident (negative value). Calibration is perfect and over/underconfidence is absent when their values are zero. These measures are further described in Lichtenstein et al. (1982).

Resolution, loosely speaking, reflects the ability of the subject to distinguish between two sets of answers, one set that is correct and one set that is incorrect. The formula for computing resolution is:

$$\text{Resolution} = 1/n\sum_{t=1}^{T} n_t(c_t - c)^2$$

Here, c is the proportion of all items for which the correct alternative was selected. To achieve maximal resolution a subject within a confidence class has to assign lower confidence to all questions answered incorrectly compared with the questions answered correctly (or vice versa). A higher value reflects better resolution than a lower.

3. Results

3.1. WORD-test over time

We first present the results for the Word-test and the DTK-test, analyzed over time and for gender. Five mixed two-way ANOVAs with the within-subject factor Time (1–3) and the between-subjects factor Gender were computed for the results from the WORD-tests for the five dependent measures, calibration, over/underconfidence, resolution, accuracy and confidence. The results are shown in Table 1.

A main effect was found for over/underconfidence for Time (Time 1 $M = -0.03$, Time 2 $M = -0.03$ and Time 3 $M = 0.01$), $F(2, 76) = 3.81$, $p < 0.05$. In the pairwise comparisons the mean differences between Time 1 and Time 3 and Time 2 and Time 3 were significant at the $p < 0.05$ level. The slight underconfidence at Time 1 and 2 changed to fairly good realism at Time 3. A close to significant main effect was found for Gender, $F(1, 77) = 3.88$, $p < 0.052$ (see Table 1). The men were fairly realistic at Time 1 and 2 and overconfident at Time 3. The women were underconfident at Time 1 and 2 and fairly realistic at Time 3.

A main effect was found for accuracy for Time (Time 1 $M = 49\%$, Time 2 $M = 50\%$ and Time 3 $M = 46\%$), $F(2, 76) = 3.11$, $p < 0.05$. In the pairwise comparison the only significant effect found

Table 1
Means for the dependent measures calibration, over/underconfidence, resolution, accuracy and confidence (F = females, M = males) for the WORD-test, Time 1, 2 and 3 ($n = 79$ in each condition)

	Time 1			Time 2			Time 3		
	F	M	Total	F	M	Total	F	M	Total
Calibration	0.067	0.068	0.068	0.063	0.068	0.065	0.066	0.067	0.066
Over/under confidence	−0.057	0.004	−0.030[a]	−0.061	0.002	−0.033[a]	−0.009	0.037	0.011
Resolution	0.077	0.080	0.078	0.082	0.079	0.081	0.070	0.072	0.071
Accuracy	0.508	0.472	0.492	0.529	0.471	0.503[a]	0.477	0.444	0.462
Confidence	0.452	0.477	0.463	0.468	0.474	0.470	0.468	0.480	0.473

[a] $p < 0.05$ compared with Time 3.

was between Time 2 and Time 3 ($p < 0.05$). It may be noted that these effects, although significant, were not very strong. No other effects were found.

3.2. DTK-test over time

The corresponding ANOVAs to the ones just reported were computed for the DTK-tests. The results are shown in Table 2. The only significant effect found was for Gender with respect to confidence, $F(1, 77) = 4.19$, $p < 0.05$. The men showed higher confidence than the women.

3.3. Within-subject stability: correlations between Time 1, 2 and 3

3.3.1. Pearson correlations for WORD

The stability of each participant for the five dependent measures was computed by means of Pearson correlations between Time 1 and 2, 2 and 3, and 1 and 3 for each of the five dependent measures. The results for the WORD-test (see Table 3) showed that all correlations were significant at $p < 0.01$. Accuracy and confidence, and to a somewhat lower extent over/under-

Table 2
Means for the dependent measures calibration, over/underconfidence, resolution, accuracy and confidence (F = females, M = males) for the DTK-test, Time 1, 2 and 3 ($n = 79$ in each condition)

	Time 1			Time 2			Time 3		
	F	M	Total	F	M	Total	F	M	Total
Calibration	0.078	0.103	0.089	0.097	0.098	0.097	0.108	0.096	0.103
Over/under confidence	0.086	0.112	0.098	0.106	0.135	0.119	0.119	0.154	0.134
Resolution	0.082	0.076	0.080	0.068	0.074	0.071	0.074	0.060	0.068
Accuracy	0.433	0.497	0.461	0.424	0.444	0.433	0.409	0.474	0.438
Confidence	0.519[a]	0.609	0.559	0.530[a]	0.579	0.552	0.528[a]	0.628	0.572

[a] $p < 0.05$ compared with the males.

Table 3
Pearson correlations between Time 1 and Time 2, Time 2 and Time 3, and Time 1 and Time 3 for the WORD and the DTK-tests ($n = 79$)

	Time 1–Time 2		Time 2–Time 3		Time 1–Time 3	
	WORD	DTK	WORD	DTK	WORD	DTK
Calibration	0.635[b]	0.319[b]	0.387[b]	0.551[b]	0.374[b]	0.227[a]
Over-underconfidence	0.528[b]	0.380[b]	0.594[b]	0.501[a]	0.530[b]	0.233[a]
Resolution	0.342[b]	0.211	0.571[b]	0.206	0.454[b]	0.384[b]
Accuracy	0.678[b]	0.568[b]	0.694[b]	0.584[b]	0.737[b]	0.578[b]
Confidence	0.870[b]	0.742[b]	0.849[b]	0.748[b]	0.871[b]	0.708[b]

[a] $p < 0.05$.
[b] $p < 0.01$.

confidence, showed the highest correlations and calibration and resolution had a tendency to show the lowest. This means that the individuals were to some extent rank-ordered more or less in the same way with respect to their level of, for example, over/underconfidence over the three measurement occasions. It is also interesting to note that confidence was more highly correlated over time than accuracy. For confidence the correlations T1–T2, T2–T3 and T1–T3 were $r = 0.870$, $r = 0.849$, and $r = 0.871$, and for accuracy the corresponding correlations were $r = 0.678$, $r = 0.694$ and $r = 0.737$, respectively.

3.3.2. Pearson correlations for DTK

The correlations for the DTK-test were not as high as those for the WORD-tests (see Table 3). But still all correlations, except for resolution, were significant on at least $p < 0.05$ and at most on $p < 0.01$. For resolution, the participants showed some individual stability only between T1–T3, $r = 0.384$, $p < 0.01$. Again, confidence showed a somewhat stronger correlation than accuracy.

3.4. Within-subject stability: Pearson correlations between WORD and DTK

The correlations between WORD and DTK when collapsed over time were all significant at the $p < 0.01$ level. For calibration $r = 0.501$, over/underconfidence $r = 0.664$, resolution $r = 0.368$, accuracy $r = 0.560$, and confidence $r = 0.568$.

3.5. Within-subject differences between the WORD-tests and the DTK-tests

Fig. 1 shows the calibration curves for the two tests (WORD, DTK), collapsed over time. The diagonal indicates perfect calibration. The curve for the DTK-test is further away from the diagonal then the curve for the WORD-test, indicating more overconfidence for the DTK-test.

Five mixed two-way ANOVAs with the within-subject factor Test (WORD, DTK) and the between-subjects factor Gender were computed for the same five dependent measures as above. The values for each test were collapsed over time. Table 4 shows the results.

A main effect was found for calibration with respect to the Test factor (WORD $M = 0.066$ and DTK $M = 0.096$), $F(1, 77) = 27.34$, $p < 0.001$, showing better calibration for the WORD-test than

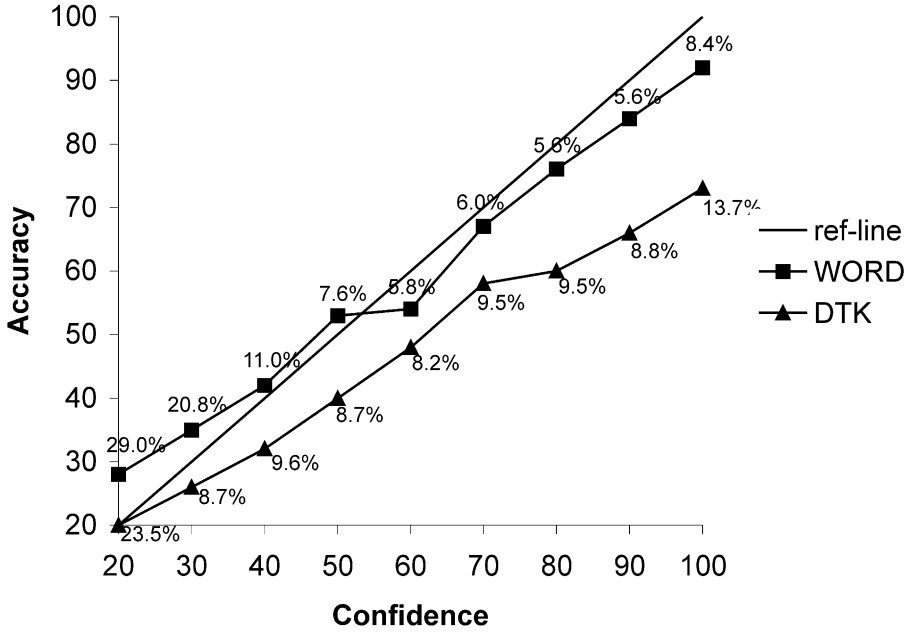

Fig. 1. Calibration curves for the WORD-test (Time 1, 2 and 3 combined) and the DTK-test (Time 1,2 and 3 combined). The percentages give the percentage for each test of all items in the test in each confidence class.

for the DTK-test. Likewise, a main effect was found for over/underconfidence for the Test factor (WORD $M = -0.017$ and DTK $M = 0.117$), $F(1, 77) = 130.9$, $p < 0.001$.[4]

The results also showed a main effect for accuracy for the Test factor (WORD $M = 48.6\%$ and DTK $M = 44.4\%$), $F(1, 77) = 4.49$, $p < 0.05$. An interaction effect was found involving WORD/women ($M = 50.5\%$), WORD/men ($M = 46.2\%$) and DTK/women ($M = 42.2\%$), DTK/men ($M = 47.2\%$), $F(1, 77) = 7.10$, $p < 0.01$. The women had quite different results on the two tests, whereas the men had fairly similar results. Finally, confidence showed a main effect for Test (WORD $M = 46.8\%$ and DTK $M = 56.1\%$), $F(1, 77) = 31.6$, $p < 0.001$.

3.6. Correlations between Need-for-Cognition and the WORD- and DTK-tests

Pearson correlations were computed between the results on the Need-for-Cognition scale (NfC-scale) and each of the five dependent measures for the score combined over the three test/occasions for each of the two tests WORD and DTK (see Table 5). We first present the results for the WORD-test. A significant correlation was found between NfC and calibration ($r = 0.223$, $p < 0.05$). Participants with high NfC showed less realism in their calibration values. A significant

[4] Kolmogorov–Smirnov tests were computed for all five dependent measures. All except calibration had a normal distribution. In order to control for outliers disturbing the normal distribution in calibration, the variable was transformed to standardized z-values. The outliers (3) who's z-values were above 3.29 were removed in accordance with Tabachnick and Fidell (2000) recommendations, in order to reduce the skewness. After this the ANOVA was recomputed for calibration. The result showed a significant difference between WORD and DTK, $F(1, 74) = 34.60$, $p < 0.001$, evidencing an even stronger effect when the skewness in calibration was reduced.

Table 4
Means for the dependent measures calibration, over/underconfidence, resolution, accuracy and confidence for females (F) and males (M) for the WORD and the DTK-test ($n = 79$ in each condition)

	WORD			DTK		
	F	M	Total	F	M	Total
Calibration	0.065	0.068	0.066[b]	0.094	0.099	0.096
Over/under confidence	−0.042	0.014	−0.017[b]	0.103	0.134	0.117
Resolution	0.076	0.077	0.077	0.075	0.070	0.073
Accuracy	0.505	0.463	0.486[a]	0.422	0.472	0.444
Confidence	0.462	0.477	0.469[b]	0.526	0.606	0.561

[a] $p < 0.05$ compared with DTK.
[b] $p < 0.001$ compared with DTK.

correlation was also found between NfC and resolution ($r = 0.337$, $p < 0.01$). The higher a participant scored on NfC the better resolution the same individual showed. Furthermore, significant correlations were found between NfC and accuracy ($r = 0.349$, $p < 0.01$) and between NfC and confidence ($r = 0.456$, $p < 0.001$).

For the DTK-test no significant correlations were found between NfC and any of the three dependent measures, calibration, over/underconfidence and resolution. However, significant correlations were found between NfC and accuracy ($r = 0.445$, $p < 0.001$) and between NfC and confidence ($r = 0.468$, $p < 0.001$).

3.7. Difference in variance and scale-end effects

In order to control for the possibility that differences in variance could explain the differences found for the two tests, Word and DTK, and for the tendencies for gender differences we computed ANOVAs of the individuals' variance in confidence for each test using Time (1–3) and Gender as the factors and for the two tests collapsed over time, using Test (WORD and DTK) and Gender as the two factors. None of these three ANOVAs showed any significant effects.

We also checked for the possibility of value of scale-end effects being the explanation for the difference found between the two tests WORD and DTK. This was done by comparing the two

Table 5
Correlations between Need-for-Cognition, the WORD-test and the DTK-test ($n = 79$)

	NfC–WORD	NfC–DTK
Calibration	0.223[a]	−0.069
Over/under confidence	0.073	0.055
Resolution	0.337[b]	0.025
Accuracy	0.349[b]	0.445[c]
Confidence	0.456[c]	0.468[c]

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.

tests' distribution of the more or less extreme confidence judgments. In this context we first created five confidence categories for each test by adding the frequencies for the most extreme confidence classes (20–29 and 100%), (30–39 and 90–99%), the less extreme confidence classes (40–49 and 80–89%), (50–59 and 70–79%) and the middle range confidence classes (60–69%). Second, the frequencies in these five confidence categories were compared between the two tests by use of Kolmogorov–Smirnov's test. No significant difference in distribution of the confidence judgments was found. Similar tests were computed for each Time occasion (1–3) and again there were no significant differences between the two tests.

4. Discussion

The present study investigated the influence on the realism of confidence judgments of four different factors, the individual, the knowledge domain, gender and cognitive style (Need-for-Cognition). Our results for these four factors will now be discussed in the order mentioned.

4.1. Individual stability

One dimension of individual stability has to do with whether individuals show stability in the realism of their confidence ratings *over time*. In the present study we investigated individual stability over time by testing the same individuals in two different domains (word knowledge as a specimen of crystallized knowledge and spatial knowledge as a specimen of fluid intelligence) on three different occasions, two weeks apart. In both domains, our results give some, but not complete, support to the idea of individual stability over time. Of the three measures of realism used, only over/underconfidence for the WORD-test showed a significant difference between the three test occasions.

However, since the individuals within the test groups could have changed their rank-order even though the mean values did not change much, we also computed the correlations between each individual's values on the different dependent measures. The results showed that nearly all of the correlations differed significantly from zero, indicating some stability for the particular individuals over the three test occasions for each of the two types of test.

Obviously, our results did not support the presence of total stability, since the correlations were far from unity; in fact, the highest correlation achieved for the three calibration measures (calibration for the WORD-test, Time 1–2) only explained about 40% of the variance. The correlations were stronger for the calibration and the over/underconfidence measures as compared with the measure for resolution and they were somewhat stronger for the WORD-test as compared with the DTK-test. These results indicate that the degree of individual stability varies somewhat over domains. Further research is needed to better understand the extent of this variation.

It is also of interest to note that the correlations were always somewhat higher for confidence than for accuracy. This finding parallels that of Stankov and Crawford (1996) and Bornstein and Zickafoose (1999).

Previous research has supported the notion that there is some individual stability of the realism in individuals' confidence judgments over different domains (e.g. Klayman et al., 1999; Pallier et

al., submitted). The fact that our results showed significant correlations between the WORD- and the DTK-tests is in line with these findings.

4.2. Domain stability

We also analyzed the effect of differences in knowledge domains on the realism of the participants' confidence judgments. This analysis was carried out within-subject in order to maintain control over the effects of individual variation. As we predicted, the results demonstrated that the participants showed better calibration and lower overconfidence for the WORD-test as compared with the DTK-test. It is also interesting to note that the value for over/underconfidence for the WORD-test was exceptionally good, showing only a very slight underconfidence. Here it should be noted that in spite of the fact that accuracy was higher for the WORD-test, compared with the DTK-test, the participants still showed lower confidence in their answers to the WORD-test, compared with the DTK-test.

It is of interest that the presentation of five answer alternatives to all questions in both tests does not appear to have influenced the outcome in realism to any great extent. Instead, the way the questions to the two types of tests were answered by the participants may have been more important. Although not explicitly tested in our study, our informal task analysis suggests that it is likely that, on the WORD-test the participants scanned the five answer alternatives in order to find a suitable answer more often than they did on the DTK-test. For the DTK-test, it is likely that the participants worked their way towards the correct alternative by means of a multi-step process in which the person utilized some, or all, of the information given in the problem. In addition, the same processes should have made them accumulate more information for the chosen alternative, compared with the WORD-test.

According to our reasoning in the introduction and according to previous research both of these features may have contributed to the higher overconfidence for the DTK-test that we found, as compared with the WORD-test. Future research could explore further the relation between realism in confidence judgments and type of cognitive processes and answer format.

Finally, it can be noted that our analyses did not support the possibility that error variance (see Erev et al., 1994) in the confidence judgments contributed substantially to the reported results since our analysis showed no significant differences in variance between the tests (for any of the three test occasions or when the test results were collapsed over time) or for gender. Likewise, when the extremity of the confidence judgments was compared for the two tests no support was found for the explanation that scale-end effects contributed substantially to the differences observed.

4.3. Gender differences

The present results do not give very clear support for gender differences in the realism of confidence judgments. No gender differences were found for the DTK-test, and for the WORD-test, the gender differences that were found to be close to significant were somewhat unstable over time.

To the extent that word knowledge as tested by the WORD-test can be seen as a feminine knowledge domain and the spatial knowledge tested by the DTK can be seen as a male knowledge domain our results differ from those reported by Beyer and Bowden (1997). Beyer and

Bowden found that females were underconfident in masculine domains (no differences in realism were found in neutral or feminine domains). In contrast, our results showed no gender differences for the male domain (the DTK-test), both genders were overconfident. For the WORD-test the females were somewhat underconfident and the males showed fairly good realism. Our conclusion is that gender differences with respect to realism in confidence judgments are unstable and that they are dependent on the knowledge domain and/or on the cognitive processes activated by the task given in a knowledge domain.

4.4. Cognitive style: Need-for-Cognition

We also analyzed the relation between a cognitive style variable, Need-for-Cognition, and our different measures of the realism of confidence judgments. Although NfC was positively correlated with both accuracy and confidence, the results showed no correlation between NfC and over/underconfidence in any of the two domains analyzed.

The results for calibration and resolution were dependent on the knowledge domain. For the WORD-test, NfC correlated positively with calibration, indicating that the more an individual's cognitive style was characterized by NfC, the worse was his or her calibration. However, the results for the WORD-test also showed that higher NfC was associated with better resolution. For the DTK-test, no significant correlations between NfC and any of the three measures of realism were found. The DTK results are in line with Allwood and Björhag (1990) who used general knowledge questions and found no significant correlation between NfC and realism as measured by calibration and resolution. Further research is needed in order to improve our understanding of how the various aspects of the cognitive processes that differ between individuals high and low in NfC contribute to these results.

5. Conclusion

All in all, our results point to the influence of a range of factors on the realism of confidence judgments. This suggests, as argued by Allwood and Granhag (1999) and Klayman et al. (1999), that no simple one- or few-factor theories will give a full explanation of how different levels of realism in confidence judgments are produced. This is clearly true if, as in the present study, more distal and global factors (such as individual, knowledge domain, gender and cognitive style) are considered. This is well illustrated by the complexity of the results in the present study where some, but far from all, variation was explained by each of the factors considered.

If the focus is on more proximal factors, such as the type of mental/cognitive processes leading to the confidence judgment, it may be possible to reduce some of the complexity by specifying what type of processes tend to be associated with different levels of realism in confidence judgments. For example, previous research has suggested that the extent to which different alternatives are elaborated and the total amount of information considered may be related to the level of realism in confidence judgments. The effect of different distal factors or situations could then be accounted for in terms of the specific cognitive processes they tend to activate. Of course, it is still an open empirical question whether the mapping between distal and proximal factors in general is many-to-one or many-to-many. However, in conclusion, we suggest that in future

research into the effect of distal factors on the realism of confidence judgments, or the effect of combinations of such factors, it may be helpful to study their effects with respect to the specific cognitive process they tend to elicit.

References

Allwood, C. M., & Björhag, C. G. (1990). Are two judges better than one?: on the realism of confidence judgments by pairs and individuals. In J. P., Caverni, J. M., Fabre, & M. Gonzalez, (Eds.), *Cognitive biases*. Amsterdam, Elsevier Science Publishers B.V.

Allwood, C. M., & Granhag, P. A. (1996). Realism in confidence judgements as a function of working in dyads or alone. *Organizational Behavior and Human Decision Processes*, *66*, 277–289.

Beyer, S., & Bowden, E. M. (1997). Gender differences in self-perception: convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, *23*, 157–172.

Bless, H., & Strack, F. (1998). Social influences on memory. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: cognitive and social dimensions* (pp. 90–106). London: Sage.

Bornstein, B. H., & Zickafoose, D. J. (1999). I know I know it, I know I saw it: the stability of the confidence—accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, *5*, 76–88.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131.

Cacioppo, J. T., Petty, R. E., Feinstien, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197–253.

Crawford, J., & Stankov, L. (1996). Age differences in the realism of confidence judgements: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences*, *6*, 84–103.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). New York: John Wiley and Sons.

Gill, M. J., Silvera, D. H., & Swann, W. B. (1998). On the genesis of confidence. *Journal of Personality and Social Psychology*, *75*, 1101–1114.

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

Gustafsson, J. E. (1992). The dimensionality of the Swedish scholastic aptitude test. *Scandinavian Journal of Educational Research*, *36*, 21–39.

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology*, *5*, 55–71.

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.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychological Review*, *107*, 384–396.

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

Keren, G. (1997). On the calibration of probability judgments: some critical comments and alternative perspectives. *Journal of Behavior Decision Making*, *10*, 269–278.

Klayman, J., Soll, J. B., Gonzáles-Vallejo, C., & Barlas, S. (1999). Overconfidence: it depends on how, what and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216–247.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology*, *15*, 321–341.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, *110*, 499–519.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 461–469.

Koehler, D. J. (2000). Probability judgment in three-category classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 28–52.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?. *Organizational Behavior and Human Performance*, *20*, 159–183.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgements under uncertainty: heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980–1994. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). New York: John Wiley and Sons.

Murphy, A. H., & Winkler, R. L. (1971). Forecasters and probability forecasts: some current problems. *Bulletin of American Meteorological Society*, *52*, 239–247.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102–116.

Pallier, G., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L. & Roberts, R.D. (submitted). The role of question format and individual differences in the realism of confidence judgments.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, *23*, 125–133.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: the roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*, 117–137.

Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tests. *Learning and Individual Differences*, *8*, 28–51.

Stankov, L. (1999). Mining in the "no man's land" between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 315–338). Washington, DC: American Psychological Association.

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*, 971–986.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on cognitive tests. *Intelligence*, *25*, 93–109.

Swann, B. W., & Gill, J. M. (1998). Beliefs, confidence and the widows Ademoski: on knowing what we know about others. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: cognitive and social dimensions* (pp. 107–125). London: Sage.

Tabachnick, B. G., & Fidell, L. S. (2000). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.

Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567.