Memory & Cognition
2010, 38 (4), 441-451
doi:10.3758/MC.38.4.441

Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome

Christoph Mengelkamp
University of Koblenz-Landau, Landau, Germany

and

Maria Bannert
Chemnitz University of Technology, Chemnitz, Germany

One aspect of metacognition is the monitoring of memory or comprehension measured with retrospective confidence judgments after test taking. The research questions of the present study were whether different measures for the accuracy of such confidence judgments are stable over learning time, whether they generalize over two different tests, and whether they predict the learning outcome. In order to answer these questions, a study was conducted in which university students ($N = 113$) learned about the basic concepts of operant conditioning for 30 min. Knowledge tests with confidence judgments presented after each item were obtained before learning, after 10 min, and at the end of the learning session. Bias and absolute bias were calculated as absolute measures of accuracy, and gamma, Pearson's r, and d_a were calculated as relative measures of accuracy. The results showed that the absolute measures were stable, but that the relative measures were not. Furthermore, absolute bias, gamma, and Pearson's r obtained 10 min after the beginning of the learning process predicted the learning outcome. The results are discussed with regard to research on different measures of accuracy for confidence judgments.

Studies on metamemory (see, e.g., Sampaio & Brewer, 2009), metacomprehension (e.g., Maki, Jonas, & Kallod, 1994; Mengelkamp & Bannert, 2009), and test taking (e.g., Kleitman & Stankov, 2007) use confidence judgments and calculate the accuracy of these judgments. Confidence judgments are made after an item in a test has been answered, in contrast with judgments about the ease of learning or judgments of knowing that are made before the items are solved (Nelson & Narens, 1990). Accuracy of judgments is interpreted as an indicator for metacognitive monitoring, which, besides metacognitive control, is a crucial process in research on metamemory (Nelson & Narens, 1990, 1992). In the present study, we examined whether the accuracy of confidence judgments that is operationalized by different measures is stable during learning and whether it generalizes over two different tests. Moreover, the predictive validity of different accuracy measures for learning outcome was investigated. Although there are studies available that have examined stability or generality using posttest confidence judgments (e.g., Jonsson & Allwood, 2003; Thompson & Mason, 1996), they mostly report only one or two measures of accuracy and do not compare the results of different measures. Since the accuracy of confidence judgments is widely used in current research on metamemory, metacomprehension, and test taking, the stability, generality, and validity of different measures of accuracy are important factors when findings are compared, and considering these factors may encourage further research. We will first describe the measures of accuracy briefly. Then, theory and studies concerning the stability, generality, and validity of such measures will be summarized.

Measures of Accuracy

One fundamental distinction of the measures is drawn between relative and absolute accuracy (Maki, 1998; Nelson, 1996; Nelson & Dunlosky, 1991). Measures of absolute accuracy are based on the difference between the judgment about an item and the actual performance for the same item. The most common measure of the absolute accuracy of judgments is bias. Bias is calculated as the signed difference between the average of judgments for a test and the average performance on this test (Yates, 1990), and therefore it allows the distinction between persons who are overconfident versus persons who are underconfident. A further measure of absolute accuracy is calculated by ignoring the direction of the difference between judgment and performance—for example, by averaging the squares of the differences (Schraw, 2009). In

C. Mengelkamp, mengelkamp@uni-landau.de

the following, the term *absolute bias* is used for measures that ignore the direction of the difference.

Measures of relative accuracy describe how accurately a learner discriminates between the test performance on correct and incorrect items. Relative measures of accuracy correlate the judgments with the performances within each person in the sample. If one accepts that judgments are on a metric level of measurement, Pearson's *r* can be calculated as a correlation coefficient. Nelson (1984, 1996) advocated the use of the nonparametric correlation gamma instead of Pearson's *r* because gamma does not require a metric but requires an ordinal scale level. The calculation of gamma is based on dyads of items; that is, to reach high accuracy, the judgments about two items have to be in the same order as the performances on these two items (see Gonzalez & Nelson, 1996). Furthermore, gamma is unaffected by the overall performance on the items and the magnitude of the metacognitive judgments (Nelson, 1984). A recent Monte Carlo study (Benjamin & Diaz, 2008) showed that d_a from signal detection theory is an equivalent or even a better measure of relative accuracy than is gamma.

The decision for one measure of accuracy should be guided by the nature of the data, especially of the data level. Measures of absolute accuracy require the calculation of differences that are meaningful only in the case of a metric data level. The same is true for the intraindividual correlation Pearson's *r* as a measure of relative accuracy. If the data are on an ordinal level, gamma is an appropriate measure of relative accuracy (Nelson, 1984, 1996). Of course, gamma can be used for metric data as well, since it is a more conservative measure than *r*. Moreover, one could use Kim's d_{xy} when ties on the judgments are not forced by the procedure of data collection (Gonzalez & Nelson, 1996). Nevertheless, all of the measures have been used in past research with confidence judgments—for example, bias (see, e.g., Maki, 1998) and absolute bias (e.g., Nietfeld & Schraw, 2002) as measures of absolute accuracy, and Pearson's *r* (e.g., Glenberg, Sanocki, Epstein, & Morris, 1987), gamma (e.g., Dunlosky, Rawson, & Middleton, 2005), and *d* (e.g., Tobias & Everson, 2000) as measures of relative accuracy. Measures of absolute accuracy in particular are favored by researchers working in classroom settings (see, e.g., Hacker, Bol, & Bahbahani, 2008; Nietfeld, Cao, & Osborne, 2005). Our aim was not to limit our results to one measure within each class of measures, but to show that the results hold true, irrespective of the measure used within each class. Therefore, we used a finely graded percentage scale to meet the assumption of metric data of the judgments.

Stability and Generality

What determines the stability or instability of the accuracy of confidence judgments? In regard to this question, Koriat's (Koriat, 2007; Koriat, Nussinson, Bless, & Shaked, 2008) work is essential. He proposed that judgments are generated on the basis of cues, and he distinguished theory-based cues from experience-based cues. Using theory-based cues, people judge items on the basis of their theories or beliefs—for example, beliefs about self-efficacy or about their own ability (Ehrlinger & Dunning, 2003) or expertise (Glenberg & Epstein, 1987). Self-efficacy is at least moderately stable over a period of 1 week (Lane & Lane, 2001), and the ability of reading comprehension is at least moderately stable over 5 months, as well (Byrne, 1986). Moreover, self-efficacy predicts academic learning outcome (Lane & Lane, 2001; Lane, Lane, & Kyprianou, 2004). Therefore, it can be expected that some of the stable between-person variance of self-efficacy will be found in the judgments, and some of the stability of ability in reading comprehension will be found in the learning outcome. Furthermore, self-efficacy may influence all judgments to almost the same extent, and ability may influence performance on test items to almost the same extent. Absolute measures of accuracy are not independent of the magnitude of the underlying judgments and performances (Nelson, 1984); thus, the stability of judgments and performances is inherited by the measures of absolute accuracy. In contrast, relative accuracies should not be affected by factors that influence the judgments and/or performances evenly (cf. Nelson, 1984).

Besides theory-based cues, people use experience-based cues to generate judgments. These cues arise from the information that they process when learning or answering test items (Koriat, 2007; Koriat et al., 2008). In the case of confidence judgments, such cues are, for example, the time needed to retrieve information (Kelley & Lindsay, 1993; Nelson & Narens, 1990), the completeness of the recall (Brewer, Sampaio, & Barlow, 2005), and the kind of questions asked (Maki, 1995, Experiment 1). Therefore, confidence judgments vary depending on factors that are not stable characteristics of a person but that are characteristics of the learning and testing situation. Thus, judgments vary within a person across tests and over time. Furthermore, if experience-based cues are not used in the same manner by all persons, this results in the instability of the accuracy of judgments in a sample of persons that is reflected in both absolute and relative accuracy. To sum up these deliberations, we may expect at least some stability in judgments and some stability in the absolute accuracy of judgments caused by theory-based cues. And, indeed, some evidence for such stability has been found and will be reported next.

Leonesio and Nelson (1990) found only small gamma correlations between judgments of learning, judgments of knowing, and feeling-of-knowing judgments. They suggested that there were different bases—experience-based cues, in the terminology of Koriat (2007)—for the judgments made at different points in time during the learning process. Another study (Kelemen, Frost, & Weaver, 2000) used English word pairs, English–Swahili word pairs, general knowledge questions, and narrative tests as materials. In addition to different domains, they used different types of judgments: ease of learning, judgments of learning, feeling of knowing, and text comprehension. They found distinctly higher Spearman correlations between the same types of judgments in comparison with the low stabilities found by Leonesio and Nelson between different kinds of

judgments. Kelemen et al. reported not only correlations between the judgments, but also correlations between the accuracies. Using bias, they found some stability and generality, but hardly any stability or generality was found for gamma. In the two studies cited so far (Kelemen et al., 2000; Leonesio & Nelson, 1990), judgments were made before the recognition test was taken, and stability was investigated over a period of weeks. In addition, Kelemen et al. used different domains and materials to investigate the generality of judgments. In contrast, our present research examined confidence judgments taken after four tests, and the judgments were made within the same domain over a short period of approximately 1 h.

Results from studies on test taking (Jonsson & Allwood, 2003; Schraw, Dunkle, Bendixen, & DeBacker Roedel, 1995; Schraw & Nietfeld, 1998) revealed that bias generalized over different tests to a considerable degree, since almost all correlations between bias calculated from different tests were significant and reached at least medium effect size (cf. Cohen, 1988). Moreover, bias showed at least some stability over a period of 2 weeks (Jonsson & Allwood, 2003). For absolute bias, the generality is lower than that for bias, but most correlations between different tests are still significant (Schraw & Nietfeld, 1998). Using relative accuracy measured by gamma, no stability was found for three different tasks administered at two times within a period of 2 weeks (Thompson & Mason, 1996, Experiments 1 and 2). Moreover, the split-half reliabilities of the tests were not significant either, indicating a lack of generality for gamma, even within the same test. This result is in accordance with another study in which no evidence was found for generality using gamma computed from a comprehension test, an analogy test, and a test of opposites (Pressley & Ghatala, 1988).

To sum up, there is some evidence for the stability and generality of accuracy of confidence judgments if absolute measures of accuracy are used, but there is no evidence for relative accuracy measured by gamma. This conclusion is somewhat limited, due to the fact that no study has compared more than two measures of accuracy, and no study compared an absolute with a relative measure. Furthermore, the stability of confidence judgments has only been investigated with regard to periods of weeks between the tests. Thus, the question remains whether there is any stability in shorter periods of time in the range of hours or minutes. Relative accuracy might be more stable over such short periods of time with less intervening study.

Validity

Why is the accuracy of monitoring important for learning processes and learning outcome? The judgments themselves are crucial for the control of learning, especially for the allocation of study time (Metcalfe, 2002; Son & Metcalfe, 2000; Thiede & Dunlosky, 1999). That is, study time is allocated to items that are judged as "least understood" (Dunlosky & Hertzog, 1998) or to items that are in a region of proximal learning (Metcalfe, 2002). But this allocation of study time is effective with regard to learning outcome only if the judgment is accurate, because the control of the learning process will fail if the accuracy is too low (Dunlosky, Hertzog, Kennedy, & Thiede, 2005). According to the transfer-appropriate monitoring hypothesis, the accuracy of judgments is a predictor of the learning outcome if the items used for the judgments are the same as those used to measure the learning outcome (see Dunlosky, Rawson, & McDonald, 2002, p. 86ff.). Additionally, there is evidence that the accuracy of pretest judgments even predicts the learning outcome for new items if the items are constructed to measure the same content (Thiede, Anderson, & Therriault, 2003).

Besides the cited study by Thiede et al. (2003), there is also some evidence for the validity of the accuracy of posttest confidence judgments. Accuracy measured by gamma and performance on a comprehension test correlated significantly (see, e.g., Maki, 1998, p. 240; Maki, et al., 1994). However, using bias as the absolute measure of accuracy, no correlation was found between bias and performance on the test (Maki, 1998). Nevertheless, the results were correlational in nature; thus, other variables—for example, general intelligence, reading ability, or prior knowledge—may explain the correlation. General intelligence was controlled in a classroom study (Nietfeld et al. 2005) that used absolute bias as the measure of accuracy. Thus, absolute bias was predicted by both general intelligence and the performance on the test. In another classroom study (Nietfeld, Cao, & Osborne, 2006), the students' prior knowledge was controlled. Path analyses showed that the effect of prior knowledge on the final test score was partially mediated by absolute bias. To sum up, there is evidence for the validity of gamma and absolute bias, and there is no evidence for bias being a valid predictor for test performance.

Research Questions and Hypotheses

Our first research question was whether the accuracy of confidence judgments is stable over time and whether it generalizes over two different tests within the same learning domain. We argued that the beliefs and theories of a person may cause some stability in judgments and the absolute accuracy of those judgments. On the basis of these thoughts and the evidence from the outlined studies, we hypothesized (1) that measures of absolute accuracy are stable over time, but measures of relative accuracy are not; and (2) with regard to generality, we anticipated no generality of relative accuracy over two different tests within the same knowledge domain. Since it is assumed that the accuracy of monitoring is important for effective learning (Dunlosky, Hertzog, et al., 2005), valid measures of the accuracy of monitoring during learning should predict the learning outcome; (3) hence, we expected the accuracy of confidence judgments during the learning process to be prognostic for the learning outcome.

METHOD

Participants

A total of 113 students from a German university participated in the study; 86 (76%) were female. The mean age of the participants

was 23.6 years old (SD = 5.2). A total of 65 (58%) participants were studying psychology; 39 (35%) were studying education, and 9 (8%) were studying other social sciences. With a mean of 2.9 semesters—1.5 years of study time—the participants were rather at the beginning of their studies. The students received either €15 for participating in the study or a certificate needed for their studies.

Learning Material

By means of a hypermedia environment that was written in German, the students learned the principles and the application of operant conditioning. The hypermedia environment has been used with comparable samples and learning settings in previous studies (Bannert, 2006), and its intermediate difficulty has been proven. The hypertext consisted of 44 nodes with about 12,500 words, 19 pictures/diagrams, and 240 links in total. The part that was relevant for the learning task involved 9 nodes including 2,300 words, 3 pictures, and 60 links. Navigation was possible by using the hierarchical navigation menu, the forward and backward buttons on each node, and the hotwords placed directly in the text.

Instruments

Knowledge tests. Three types of learning outcomes are distinguished using Bloom's (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1973) taxonomy of learning objectives as a framework to generate items about operant conditioning:

1. Knowledge about facts: To answer these items, it was necessary to remember the text base (see Graesser, Millis, & Zwaan, 1997, for levels of comprehension). The participants were usually able to answer the questions without knowing the exact wording that had been used in the learning material, and it was not necessary to infer anything that was not written in the learning material. The students had to select one out of four responses to solve the items.

2. Comprehension: To answer these questions, it was necessary to know at least two facts and to draw an inference from them, because the right answers to these questions could not be found directly in the text. According to Graesser et al. (1997), these questions mostly refer to situational models; that is, a situational model was needed or had to be constructed to solve the tasks. Again, an answer format with three distractors was chosen.

3. Transfer: Questions concerning transfer required the application of knowledge to a new situation that was not mentioned in the learning material. Short texts were presented that described a situation in which operant conditioning occurred. Here, the answer format was open: The participants had to write one or two words to indicate who was conditioned by whom, what the stimulus and what the reaction were, and which principle of operant conditioning was used. Answers were rated by two raters, independently.

No items were constructed for the three highest objectives in Bloom et al.'s (1973) taxonomy because the learning goals did not include analysis, synthesis, or evaluation objectives. Three tests that were used at different times during the study were composed from the items on knowledge about facts and comprehension. The pretest (PT) consisted of 10 items. The intermediate test (IT) consisted of 20 new items. The final comprehension test (FCT) included the 10 items from the PT and 10 new items that had been used neither in the PT nor in the IT. Each item of the FCT corresponded to an item of the IT; that is, they required the same information that was given at the same node of the hypermedia. Additionally, a final transfer test (FTT) with 11 transfer items followed. Examples for the items from the comprehension tests and the FTT can be found in the Appendix.

Retrospective confidence judgments. For each item on the tests, the participants were asked retrospectively to judge their confidence that their answers were correct. Following studies on test taking (e.g., Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 1998), percentage scales were used to assess the confidence judgments. That is, the lowest possible percentage was 25% for the four-alternative multiple choice test, ranging up to 100% as the highest possible percentage. In the case of the transfer test with its open-answer format, the scale ranged from 0% up to 100%.

Design and Procedure

First, the participants practiced navigating in the hypertext using another chapter about educational psychology. After that, the PT was filled in, and the participants began to study the chapter about operant conditioning. The participants learned for 10 min; then, they took the IT before learning again for another 20 min. After that, the two final tests (FCT and FTT) took place, and the participants were informed about the aims of the study and were offered feedback about their performance.

Remarks on Statistical Analyses

All of the statistics were calculated using R. Two measures of absolute accuracy and three measures of relative accuracy were calculated. Bias and absolute bias were calculated for each person using the two following formulas:

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n}\left(c_i - p_i\right)$$

and

$$\text{abs.bias} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(c_i - p_i\right)^2},$$

where n = number of items; c_i = confidence judgment about the item i; and p_i = performance on the item i.

In the case of the FTT with an open-answer format, no measures of absolute accuracy were calculated. Although the FTT yielded an objective and reliable score, those data were obtained on an interval level of measurement instead of the ratio level of measurement obtained by the multiple-choice tests. Therefore, the value of 0 is not fixed on the performance scale, and calculating differences between judgment and performance makes no sense.

As a first measure of relative accuracy, the nonparametric correlation gamma between confidence judgments and performance in the items was calculated within each person (see Gonzalez & Nelson, 1996). Second, the within-person correlation Pearson's r was calculated as a relative measure. Third, d_a was calculated as proposed by Benjamin and Diaz (2008). Using their procedure, the confidence judgments and items were recoded into four bins. Each bin contained one quarter of the items, ordered from the items judged lowest to the items judged highest. The first row contained the number of correct answers, and the second row contained the number of incorrect items. In cases in which a person had values of 0 in one cell of the 2 × 4 data array—for example, no incorrect answer in the first bin—no value for d_a could be calculated. Furthermore, if a person had equal values for all judgments or had solved all items correctly, no measures of relative accuracy could be calculated. Thus, these participants were excluded from some analyses.

RESULTS

Instruments

The statistics for the scales of the instruments used are listed in Table 1. The alphas for the knowledge tests were far from optimal but were sufficiently reliable for research purposes. All of the alphas for the confidence judgments were good or very good (D. H. Rost, 2005, p. 132). The FTT, with its open-answer format, yielded intraclass coefficients adjusted for systematic variance between raters (ICC_a) between .86 and .97, with a mean ICC_a of .93 for all 11 items. Thus, the rater reliability was rather good, and the internal consistency of .79 was also good.

Table 1
Descriptive Statistics of the Instruments ($N = 113$)

	Number of Items	M[a]	SD	α
Knowledge				
Pretest (PT)	10	.58	.21	.55
Intermediate test (IT)	20	.65	.15	.62
Final comprehension test (FCT)	20	.79	.16	.73
Final comprehension test (repeated)[b]	10	.80	.18	.56
Final transfer test (FTT)[c]	11	.75	.09	.79
Confidence Judgments[d]				
PT	10	.56	.19	.89
IT	20	.62	.16	.90
FCT	10	.82	.13	.89
FTT[e]	10	.61	.22	.89

[a]Proportion of items solved correctly or, for confidence judgments, estimated by participants. [b]Contains the same 10 items as the PT. [c]Proportion of points achieved (a maximum of 10 points per item was assigned). [d]Scale from .25 up to 1.00. [e]Scale from 0 up to 1.00.

Descriptive statistics for the measures of accuracy can be found in Table 2. For bias and absolute bias, a value of 0 indicates perfect accuracy. For bias, the value of one indicates a maximum of overconfidence, and a value of −1 indicates a maximum of underconfidence. For the relative-measures gamma and Pearson's r, perfect accuracy is indicated by a value of 1, and 0 indicates no accuracy, whereas negative values indicate that a person judges easier items as difficult, and vice versa. The value of d_a is bounded at 0 and ∞, with zero indicating no accuracy.

Shapiro–Wilk normality tests indicated a significant deviation from the normal distribution for the values of some scales and measures of accuracy. But an inspection of the histograms showed that the distributions violate the assumption of normality only slightly. Nevertheless, all further analyses were calculated using nonparametric correlations.

Initial Analyses

Before we discuss the testing of our hypotheses, some initial analyses concerning the performances, the judgments, and the resulting accuracy measures are reported. To inspect the data, we plotted calibration curves that can be seen in Figure 1 (see Stankov, 1998, for calibration curves). The deviance of the calibration curve from the dotted line indicates over- and underconfidence. In all three tests, a slight amount of overconfidence for the lower judgments and a marginal amount of underconfidence for the upper judgments were found. Over- or underconfidence was tested using two-tailed paired-sample t tests to compare the actual performance with the judged performance. Neither over- nor underconfidence was found in the PT [bias = −.02, $t(112) = -1.14$, n.s.], or in the IT [bias = −.02, $t(112) = -1.59$, n.s.]. But, there was a small, statistically significant amount of overconfidence in the FCT [bias = .03, $t(112) = 2.58$, $p < .05$, $\eta^2 = .06$]. Thus, the calibration curves and the t tests show that the participants were able to judge their performance quite accurately.

Furthermore, the shape of the curves remained the same over all three tests, but the distribution of judgments changed over time, as can be seen from the frequencies of the judgment categories.[1] That is, before learning, more low judgments were made, and after learning, more high judgments were given. According to a Monte Carlo study by Weaver and Kelemen (1997), a shift to the end of the distribution may alter the measure of relative accuracy, even when there is no change in the genuine metacognitive accuracy of the judgments. With this potential artifact in mind, we analyzed the measures of relative accuracy. All values for the three relative measures (gamma, Pearson's r, and d_a) in all tests reached values significantly different from 0, as can be seen in Table 3. Thus, for all tests, the judgments were considerably accurate, indicating that the participants were able to judge their knowledge at least better than chance. However, the gain in accuracy from the PT to the FCT and FTT may be at least partly due to the artifact described by Weaver and Kelemen. Since the increase in accuracy is of no interest for the calculation of correlations, we will not consider this problem any further.

As was described earlier, the items in the IT and the FCT were not the same, but one item from each test refers to the same knowledge. We correlated all items between all persons to each other. According to a two-tailed Welch test for independent samples, the corresponding items yielded a greater mean gamma correlation [$M(\gamma) = .44$, $SD(\gamma) = .25$] than the noncorresponding items [$M(\gamma) = .36$, $SD(\gamma) = .23$] [$t(28) = 2.77$, $p < .01$]. Thus, even though the items in the two tests were not the same, we have evidence that the items corresponded to each other.

Stability

The investigation of strict stability is not appropriate in learning processes because it is inherent to learning

Table 2
Descriptive Statistics of the Measures of Accuracy

Measure of Accuracy	Pretest			Intermediate Test			Final Comprehension Test			Final Transfer Test		
	N	M	SD	N	M	SD	N	M	SD	N	M	SD
Bias	113	−.02	.18	113	−.02	.15	113	.03	.14	–	–	–
Absolute bias	113	.47	.11	113	.45	.08	113	.37	.12	–	–	–
Gamma	101	.37	.51	112	.57	.27	105	.62	.38	110	.30	.33
Pearson's r	101	.25	.34	112	.37	.20	105	.39	.23	110	.30	.31
d_a	72	.25	.55	62	.54	.38	52	.57	.38	–	–	–

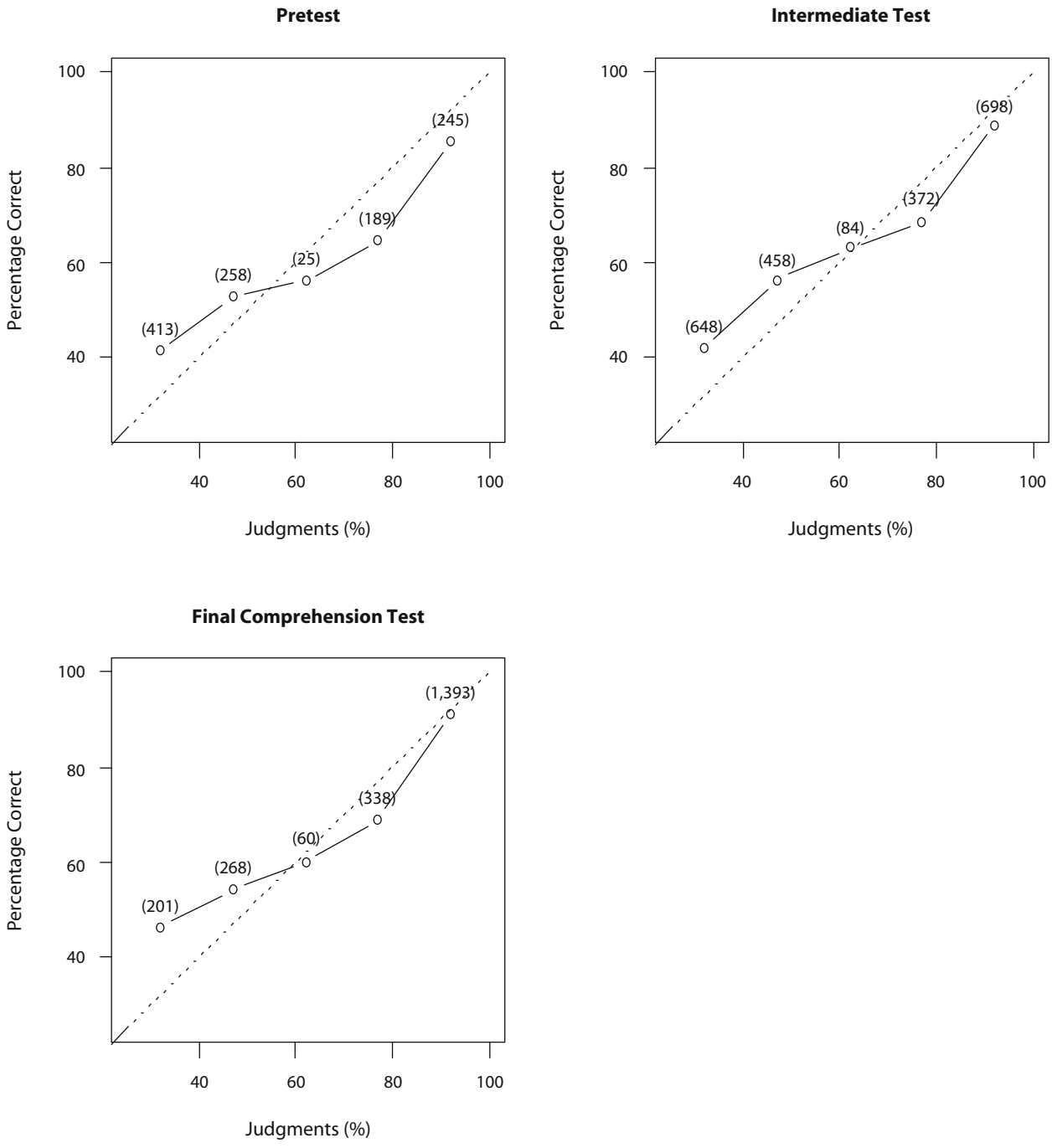

Figure 1. Calibration curves with the dotted line indicating perfectly accurate judgments. Frequencies are printed in parentheses.

that knowledge increases and is not on a stable level. This is also the case in the present study. The participants' knowledge increased during the learning process, as was demonstrated by a comparison of the PT (M = .58, SD = .21) with the repeated version of the FCT (M = .80, SD = .18) (Wilcoxon z = 8.01, $p < .001$). But, the investigation of monotonic stability is reasonable; that is, the rank order of learners according to their knowledge may be stable. The argument against the assumption of strict stability is true for the accuracy of the judgments as well as, since all measures of accuracy were based on different tests with different difficulties (see Table 1), and since test difficulty affects the value of relative accuracy (Pressley & Ghatala, 1988). Furthermore, the length of the test also influences the relative accuracy (Weaver, 1990, Experiment 2). Therefore, no strict stability can be expected in the accuracy measures, but correlations between measures of accuracy at differ-

Table 3
Independent One-Sample *t* Tests for Gamma, Pearson's *r*, and d_a

	Gamma			Pearson's *r*			d_a		
	γ	*t*	*df*	*r*	*t*	*df*	d_a	*t*	*df*
PT	.37	7.25***	100	.25	7.34***	100	.25	3.79***	71
IT	.57	22.64***	111	.37	19.43***	111	.54	11.25***	61
FCT	.62	16.91***	104	.39	17.35***	104	.57	11.35***	54
FTT	.30	9.56***	109	.30	10.09***	109	–	–	–

Note—PT, pretest; IT, intermediate test; FCT, final comprehension test; FTT, final transfer test. ***$p < .001$.

ent points in time can be calculated to test whether there is any monotonic stability.

Before the stability of the accuracies is reported, the stability of the underlying judgments and performances will be described. In Table 4, the correlations between the performances across all tests are listed in the second column. All correlations between tests at different points in time were significant, indicating that despite the absolute gain in knowledge, the participants' rank order of knowledge was at least partly stable over time. Furthermore, Table 4 contains the correlations between judgments across all tests in the third column. Descriptively speaking, the correlations were consistently higher than those for the performances, indicating an even higher stability in the confidence judgments about knowledge than in the knowledge itself.

To test our first hypothesis concerning the stability of accuracies, gamma correlations were calculated between the measures of accuracy derived before, during, and after learning. Correlations between biases and absolute biases are listed in Table 4, in columns four and five. All correlations between biases were significant, and all correlations between absolute biases were significant as well, indicating some stability over time. For relative accuracy, the gamma correlations are presented in the upper part of Table 5. Most of these correlations were not significantly greater than 0, indicating no stability at all. Two correlations revealed significantly negative values that do not support any stability, either. There was only one significant positive correlation between the IT and the FCT for gamma, indicating some stability over time. To sum up, relative accuracy was not stable over time at all, but absolute accuracy was.

Comparing the correlations for gamma and Pearson's *r* reveals few differences between the two measures of relative accuracy. But for d_a as a measure of relative accuracy, the picture changes somewhat. Descriptively, the correlation between d_a in the PT and the IT was not 0 as it was for gamma and *r*, but was negative. This result may be due to the reduced sample. A comparison between the 48 participants for which d_a could be calculated and the remaining 65 participants showed that the 48 participants had less knowledge ($M = .22$ vs. $M = .36$) in the PT (Wilcoxon $W = 1{,}100.5$, $p < .001$). Furthermore, they had less knowledge ($M = .24$ vs. $M = .40$) in the IT ($W = 828.0$, $p < .001$); their judgments in the PT were lower ($M = .20$ vs. $M = .36$) ($W = 898.0$, $p < .001$), and their judgments in the IT were also lower ($M = .23$ vs. $M = .39$) ($W = 865.5$, $p < .001$). Accordingly, the reduced sample differed from the remaining participants in all four measures, and the reduced sample is not representative of the whole sample. Furthermore, none of the correlations between the values for d_a reached significance. This may also be due to the nonrepresentativeness of the sample. Additionally, the power of the tests is reduced, since the *N*s are clearly smaller than these for gamma and Pearson's *r*. To support the explanation that the different results for d_a were caused by the reduced sample, the three measures of relative accuracy were correlated to each other for the PT. The PT was chosen because the reduction of the sample was less than that in the IT and the FT; that is, a d_a value could be calculated for 72 participants. Results revealed Spearman correlations of .95 for all combinations of the three relative measures. Therefore, one may expect very similar results concerning stability and generality for all three relative measures of accuracy if d_a can be calculated for the complete sample.

Generality

Before the hypothesis concerning the generality of accuracies is addressed, the generality of the underlying two final tests and the confidence judgments shall be considered. As can be seen in the lower part of Table 4, the correlation between the FCT and the FTT was significant. This indicates that the two tests at least partly measured the same knowledge. Descriptively speaking, the correlation between the judgments was even higher than the correlation between the tests themselves.

In order to test our second hypothesis concerning generality, the relative accuracies in the two final tests were

Table 4
Gamma Correlations Across Tests for Performance, Judgments, and Absolute Measures of Accuracy ($N = 113$)

Tests	Performance	Judgment	Bias	Absolute Bias
		Stability		
PT–IT	.47***	.64***	.21**	.23***
PT–FCT	.53***	.53***	.20**	.23***
PT–FTT	.34***	.36***	–	–
IT–FCT	.50***	.61***	.26***	.34***
IT–FTT	.41***	.47***	–	–
		Generality		
FCT–FTT	.47***	.54***	–	–

Note—PT, pretest; IT, intermediate test; FCT, final comprehension test; FTT, final transfer test. **$p < .01$. ***$p < .001$.

Table 5
Gamma Correlations Across Tests for Relative Measures of Accuracy

	Gamma		Pearson's *r*		d_a	
Tests	*N*	γ	*N*	γ	*N*	γ
			Stability			
PT–IT	101	.06	101	−.02	48	−.17
PT–FCT	96	.03	96	−.04	40	.08
PT–FTT	99	−.16*	99	−.14*	–	–
IT–FCT	105	.15*	105	.09	37	.17
IT–FTT	110	−.11	110	−.05	–	–
			Generality			
FCT–FTT	103	.15*	103	.13*	–	–

Note—PT, pretest; IT, intermediate test; FCT, final comprehension test; FTT, final transfer test. *$p < .05$.

correlated. The correlations can be found in the lower part of Table 5. Correlations between relative accuracy in the FCT and the FTT were significant for gamma as well as for Pearson's *r*. These correlations indicate some generality of relative accuracy between the two tests. Note that these tests measured comprehension and transfer on the basis of the same learning material, but with different items and different answer formats. Thus, relative accuracy generalized at least partly between different tests administered at the same time.

Predictive Validity

To avoid the methodological problem of part–whole correlations mentioned by Hasselhorn and Hager (1989), no performance was predicted with absolute measures of accuracy that were derived from the same test. Thus, the aim of the analyses was to predict the performance in the FCT and FTT from measures of accuracy in the IT—that is, from measures of accuracy during the learning process.

For absolute accuracy, the bivariate correlation between bias in the IT and the FCT was $r_s(111) = -.18$, n.s. For absolute bias, the correlation was $r_s(111) = -.42$, $p < .001$. Taking the FTT as a criterion changed the correlations for bias [$r_s(111) = -.11$, n.s.] and for absolute bias [$r_s(111) = -.43$, $p < .001$]. Thus, bias was not a valid predictor for learning outcome, but absolute bias was. For the relative measures of accuracy, the bivariate correlation between gamma in the IT and the FCT was $r_s(110) = .20$, $p < .05$; for Pearson's *r* the correlation was $r_s(110) = .12$, n.s.; and for d_a, the correlation was $r_s(60) = -.05$, n.s. For the FTT, the predictive validity for gamma was $r_s(110) = .26$, $p < .01$; for Pearson's *r*, the validity was $r_s(110) = .21$, $p < .05$; and for d_a, the validity was $r_s(60) = -.11$, n.s. Thus, relative accuracy was a valid predictor for the learning outcome for gamma and partly for Pearson's *r*, but d_a failed to show its predictive validity. As was argued above, this may be due to the restricted sample that was not representative of the whole sample.

DISCUSSION

In the present study, we analyzed whether different measures for the accuracy of confidence judgments are stable over learning time, whether they generalize over different tests within the same domain, and whether they predict the learning outcome. Concerning stability over time, our results confirm our first hypothesis assuming that absolute accuracies were considerably stable, since all correlations were significant. In contrast, all three relative accuracies were not stable over time, with the exception of gamma correlating significantly between the IT and the FCT. This finding is not due to the use of a specific measure within these two classes of accuracy measures.

Moreover, bias derived from predictive judgments has some stability as well (Kelemen et al., 2000). One reason for the stability of absolute accuracies could be that they are not independent of the two underlying measures of knowledge and confidence (see Nelson, 1984). That is, variance in the performance test and variance in the confidence judgments are part of the measures of absolute accuracy, due to their calculation. Therefore, the stability in absolute accuracy may purely reflect the stability of the performance and the judgments. In accordance with results from studies on test taking (see, e.g., Jonsson & Allwood, 2003; Schraw & Nietfeld, 1998), the rank order of judgments in this sample was even more stable than the rank order of the performance itself. So far, there is no method, to the authors' knowledge, that solves the described problem of the results potentially being a statistical artifact.

In contrast with absolute measures of accuracy, relative measures of accuracy do not suffer from the problem of inherited variance, as was shown for gamma (Nelson, 1984). Thus, the findings for relative accuracy are not artifacts but can be interpreted unrestrictedly. Almost no evidence for the stability of relative accuracy was found, either with retrospective judgments in the present study or with predictive judgments and gamma in the study by Kelemen et al. (2000). Instead, the accuracy of monitoring seems to depend strongly on the situation. This is in accordance with our assumption that relative accuracy is dependent on experience-based cues rather than on theory-based cues, and it is true despite the considerable stability of the judgments themselves. Low stability enables the manipulation of relative accuracy, as was done in the study by Thiede et al. (2003). Further interventions that increase the accuracy of judgments have been reviewed recently (Dunlosky & Lipko, 2007).

The second hypothesis stated that relative accuracy does not generalize over different tests. Results showed that there was a significant correlation between the gammas and Pearson's *r* for the FCT and the FTT. At first glance, this result contradicts the findings with gamma and retrospective judgments (Pressley & Ghatala, 1988), as well as findings with predictive judgments (Glenberg & Epstein, 1987; Kelemen et al., 2000). However, all of these studies used tests of different knowledge domains from which to calculate the gammas. Our interpretation of the findings is that relative accuracy is domain specific; thus, relative accuracy generalizes over different tests within a domain, but not between domains. This interpretation is supported by findings about theory-based monitoring (see Koriat, 2007), indicating that, among

other sources, domain knowledge can be used to generate judgments. Our interpretation is supported further by a metacomprehension study (Moore, Lin-Agler, & Zabrucky, 2005) in which the retrospective judgments of comprehension over three trials were influenced much more by prior judgments than by the performance within each trial. Moore et al. (2005, p. 261) concluded that the judgments are based on "a general perception of oneself as being a good or poor reader."

The third hypothesis of the present study stated that the accuracy measures should be predictive for learning outcome. For bias, no significant bivariate correlations between the accuracy in the IT and in the final tests were found, but the correlations were significant for absolute bias. The former result was also found in a laboratory study (Maki, 1998). The latter result parallels findings from classroom studies that showed predictive validity for absolute bias (Nietfeld et al., 2005, 2006). One explanation for the nonsignificance of bias is that bias reflects over- and underconfidence. We based our assumption of predictive validity on the model of effective learning (Dunlosky, Hertzog, et al., 2005). The model does not differentiate between over- and underconfidence; that is, in both cases, monitoring is inaccurate, and the learning process is not controlled effectively. Therefore, the model does not predict the validity of bias, but the validity of absolute bias; hence, the results in the present study are in accordance with the hypotheses drawn from the model.

For relative accuracy, significant correlations between gamma in the IT and in both final tests were found. For Pearson's *r*, the correlation with the FCT was not significant, but the correlation with the FTT was. The finding that relative accuracy is a valid predictor for learning outcome corresponds with results from metacomprehension research, which also have shown significant correlations for gamma (see, e.g., Maki, 1998; Maki et al., 1994), and is in accordance with the model by Dunlosky, Hertzog, et al. (2005). For d_a, no significant correlation was found, but this may be due to the smaller sample size.

A methodological problem concerning the stability, generality, and validity of the measures of accuracy arises from the confounding of these coefficients with the reliability of the measures. That is, a correlation cannot exceed the minimum of the square roots of the reliabilities (J. Rost, 2004, p. 389). Thus, for future research on stability, generality, and validity of accuracy measures, we suggest using structural equation models to estimate the accuracy at a latent level, thereby eliminating errors of measurement (see Schilling, 2005). Doing this requires at least four measures in time for each person and larger samples than in the present study. Bias can be estimated as a latent variable using models of true change (see, e.g., Little, Bovaird, & Slegers, 2006). To gain an estimation of latent relative accuracy, we suggest obtaining two relative accuracies at one point in time using two parallel tests.

Despite this statistical discussion, our present study highlighted the role of different classes of measures of accuracy. Apparently, absolute and relative measures of accuracy produced different results concerning their stability, and this has not been investigated systematically so far. That is, the absolute accuracy of confidence judgments was stable, whereas the relative accuracy showed no stability. In our view, it seems promising to investigate the idea of theory-based versus experience-based cues (Koriat, 2007) more deeply; for example, measures of self-efficacy should be used to test our rationale that theory-based cues cause stability in absolute accuracies.

AUTHOR NOTE

The present article was mainly written during a research stay of M.B. at the CoCo Research Centre, University of Sydney, Australia, which was supported by funds from the German Science Foundation (DFG: BA 2044/5-1). We thank Susanne Narciss and the anonymous reviewers for their valuable advice and comments on the manuscript. Address correspondence to C. Mengelkamp, University of Koblenz-Landau, Fortstrasse 7, 76829 Landau, Germany (e-mail: mengelkamp@uni-landau.de).

REFERENCES

BANNERT, M. (2006). Effects of reflection prompts when learning with hypermedia. *Journal of Educational Computing Research*, **35**, 359-375. doi:10.2190/94V6-R58H-3367-G388

BENJAMIN, A. S., & DIAZ, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 73-94). New York: Psychology Press.

BLOOM, B. S., ENGELHART, M. D., FURST, E. J., HILL, W. H., & KRATHWOHL, D. R. (1973). *Taxonomie von Lernzielen im kognitiven Bereich* [Taxonomy of educational objectives] (E. Füner & R. Horn, Trans., 3rd ed.). Weinheim, Germany: Beltz.

BREWER, W. F., SAMPAIO, C., & BARLOW, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory & Language*, **52**, 618-627. doi:10.1016/j.jml.2005.01.017

BYRNE, B. M. (1986). Self-concept/academic achievement relations: An investigation of dimensionality, stability, and causality. *Canadian Journal of Behavioural Science*, **18**, 173-186.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

DUNLOSKY, J., & HERTZOG, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 249-275). Mahwah, NJ: Erlbaum.

DUNLOSKY, J., HERTZOG, C., KENNEDY, M. R. T., & THIEDE, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology*, **10**, 4-11.

DUNLOSKY, J., & LIPKO, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, **16**, 228-232. doi:10.1111/j.1467-8721.2007.00509.x

DUNLOSKY, J., RAWSON, K. A., & MCDONALD, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 68-92). Cambridge: Cambridge University Press.

DUNLOSKY, J., RAWSON, K. A., & MIDDLETON, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory & Language*, **52**, 551-565. doi:10.1016/j.jml.2005.01.011

EHRLINGER, J., & DUNNING, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality & Social Psychology*, **84**, 5-17. doi:10.1037/0022-3514.84.1.5

GLENBERG, A. M., & EPSTEIN, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, **15**, 84-93.

GLENBERG, A. M., SANOCKI, T., EPSTEIN, W., & MORRIS, C. (1987). En-

hancing calibration of comprehension. *Journal of Experimental Psychology: General*, **116**, 119-136. doi:10.1037/0096-3445.116.2.119

GONZALEZ, R., & NELSON, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin*, **119**, 159-165. doi:10.1037/0033-2909.119.1.159

GRAESSER, A. C., MILLIS, K. K., & ZWAAN, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, **48**, 163-189. doi:10.1146/annurev.psych.48.1.163

HACKER, D. J., BOL, L., & BAHBAHANI, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition & Learning*, **3**, 101-121. doi:10.1007/s11409-008-9021-5

HASSELHORN, M., & HAGER, W. (1989). Prediction accuracy and memory performance: Correlational and experimental tests of a metamemory hypothesis. *Psychological Research*, **51**, 147-152. doi:10.1007/BF00309310

JONSSON, A.-C., & ALLWOOD, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality & Individual Differences*, **34**, 559-574. doi:10.1016/S0191-8869(02)00028-4

KELEMEN, W. L., FROST, P. J., & WEAVER, C. A., III (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, **28**, 92-107.

KELLEY, C. M., & LINDSAY, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory & Language*, **32**, 1-24. doi:10.1006/jmla.1993.1001

KLEITMAN, S., & STANKOV, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology*, **15**, 321-341. doi:10.1002/acp.705

KLEITMAN, S., & STANKOV, L. (2007). Self-confidence and metacognitive processes. *Learning & Individual Differences*, **17**, 161-173. doi:10.1016/j.lindif.2007.03.004

KORIAT, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289-325). Cambridge: Cambridge University Press.

KORIAT, A., NUSSINSON, R., BLESS, H., & SHAKED, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 117-135). New York: Psychology Press.

LANE, J., & LANE, A. [M.] (2001). Self-efficacy and academic performance. *Social Behavior & Personality*, **29**, 687-693. doi:10.2224/sbp.2001.29.7.687

LANE, J., LANE, A. M., & KYPRIANOU, A. (2004). Self-efficacy, self-esteem and their impact on academic performance. *Social Behavior & Personality*, **32**, 247-256. doi:10.2224/sbp.2004.32.3.247

LEONESIO, R. J., & NELSON, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 464-470. doi:10.1037/0278-7393.16.3.464

LITTLE, T. D., BOVAIRD, J. A., & SLEGERS, D. W. (2006). Methods for the analysis of change. In D. K. Mroczek & T. D. Little (Eds.), *Handbook of personality development* (pp. 181-211). Mahwah, NJ: Erlbaum.

MAKI, R. H. (1995). Accuracy of metacomprehension judgments for questions of varying importance levels. *American Journal of Psychology*, **108**, 327-344. doi:10.2307/1422893

MAKI, R. H. (1998). Metacomprehension of text: Influence of absolute confidence level on bias and accuracy. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 38, pp. 223-248). San Diego: Academic Press.

MAKI, R. H., JONAS, D., & KALLOD, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review*, **1**, 126-129.

MENGELKAMP, C., & BANNERT, M. (2009). Judgements about knowledge: Searching for factors that influence their validity. *Electronic Journal of Research in Educational Psychology*, **7**, 163-190.

METCALFE, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, **131**, 349-363. doi:10.1037/0096-3445.131.3.349

MOORE, D., LIN-AGLER, L. M., & ZABRUCKY, K. M. (2005). A source of metacomprehension inaccuracy. *Reading Psychology*, **26**, 251-265. doi:10.1080/02702710590962578

NELSON, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, **95**, 109-133. doi:10.1037/0033-2909.95.1.109

NELSON, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology*, **10**, 257-260.

NELSON, T. O., & DUNLOSKY, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, **2**, 267-270. doi:10.1111/j.1467-9280.1991.tb00147.x

NELSON, T. O., & NARENS, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125-173). New York: Academic Press.

NELSON, T. O., & NARENS, L. (1992). Metamemory: A theoretical framework and new findings. In T. O. Nelson (Ed.), *Metacognition: Core readings* (pp. 117-130). Needham Heights, MA: Allyn and Bacon.

NIETFELD, J. L., CAO, L., & OSBORNE, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal of Experimental Education*, **74**, 7-28.

NIETFELD, J. L., CAO, L., & OSBORNE, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition & Learning*, **1**, 159-179.

NIETFELD, J. L., & SCHRAW, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *Journal of Educational Research*, **95**, 131-142.

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. *Journal of General Psychology*, **129**, 257-299.

PRESSLEY, M., & GHATALA, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly*, **23**, 454-464. doi:10.2307/747643

ROST, D. H. (2005). *Interpretation und Bewertung pädagogisch-psychologischer Studien. Eine Einführung* [Interpretation and evaluation of studies in educational psychology. An introduction]. Weinheim, Germany: Beltz.

ROST, J. (2004). *Lehrbuch Testtheorie—Testkonstruktion* [Test theory and test construction textbook]. Bern, Switzerland: Huber.

SAMPAIO, C., & BREWER, W. F. (2009). The role of unconscious memory errors in judgments of confidence for sentence recognition. *Memory & Cognition*, **37**, 158-163. doi:10.3758/MC.37.2.158

SCHILLING, O. (2005). Strukturgleichungsmodelle zur Analyse von Simplex-Strukturen [Structural equation models for the analysis of simplex structures]. In J. Werner (Ed.), *Zeitreihenanalysen mit Beispielen aus der Psychologie* (pp. 37-72). Berlin: Logos.

SCHRAW, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition & Learning*, **4**, 33-45. doi:10.1007/s11409-008-9031-3

SCHRAW, G., DUNKLE, M. E., BENDIXEN, L. D., & DEBACKER ROEDEL, T. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology*, **87**, 433-444. doi:10.1037/0022-0663.87.3.433

SCHRAW, G., & NIETFELD, J. [L.] (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology*, **90**, 236-248. doi:10.1037/0022-0663.90.2.236

SON, L. K., & METCALFE, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 204-221. doi:10.1037/0278-7393.26.1.204

STANKOV, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning & Individual Differences*, **10**, 29-50. doi:10.1016/S1041-6080(99)80141-1

THIEDE, K. W., ANDERSON, M. C. M., & THERRIAULT, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, **95**, 66-73. doi:10.1037/0022-0663.95.1.66

THIEDE, K. W., & DUNLOSKY, J. (1999). Toward a general model

of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1024-1037. doi:10.1037/0278-7393.25.4.1024

THOMPSON, W. B., & MASON, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition*, **24**, 226-234.

TOBIAS, S., & EVERSON, H. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 147-222). Lincoln, NE: Buros Institute of Mental Measurements.

WEAVER, C. A., III (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 214-222. doi:10.1037/0278-7393.16.2.214

WEAVER, C. A., III, & KELEMEN, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *Psychological Science*, **8**, 318-321. doi:10.1111/j.1467-9280.1997.tb00445.x

YATES, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

NOTE

1. The low frequencies for the middle category of judgments are due to the construction of the categories. Participants tended to use judgments such as 25%, 50%, or 75%. As the middle category reaches from 55% up to 69%, neither 50% nor 75% is included in the interval; thus, the frequencies are rather low.

APPENDIX

Example of an Item From the Comprehension Test

Positive reinforcement can be obtained with…

- ☐ a positive consequence, but not with the withdrawal of an aversive stimulus.
- ☐ the withdrawal of an aversive stimulus, but not with a positive consequence.
- ☐ a positive consequence as well as with the withdrawal of an aversive stimulus.*
- ☐ a positive consequence as well as with an aversive stimulus.

*Correct answer

Example of an Item From the Transfer Test

In many families, getting the children to bed is not easy. They get out of bed, want something to eat or drink, want their parents to sing a song for them, etc. This scenario is repeated several times in one evening, and the parents become annoyed. If the parents decide not to respond to the wishes of their children but do not do this consistently, the children will behave in an even more bothersome manner.

Acting persons:	parents influence children
Influenced behavior:	getting out of bed, wanting something to drink, etc.
Stimulus:	eating, drinking, hearing a song
Principle:	variable interval periodic reinforcement

(The underlined passages had to be filled in by the participants.)

(Manuscript received May 28, 2009;
revision accepted for publication October 29, 2009.)