EDUCATION & PSYCHOLOGY I+D+i
Research, Innovation and Solutions on-line®
Electronic Journal of Research in Educational Psychology
ejrep
Editorial EOS

Judgements about Knowledge: Searching for Factors that Influence their Validity[1]

Christoph Mengelkamp[1], Maria Bannert[2]

[1] Arbeiststelle Multimedia, University of Koblenz-Landau, Landau
[2] Educational Media, Chemnitz University of Technology, Chemnitz

Germany

Christoph Mengelkamp. Universitaet Koblenz-Landau. Arbeitsstelle Multimedia. Thomas-Nast-Strasse 44. 76829 Landau. Germany. E-mail: mengelkamp@uni-landau.de

[1] Paper was submitted during a research stay of the second author at the CoCo Research Centre, University of Sydney, Australia, which was supported by funds from the German Science Foundation (DFG: BA 2044/5-1).

Abstract

Introduction. Research in the field of metacomprehension often uses the accuracy of judgements of own knowledge as a measure of monitoring which is a central component of metacognition. One aim of this study is to investigate if the accuracy above chance usually found in studies using traditional texts can be replicated with hypermedia. More importantly, the study investigates differences between the accuracy of monitoring in two kinds of tests: a comprehension test and a transfer test.

Method. To investigate these questions, 126 university students learnt the basic concepts of operant conditioning presented in a hypermedia system within 30 minutes. Immediately afterwards, a comprehension test and a transfer test were administered. For each test, the students predicted their knowledge just before it was administered and retrospectively judged their confidence regarding each item.

Results. The results showed that the students' accuracy of monitoring was above chance. Further, accuracy of retrospective judgements was higher in the comprehension test than in the transfer test, and the predictive judgement was less accurate than retrospective judgements in the comprehension test, but not in the transfer test. Structural equation modelling rendered no evidence that this difference in accuracy between the two tests was due to differences in the use of experience-based cues.

Discussion and Conclusion. The results are discussed with regard to the distinction between theory-based and experience-based cues and with regard to the accessibility hypothesis.

Keywords: metacognition, monitoring, accuracy, judgements, accessibility hypothesis

Received: 12/15/08 *Initial Acceptance:* 12/15/08 *Final Acceptance:* 03/15/09

Resumen

Introducción. La investigación en el campo de la metacomprensión utiliza normalmente la precisión de los juicios sobre el conocimiento propio como medida de supervisión, que es es un componente central de la metacognición. Un objetivo de este estudio es investigar si la precisión de tales juicios usualmente utilizada con textos tradicionales puede replicarse con hipertextos. Aún más importante, el estudio investiga las diferencias entre la precisión de la supervisión a través de dos tipos de pruebas: una de comprensión y otra de transferencia.

Método. Para estudiar tales cuestiones, 126 estudiantes universitarios aprendieron, durante 30 minutos, los conceptos básicos del condicionamiento operante presentados a través de hipertextos. Inmediatamente después, realizaron un test de comprensión y otro de transferencia. En cada test, los estudiantes predijeron su conocimiento justo antes de realizarlo y, retrospectivamente, juzgaron su certidumbre respecto a cada ítem.

Resultados. Los resultados mostraron que la precisión de los estudiantes no se debió a al azar. Además, la precision de los juicios retrospectivos fue mayor en los test de comprensión que en los test de tranferencia, y el juicio predictivo fue menos certero que los juicios restrospectivos en el test de comprensión aunque no en el de transferencia. El modelo de ecuaciones estructurales no apunta evidencias de que la diferencia, en relación con la precisión, entre ambos test se deba a diferencias debidas a experiencias previas.

Discusión y Conclusiones. Los resultados se discuten respecto a la distinción entre indicadores teóricos y experienciales, y, con respecto a la hipótesis de accesibilidad.

Palabras Clave*:* metacognición, supervisión, precisión, juicios, hipótesis de accesibilidad.

Recibido: 15/12/08 *Aceptación inicial:* 15/12/08 *Aceptación final:* 15/03/09

Introduction

The focus of this study lies on *metacognitive skills*, which are the procedural part of metacognition (Desoete, 2007; Hasselhorn, 2001; Veenman, 2005). Different terms have been used, e.g. metacognitive skills (Brown, 1980), executive control (Simons, 1986), and self-management of thinking (Jacobs & Paris, 1987). The common factor of all these definitions and terms is that they concern the executive part of the human memory consisting of processes such as "predicting, checking, monitoring, reality testing, and coordination and control of deliberate attempts to study, learn, or solve problems" (Brown, 1980, p. 454). Metacognitive skills can be split into two basic processes (Nelson & Narens, 1990, 1992): *monitoring* the memory means that information from an object-level is used to inform a higher meta-level about "what's going on" at the object-level and therefore enables one to observe, reflect on, and experience one's own cognitive processes. For example, a student may ask himself whether he has understood the last paragraph he read or is able to remember the facts that he should have learned so far. *Controlling* the memory means that information from the meta-level is used to alter cognitive processes at the object-level, e.g. changing to another cognitive learning strategy. During the learning process monitoring and control often occur simultaneously (Torrano Montalvo & Gonzáles Torres, 2004). Monitoring and control will only be effective will only be effective with respect to learning outcome if learners are able to monitor their memory accurately during learning, because otherwise the learner would not have a solid basis from which to control his learning process (Dunlosky, Hertzog, Kennedy, & Thiede, 2005). For example, if a learner judges that he did not understand a text passage but in fact did, he will allocate learning time ineffectively in re-reading the topic he has already grasped (cf. e. g.Metcalfe, 2002; Thiede, Anderson, & Therriault, 2003, for studies on study-time-allocation).

As far as the authors know, all studies on metacomprehension used conventional linear texts as learning material. In contrast, the present study used a hypermedia learning setting, Thus, one aim of the present study is to investigate whether accuracy above chance usually found in research with traditional texts (e.g. Dunlosky & Lipko, 2007; Thiede et al., 2003) can be replicated in a hypermedia learning environment. Because accuracy of monitoring is essential for effective self-regulated learning, several studies in research on metacomprehension investigated factors that influence the accuracy of monitoring, e.g. expertise of participants (Glenberg & Epstein, 1987), difficulty of texts (Weaver & Bryant, 1995), number of test

items (Weaver, 1990), and kinds of judgements (Dunlosky, Rawson, & Middleton, 2005). The second aim of the present study is to investigate the influence of the type of knowledge (comprehension vs. transfer) on the accuracy of monitoring. Before delving more deeply into theory and empirical findings, the paradigms of research on metacomprehension and how accuracy of monitoring is assessed are briefly described.

Paradigms in Research on Metacomprehension

Two paradigms from research on metacomprehension are similar to the methods used in the present study. The first paradigm is the *prediction of performance* (Glenberg, Sanocki, Epstein, & Morris, 1987; Morris, 1995), which is often called *calibration of comprehension* (Lin & Zabrucky, 1998, p. 346). In this paradigm, a text is presented to participants. After reading the complete text, or after each text paragraph, they have to judge how well they will perform in a comprehension test. The true performance is then measured, and the accuracy of the predictive judgement is calculated, mostly using gamma as a within-subject correlation between judgement and actual performance.

The second paradigm is called *calibration of performance* (Glenberg & Epstein, 1985, 1987). When using this paradigm, a text is presented to participants. After the text or passages in the text have been read a test is administered, and the participants are asked to judge the likelihood that their answers were correct retrospectively (confidence judgement).

Thus, calibration of performance uses *retrospective judgements* in contrast to the prediction of performance paradigm. Another possibility is to differentiate judgements with regard to the content they are made about. A *global judgement* is made about a wide range of content that is tested with a couple of items. For example, the test score can be judged before a knowledge test is taken. In contrast, *specific judgements* are made about each item of the test. Often, predictive judgements are also global judgements and retrospective confidence judgements are specific ones.

In order to understand how accuracy of monitoring is influenced by the kind of knowledge, respectively the kind of test, the hypotheses by Koriat and collegues (Koriat, 2007; Koriat, Nussinson, Bless, & Shaked, 2008) about the generation of judgements will be described shortly in the next section.

The Generation of Metacognitive Judgements

The *cue-utilisation-hypothesis* proposes that metacognitive judgements are not directly based on memory traces, but that inferences are drawn from different cues to generate judgements (Koriat, 2007). These inferences may be *theory-based* or *experience-based*. Theory-based judgements are generated on the basis of beliefs or memories whereas experience-based judgements are based on heuristics. One example for the former is the self-classification hypothesis. That is, "subjects are not actually assessing knowledge gained from a particular text; instead they are responding on the basis of beliefs about their abilities within a given domain" (Glenberg & Epstein, 1987, p. 91). Glenberg and Epstein argue that self-classification is one source to generate a metacognitive judgement from, especially when the comprehension test has not yet been taken, thus for predictive judgements. An example for experience-based metacognitive judgements is using the terms in the item (Reder & Ritter, 1992): the more familiar the terms in the items are, the more likely it is participants will have the feeling of knowing the answer. For example, if one is asked about Skinner's definition of the term "operant conditioning", one will feel as if one knows the answer to the item if one has heard or read the terms "operant conditioning" and "Skinner" together before. One fundamental difference is that theory-based judgements are not based on specific knowledge, but on general beliefs or domain knowledge, whereas experience-based judgements deploy cues from the specific information processing while learning or answering a question (cf. Koriat, 2007; Koriat et al., 2008).

In comprehension tests, terms that can be used as cues by the participants when generating a judgement are usually included in the item. For example, the item may contain terms like "positive reinforcement" and "positive consequences" the participant will remember from the learning material. In contrast, transfer tests hardly include any of these cues because transfer means the application of acquired knowledge to new situations not mentioned in the learning material. For example, the participant is asked how to act adequately in an educational situation according to learning theory. Terms used in the item like "child" or "parents" are of no use to the participant as a cue to generate a judgement about his answer because the terms were not included the learning material. Thus, the participants have to rely on theory-based sources only to generate the requested judgement rather than using cues from experience-based sources like the terms in the question.

In the next section, major findings from research on metacomprehension with traditional texts are reviewed, and it is argued that the replication of these results with hypermedia material should be tested empirically. Furthermore, it is argued that differences in accuracy between kinds of judgements are not expected to be found with a transfer test, but are with a comprehension test.

Findings from Research on Metacomprehension and their Replication with Hypermedia

Some research with conventional texts revealed that confidence ratings correlate with performance only poorly with a maximum gamma correlation of .20, leading to the term "illusion of knowing" (Glenberg & Epstein, 1985; Morris, 1995). However, other research using different texts revealed that gammas up to .69 can be reached (Weaver & Bryant, 1995, experiments 2 & 3), and further studies usually found gammas differing from zero, indicating that monitoring of comprehension is possible (see Maki & McGuire, 2002, for a review). As argued above, a certain amount of accuracy of monitoring is necessary in order to regulate one's learning effectively (see Dunlosky, Hertzog et al., 2005). Moreover, we argue that for hypermedia the accuracy of monitoring is of even more importance than for conventional texts since hypermedia learning poses higher demands to self-regulation due to the higher degree of learning control (e.g. Bannert, 2007; Dillon & Gabbard, 1998; Tergan, 2002; Unz & Hesse, 1999): hypermedia learning requires more searching for information and decision-making concerning the selection of pages than learning with linear texts because the author of the texts has already made these decisions. Therefore, in hypermedia learning accurate monitoring is even more necessary to meet these demands. Besides, in a hypermedia environment all learners do not necessarily process the learning material in the same sequence, and furthermore, the same material is not necessarily processed by all learners because they are completely free in their navigation. Consequently, terms from the items of a test may not be useful as cues for some learners if they have not read these terms whilst studying the hypermedia. Thus, when judging their performance, learners may rely more on theory-based cues such as their overall knowledge of the topic than on experience-based cues such as the terms from the specific question. Therefore, we want to investigate whether the findings that learners are able to monitor their learning indicated by accuracy above chance can be replicated with hypermedia, at least to some degree.

One further empirical finding with conventional texts concerns the accuracy of predictive versus retrospective judgements. Studies using the within-person correlation gamma as

the measure of accuracy found that retrospective judgements were more accurate than predictive judgements (Glenberg & Epstein, 1987; Maki, Foley, Kajer, Thompson, & Willert, 1990). But if accuracy of monitoring is calculated using a within-person absolute measure of accuracy (see Nelson, 1984; 1996, for the difference between relative and absolute measures of accuracy), the results are no longer that clear. For example, Pressley, Snyder, Levin, Murray, and Ghatala (1987) found no difference in accuracy between retrospective judgements made after testing and predictive judgements made before the test, with the exception of one experimental condition in experiment three. In contrast, the results of another study (Gillström & Rönnberg, 1995) using an absolute measure of accuracy showed that, descriptively speaking, more students were classified as accurate on the basis of retrospective judgements than on the basis of predictive judgements. The calculation of a between-person measure of accuracy showed mixed results, too. Commander and Stanwyck (1997) found that about 58 % of the students were classified as accurate on the basis of predictive judgements, whereas 63 % were classified as accurate on the basis of retrospective judgements. In a classroom study (Hacker, Bol, Horgan, & Rakow, 2000), global predictive and global retrospective judgements over three exams were compared. A regression analysis with the judgements as the predictors and the performance as the criterion showed that, descriptively speaking, predictive judgements were more accurate than retrospective judgements, at least for the first two exams.

To sum up, there is clear evidence of retrospective judgements being more accurate than predictive judgements if the within-person correlation gamma is used as a measure of accuracy, and there is some evidence from studies using within-person measures of absolute accuracy or between-person measures of accuracy. The latter results are somewhat limited because inferential statistics were not often used in the studies, but results were reported only descriptively. One explanation for retrospective judgements being more accurate than predictive ones is that experience-based cues like terms in the questions of a test are available for retrospective judgements only. As we argued above, more experience-based cues are given in the items of a comprehension test than in the items of a transfer test. Therefore, we expect that in a comprehension test, retrospective judgements are more accurate than predictive judgements, but not in a transfer test.

The studies cited so far investigated differences between predictive and retrospective judgements. There are also some studies that investigate differences between global and spe-

cific judgements. Interestingly, specific judgements were not always better than global judgements in the case of predictive judgements (Dunlosky, Rawson, & McDonald, 2002, experiment 2). In this study, participants read six text paragraphs with each paragraph including four key terms. After reading, the participants made one global judgement and four specific judgements for each key term. Then the participants had to define each of the key terms, and the accuracy of their judgements was calculated using the within-person correlation gamma. No difference was found between global judgements and specific judgements. In contrast, specific judgements were more accurate than global judgements when the key terms had to be defined before the specific judgement was made (Dunlosky, Rawson et al., 2005). These findings were explained with the *accessibility hypothesis*, which was proposed by Koriat (1995) in connection with research on feeling-of-knowing-judgements. With reference to this hypothesis, Dunlosky et al. (2005) assumed that the accuracy of judgements is a function of the total amount of information activated immediately prior to making such a judgement. In contrast to specific judgements, global judgements were made quite quickly and there was little time to activate enough information prior to the judgement in order to obtain an accurate estimation of one's own knowledge. When specific judgements were recorded immediately after the key term had to be defined, more information had been accessed before the judgement. This lead to a more accurate estimation compared with global judgements.

In accordance with the results of Dunlosky et al. (2005), it can be expected that a predictive global judgement before a comprehension test is less accurate than specific retrospective judgements taken after each item of the test because participants use the terms of the items as cues. These experience-based cues activate some of the knowledge that has been learnt and this knowledge will enhance the level of judgement and the probability of solving the item correctly. As we argued above, there are almost no cues corresponding to the learning material in the case of a transfer test and, hence, the mechanism will not work in the same manner. Of course, knowledge is activated whilst taking the transfer test, too, but this activation is less specific than in the case of the comprehension test and may also be misleading. Thus, in the case of transfer tests, retrospective specific judgements can not be expected to be more accurate than the predictive global judgement.

Research Questions and Hypotheses

All cited studies from metacomprehension research used chapters or paragraphs from textbooks as learning material. In contrast, in the present study, participants learned with a

hypermedia learning environment in which the students had the opportunity to browse freely among all text and picture material available. That is why it can not be assumed that all learners processed the learning material in the same sequence, and furthermore, that the same material had been processed by all learners. Thus, terms from the items of a test may perhaps be of less use when generating judgements compared with conventional linear texts, leading to the question if (H1) the learners estimate their knowledge above chance.

The major aim of this study is to compare the accuracy of monitoring for two kinds of knowledge: comprehension and transfer to new situations. The tests measuring these kinds of knowledge were based on the same learning material, but were different concerning the kind of task and the answer format (multiple choice vs. open answer format). As the comprehension test was tied more closely to the learning material than the transfer test, the items in the comprehension test included many terms from the learning material, whereas the transfer test did not. We assume that these terms can be used as experience-based cues for generating retrospective judgements. Hence, many more experience-based cues were available to the learners in the comprehension test than in the transfer test. Thus, we (H2) expect a higher retrospective accuracy for the comprehension test than for the transfer test, and this difference should (H3) be due to more experience-based cues compared to the transfer test. Further, the terms are available as cues for specific retrospective judgements, but not for global predictive judgements, and therefore, global predictive judgements should be less accurate than specific retrospective ones. Since this argument remains true for the comprehension test but not for the transfer test in which almost no cues are given in the items, we (H4) expect almost no difference in the accuracy of predictive global judgements and retrospective specific judgements for the transfer test, but there should be a considerable difference between the two kinds of judgements for the comprehension test.

Method

Participants

Participants were 126 University students, 98 (77.8%) female and 28 (22.2%) male. The average age was 23.0 years with a standard deviation of 4.35. The youngest student was 18 years old, the eldest 42. The average number of terms studied was 2.8 with a standard deviation of 2.61. The students' study disciplines were Education or Psychology. The participants received

either 20 Euros for participating in the study or, in the case of psychology students, a certification needed for their studies.

Material, Instruments, and Procedure

The participants learned the basic concepts of operant conditioning presented in a hypermedia system within 30 minutes in individual sessions. The hypermedia learning environment consisted of 44 nodes with about 12.500 words, 19 pictures/diagrams, and 240 links in total. Participants were requested to learn about the principles of operant conditioning. The relevant learning material involved 9 nodes, including 2300 words, 3 pictures and 60 links. Thus, participants had to search and read these nodes, and potentially, learners could also miss out some of the relevant nodes. Navigation was made possible by using a hierarchical navigation menu, forward- and backward-buttons on each node, and hotwords directly placed in the text.

After learning, the students had to fill in a comprehension test about operant conditioning which contained 22 items with 5 answer options, including the option "don't know the answer". The questions covered most of the relevant pages in the hypermedia system and contained items inquiring the comprehension of definitions mentioned in the learning material, but also items requiring comprehension of the concepts of operant conditioning. Right before taking the test, participants made a global prediction by answering the question "What percentage of the following items will you solve correctly?". Additionally, after each test item, they made a judgement about their confidence that they had solved the item correctly on a six-point rating scale ranging from 1 ("very low confidence") to 6 ("very high confidence"). This kind of scale has been successfully used for obtaining confidence judgements in various studies by Glenberg and colleagues (Glenberg & Epstein, 1985, 1987; Glenberg et al., 1987) before.

As a second performance measure, a transfer test was conducted using the same judgement procedure. In contrast to the comprehension test, the transfer test had an open answer format: the students had to write down short solutions to educational problems which were not included in the learning material. Thus, in contrast to the comprehension test, not only was the comprehension of definitions and theory of operant conditioning required, but transfer and application of that knowledge and the formulation of an answer in a few sen-

tences. Answers were coded by two raters independently, and afterwards, both raters discussed and decided about cases in which no consensus was achieved. Examples for both tests can be found in the appendix.

Results

Calculation of Accuracy Measures

In the comprehension test, there was the option to answer "don't know the answer". When this option was chosen, students were not asked to make a retrospective confidence judgement about this item. Therefore, the data set contains a number of quasi "missing data". The maximum of missing values was observed for one item with 62 participants choosing "don't know the answer". Similarly, but less often, this was the case in the transfer test when participants gave no answer to a question. The maximum of 25 missing values was observed for one item in this test. There are two ways to treat the problem: (1) calculating the accuracy of judgements only for those participants who filled in all judgements, (2) setting the judgement to the value of zero because the alternative "don't know the answer" indicates that the person is very unsure about the right answer. The first way of treating the missing values leads to an underestimation of accuracy because of the restricted range of the judgements and test scores (see also Schwartz & Metcalfe, 1996): there are many more missing values for participants with low test scores than for participants who achieved high test scores. Those low achievers would be excluded from analyses, causing a restricted range and therefore an underestimation of accuracy. In addition, the smaller sample size reduces the power of inference tests. The second alternative may lead to an overestimation of accuracy because the option "don't know the answer" determines both the test score and the judgement, and, therefore, pushes the correlation between these variables. In the calculation of the following results, the second procedure was used because too much data would have been lost using the first alternative. So, one has to keep in mind that the presented results have to be interpreted as an upper limit. To prevent reporting statistical artefacts, the first alternative was calculated, too. Thus, these results can be interpreted as a lower limit. Whenever a result was not replicated with this reduced sample (N =28), we will mention it in the results.

When calculating the accuracy of monitoring, measures of within-person accuracy and between-person accuracy have to be distinguished (see e.g. Dunlosky & Hertzog, 2000, for a

discussion of measures). Between-person measures are calculated by correlating a scale built from the judgements with a scale built from the test items. Therefore, accuracy is a measure for the extent to which individual differences in the judgements are correlated with individual differences in performance. In contrast, within-person accuracy is calculated for each single participant: in the case of *relative accuracy*, the rank order of judgement is correlated with the rank order of the performance in the items. In the case of *absolute accuracy*, the judgement about an item is compared with the performance in that item, e.g. using the absolute value of their difference (e.g. in the studies by Nietfeld, Cao, & Osborne, 2005; G. Schraw, 1994) or calculating bias as the signed difference between judgements and performance (Yates, 1990). The most commonly used measure of relative accuracy is Goodman and Kruskal's gamma because of its statistical properties, above all its independence from the absolute magnitude of judgements and performance (Nelson, 1984, 1996).

Since only one item was used for predictive judgements in this study, it was not possible to calculate gamma as a measure of relative accuracy for predictions because a minimum of two pairs of judgement and corresponding test item would have been necessary. For retrospective items, absolute accuracy can not be calculated because the answers were not obtained using a percentage scale. Instead, between-person accuracy is calculated using Pearson's r as it is done in self-assessments of performance in intelligence tests, too (e.g. Furnham & Rawles, 1999; Holling & Preckel, 2005). That is, firstly, scales were built from the items and secondly, Pearson's r was calculated between all participants. This proceeding implies that judgement and test performance are two latent constructs that are measured by some manifest items. In addition, within-subject relative accuracy was calculated for retrospective item-specific judgements.

Analysis of Scales

For each test, one scale for retrospective judgements and one scale for test performance were calculated taking the mean value, respectively the sum of all items per person. Statistics for both tests and for the retrospective judgements are listed in table 1. For predictive judgements, it is not possible to build a scale because these judgements consist of one item only. The statistics for predictive judgements are also given in the table.

Table 1. ***Statistics for Scales***

Scale	Items	M	SD	Alpha
Comprehension Test				
Performance [a]	22	15.06	3.82	.75
Predictive judgement [b]	1	54.96	20.87	----
Retrospective judgements [c]	22	3.95	0.95	.88
Transfer Test				
Performance [d]	8	20.08	5.37	.74
Predictive judgement [b]	1	49.13	19.32	-----
Retrospective judgements [c]	8	3.57	1.16	.84

Notes.
[a] theoretical minimum 0, maximum 22
[b] predicted percentage of correct items
[c] 0 = "don't know", 1 = "very unsure", 6 = "very sure"
[d] theoretical minimum 0, maximum 40

The comprehension test has a sufficient Cronbach's alpha of .75. A Cronbach's alpha of .74 indicates sufficient reliability of the transfer test, too. The intercoder reliability was calculated as kappa = .86. For the retrospective confidence judgments, a Cronbach's alpha of .88 for the comprehension test and .84 for the transfer test shows a good reliability.

Above-Chance Accuracy of Judgements

Between-person accuracy. The first hypothesis (H1) states that participants judge their performance above chance, that is, correlations between judgement and performance have to be greater than zero. Correlations for each kind of judgement with the corresponding test score were calculated on the basis of the scales described above (see table 2). Squared correlations as a measure for effect size are .27 up to .61. Accordingly, participants were able to predict their test performance above chance, and they were able to judge their performance retrospectively, too. Using the reduced sample ($N = 28$), the correlation between the transfer test and the retrospective judgement was no longer significant.

Within-person relative accuracy. The first research question may be answered using the within-person correlation gamma as measure of relative accuracy for the retrospective judgements, too. Because it is not possible to calculate gamma when all judgements are equal or all items are solved, respectively all items are not solved, gamma was not calculated for four persons in the case of the comprehension test and one person in the case of the transfer test. For comprehension, the mean of gammas ($M = .76$, $SD = .25$) differed from zero signifi-

cantly, $t(121) = 34.2$, $p < .001$, $d = 3.10$, as did the mean of gammas ($M = .44$, $SD = .48$) for transfer, $t(124) = 10.2$, $p < .001$, $d = .91$.

Table 2. Pearson's Correlations between Judgement and Test Performance

	Test Performance	
	Comprehension	Transfer
Predictive Judgement	.60	.52
Retrospective Judgements	.78	.49

Notes.
All correlations were significant with $p < .001$.

Differences of Accuracy Between the Comprehension and Transfer Test

Between-person accuracy. To investigate whether accuracy varies with the type of test (H2), the differences between accuracy of comprehension and accuracy of transfer were tested using a test by Steiger (1980, c.f. Diehl & Arbinger, 1992, p. 385). Accuracy of predictive judgement for comprehension ($r = .60$) was not significantly higher than accuracy of predictive judgement for transfer ($r = .52$), $z(N = 126) = 1.22$, n.s. But accuracy of retrospective judgements for comprehension ($r = .78$) was significantly higher than accuracy of retrospective judgements for transfer ($r = .49$), $z(N = 126) = 4.53$, $p < .001$. Using the reduced sample ($N = 28$), the difference between the accuracies for retrospective judgements was no longer significant, but the amount of the difference was even higher, indicating that the results are quite the same with the reduced sample.

Within-person relative accuracy. Using gamma as a measure, retrospective judgements for comprehension ($M = .76$, $SD = .25$) were significantly more accurate than for transfer ($M = .44$, $SD = .48$), $t(121) = 6.81$, $p < .001$, $\eta^2 = .28$. Thus, the result concerning retrospective judgements can be replicated using a measure of within-person relative accuracy.

Cues Used for the Generation of Judgements

As stated in the third hypothesis (H3), we expected that the differences between the two tests were due to the use of different cues. To split the between-person variance of the judgements into variance explained by the domain knowledge (a theory-based cue) and the variance explained by the cues from the questions (an experience-based cue), structural equation modelling was used. All models were calculated with LISREL, version 8.50.

Three subscales were built for each of the retrospective judgements, the comprehension test, and the transfer test. For each latent construct, the first variable was put into the first

subscale, the second into the second subscale, the third into the third subscale, the fourth into the first subscale again etc. Thus, all three scales measured the same construct with different items, and further, the first subscale of the retrospective judgements corresponded to the first subscale of the knowledge, respectively the transfer test, the second subscale of the judgements corresponded to the second subscale of the tests, and the same was true for the third subscale, too. This construction of subscales ensures that variance between the subscales can be explained in two ways: (1) Common variance due to the latent constructs, that is, the generation of retrospective judgements is explained by the comprehension of operant conditioning as a latent construct. Thus, a correlation between the latent constructs would support a theory-based generation of judgements. (2) Common variance between corresponding subscales that can not be explained by latent constructs can be attributed to variance specific to the items included in the two subscales. That would support the experience-based generation of judgements independently from domain knowledge represented in the latent construct.

An inspection of the histograms showed approximately normally distributed data for the subscales. Therefore, all models were estimated using the maximum likelihood (ML) function. Firstly, a model was estimated for the comprehension test (see figure 1). Latent variables comprehension test *(ct)* and retrospective judgement of comprehension *(rj)* were each measured by three manifest subscales built as described above. Because there was only one manifest variable for predictive judgement of comprehension *(pj)*, factor loading was set to one and residual variance to zero. Consequently, predictive judgement of comprehension *(pj)* was not measured at latent but at manifest level. Covariances between the corresponding subscales for comprehension and retrospective judgements of comprehension at manifest level were set freely and, therefore, were estimated in the model. The model showed a good fit with the data with $\chi^2(10, N = 126) = 16.72$, $p = .081$, $RMSEA = .073$, $p = .236$ ($CI = .000, .133$), $CFI = .988$, and $NNFI = .975$ (for the evaluation of model fit indices, see Schermelleh-Engel, Moosbrugger, & Müller, 2003). Factor loadings were rather high except for *ct2* with a factor loading of .58 caused by keeping items in the test with non-sufficient discrimination coefficients. The covariances of the subscales of comprehension and subscales of retrospective judgements reflected common variance due to the same specific items in these subscales. The correlations of .13 and .21 were much lower than at latent level, but still significant. Thus, most of the covariance between retrospective judgements and comprehension was explained via the latent path, that is, by a judgement of comprehension irrespective of the single items.

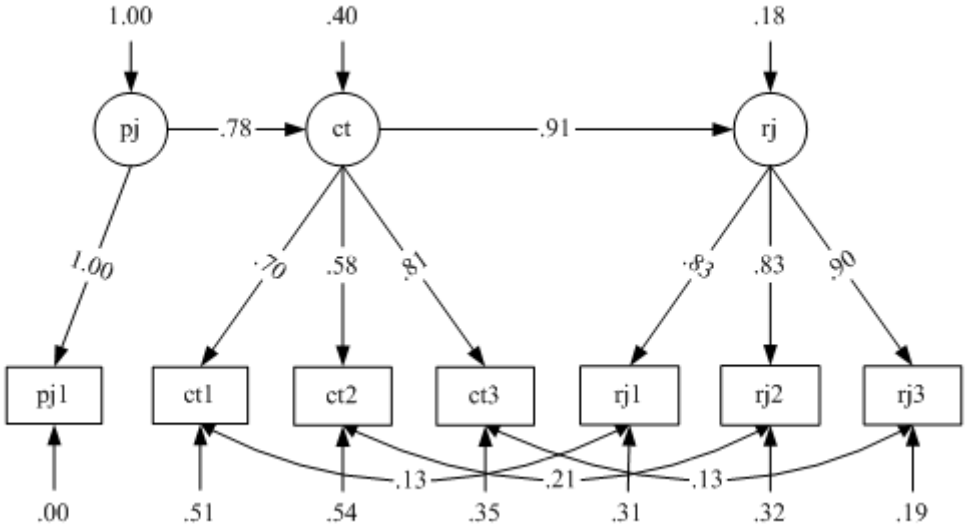

Figure 1. **Completely standardised solution for predictive judgement (pj), comprehension test (ct), and retrospective judgements of knowledge (rj); all coefficients are significant with at least p < .05**

Secondly, a model was estimated for transfer test, too (see figure 2). As in the first model, transfer test *(tt)* was measured by three subscales, and retrospective judgement of transfer *(rj)* was measured by three subscales as well. For predictive judgement *(pj)*, factor loading was set to one and residual variance to zero because there was only one manifest item in the data for predictive judgement. The model showed an insufficient fit to the data with $\chi^2(10, N = 126) = 37.30$, $p < .001$, *RMSEA* = .148, $p = .236$ (*CI* = .099, .200), *CFI* = .906, and *NNFI* = .804. Factor loadings were somewhat smaller than for the comprehension test, and residual variances of the manifest variables were higher. The correlation of .53 between transfer *(tt)* and retrospective judgement *(rj)* was much lower compared with the corresponding correlation of .91 regarding comprehension. Correlations of .20, .24, and .26 indicated slightly more covariance between the manifest subscales of transfer and the retrospective judgement subscales than for the knowledge test. Thus, the results were not in accordance with our hypothesis (H3), which claims that there is no less use of experience-based cues for the transfer test than for the comprehension test. This even remains true if the different residual variances of the manifest variables in the two models were taken into account, that is, if the correlations were adjusted using the mean of the residual variances.

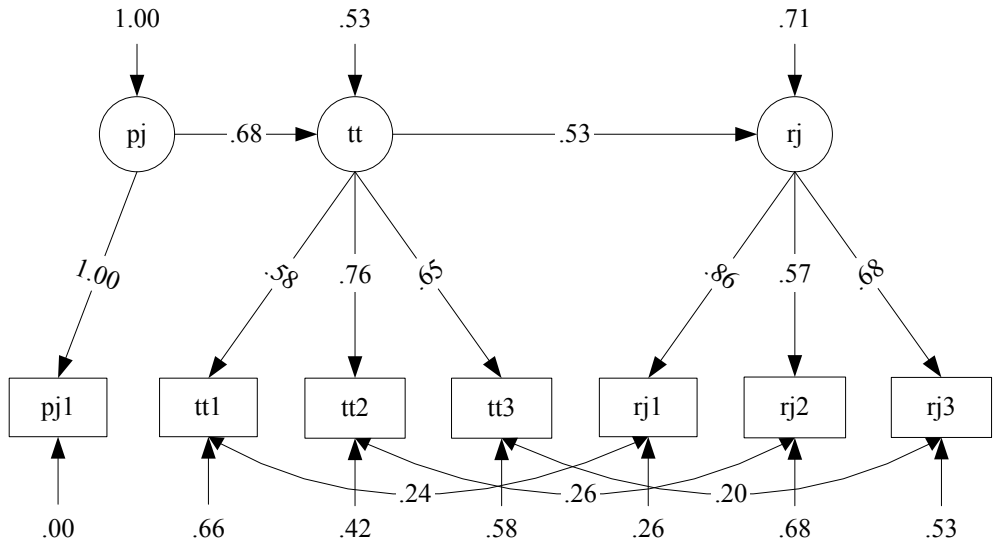

Figure 2. **Completely standardised solution for predictive judgement (pj), transfer test (tt), and retrospective judgements (rj); all coefficients are significant with at least p < .05**

The argument of less use of experience-based cues in the transfer test than in the comprehension test was crucial for our fourth hypothesis, too. As stated above, global predictive judgements should be as accurate as retrospective specific judgements for the transfer test, but not for the comprehension test (H4). To test whether the between-person correlations of judgements and performance were of significant difference, the test of Olkin and Siotani (1964, c.f. Bortz, 1993, p. 205) was calculated. This statistical test is used to calculate the significance of the difference of a correlation r_{ab} between the variables *a* and *b* and a second correlation r_{ac} between the variables *a* and *c* obtained from the same sample. The test statistic is calculated as the difference of the Fisher-z-transformed correlations adjusted for sample size and for covariance of the distributions of the two correlation coefficients. This results in a z-Value that can be tested for significance easily. Accordingly, accuracy of retrospective judgements ($r = .78$) was higher than that of predictive judgements ($r = .60$) for comprehension, $z(N = 126) = 3.60$, $p < 001$. In contrast, accuracy of retrospective judgements ($r = .49$) did not differ from accuracy of predictive judgement ($r = .52$) for transfer, $z(N = 126) = 0.42$, n.s. For the reduced sample ($N = 28$), the picture changed: the test was no longer significant in the case of the comprehension test, but was for the transfer test. Thus, the fourth hypothesis can be maintained for the whole sample only, that is, a difference in accuracy was found between global predictive judgements and specific retrospective judgements for comprehension, but not for transfer.

Post-hoc Analysis: Accessibility Hypothesis

According to the accessibility hypothesis (see e.g. Dunlosky, Rawson et al., 2005; Koriat, 1995), the accuracy of monitoring is a function of the amount of knowledge activated in the memory immediately before a judgement is made. Thus, not the correctness of the activated information, but the amount of this information is crucial for the judgement. Therefore, one reason for specific retrospective judgements being lower for the transfer test than for the comprehension test may be that participants who wrote down an incorrect answer judged their confidence equally to those who wrote down a partially correct answer because in both cases, a similar amount of information was assessed before the judgement was made. This would lead to a decrease of accuracy compared to the accuracy in the comprehension test in which no possibility was given to write down incorrect answers. To test this explanation, we computed the mean of retrospective judgements for incorrect, partly correct and correct answers for each participant. For omissions, no values were calculated because in this case, participants skipped the judgement according to the instructions (see method section). There were $N = 69$ participants whose answers fell into all three categories.

The multivariate analysis of contrasts showed that judgements differed as a function of the category of the answers, $F(2,67) = 26.49$, $p < .001$, $\eta^2 = .44$. This significance is due to differences between all of the three categories, that is, judgements for incorrect answers ($M = 3.45$, $SD = 1.10$) were lower than judgements for partly correct answers ($M = 3.99$, $SD = 1.06$), $F(1,68) = 19.88$, $p < .001$, $\eta^2 = .24$, and judgements for partly correct answers were lower than judgements for correct answers ($M = 4.49$, $SD = 1.11$), $F(1,68) = 16.98$, $p < .001$, $\eta^2 = .18$.

Summary and Discussion

First of all, the results showed that the students were able to judge their amount of knowledge quite adequately, as the significant correlations between the judgments and the tests of comprehension and transfer show. Thus, we maintain our *first hypothesis*. The highest between-person correlation obtained was .78 for retrospective judgements about comprehension, the lowest correlation was .49 for retrospective judgements of transfer. In a meta-analysis (Mabe & West, 1982), a non-weighted mean correlation of .42 with a standard deviation of .20 for self-assessment of academic performances was obtained. The smallest correlation in the studies contained in the meta-analysis was zero, the maximum correlation was .80.

Desoete and Roeyers (2006) calculated a correlation of .17 for primary children making global predictions about their performance in an arithmetic test. However, the low correlation found by Desoete and Roeyers may be due to the low age of their participants. Compared with these results, the correlations found in this study were rather high, but one has to consider that most of the studies included in the meta-analysis were conducted in the field, whereas our study was conducted in the laboratory, which allowed more control of variables and therefore enhanced the correlation. For the retrospective judgements, within-person-correlations gamma were calculated, indicating an above chance-accuracy, too. Compared with gammas found in studies on metacomprehension with traditional texts, the gammas in this study were rather high. The highest gammas the authors know from the literature were about .69 (Weaver & Bryant, 1995, experiments 2 & 3), and Dunlosky and Lipko (2007) reported a mean gamma of .56 from twelve studies under conditions that successfully enhanced accuracy, whereas we found .76 for the comprehension test. As the value of .65 for the reduced sample showed, this was not a pure artefact caused by the "don't know" answer option in the items. The value of .76 may be seen as an upper limit, whereas the value of .65 is a lower limit because of the restricted variance in the reduced sample. One explanation for the high gammas could be the fact that hypermedia requires more monitoring than traditional texts do, as there are higher demands of deciding what page should be selected next during the learning process, as mentioned above (e.g. Bannert, 2007; Lin, Hmelo, Kinzer, & Secules, 1999). Since there are no studies using gamma as a measure of monitoring during hypermedia learning, further research should explicitly compare metacomprehension with traditional texts as well as hypermedia by using the same learning material and tests.

As we expected in our *second hypothesis*, the accuracy for the specific retrospective judgements was higher for comprehension than for transfer. This result was found for between-person correlation as measure of accuracy and for the within-person-correlation gamma, too. Our assumption was that there were more cues in the comprehension test than in the transfer test and that these experience-based cues would be used to generate judgements (see *third hypothesis*). Unfortunately, our structural equation models did not support this hypothesis, that is, there were approximately the same correlations between manifest variables in both tests if the residual variances were taken into account. Another explanation could be the answer format. That is, according to the accessibility hypothesis (e.g. Dunlosky, Rawson et al., 2005; Koriat, 1995) with open answer formats, people tend to judge their answer to be correct irrespective of the actual correctness of their response. If this was true for the present

data, a decrease of the accuracy of monitoring would be the result. But in contrast to the accessibility hypothesis and in accordance with the results by Dunlosky et al. (2005), results showed significant differences in the judgements made after an incorrect answer, a partly correct answer and a correct answer. Thus, the participants were able to judge whether they wrote something incorrect or partly correct and therefore, the accessibility hypothesis does not explain our results.

The *fourth hypothesis* stated that there should be no difference between global predictive and retrospective specific accuracy for the transfer test, but there should be a difference for the comprehension test. Indeed, results using a between-person measure of accuracy revealed that the global predictive accuracy for comprehension was significantly lower than the specific retrospective accuracy. For the transfer test, the accuracies did not differ from each other significantly. The idea leading to the hypothesis was that for retrospective specific judgments, participants were able to use experience-based cues from the comprehension test, but not from the transfer test. Nevertheless, our structural equation modelling did not support this explanation. Another explanation based on the models may be that the latent domain knowledge does not serve well as a theory-based cue for the generation of retrospective judgements for the transfer test, but does for the comprehension test: the correlation between comprehension test *(ct)* and retrospective judgements *(rj)* at latent level was higher than the correlation between transfer test *(tt)* and retrospective judgements *(rj)*. The use of domain knowledge may be more appropriate for the comprehension test because this test is more closely tied to the learning material than the transfer test. The problem with this explanation is that for predictive judgements, there is no significant difference between the two tests as one would expect if the gap between learning material and test is causal for the different results for the specific retrospective judgements.

To explain the results completely, it may be necessary to investigate the generation of judgements more deeply, adding some more sources that potentially explain variance in the judgements. Both structural equation models showed that a considerable amount of variance was explained, but the model fit for transfer indicated that domain knowledge together with specific experiences whilst answering the individual items could not explain the generation of retrospective confidence judgements sufficiently. Here, we recommend the incorporation of an individual trait into the modelling in future research. There are some studies that investigate the generality of accuracy over different tasks (Pallier et al., 2002; Gregory Schraw,

1997; Gregory Schraw, Dunkle, Bendixen, & DeBacker Roedel, 1995; Gregory Schraw & Nietfeld, 1998; West & Stanovich, 1997), and at least one study that investigates stability over time (Jonsson & Allwood, 2003), suggesting that there is a general metacognitive ability across tasks from different domains if metacognitive ability is measured using bias as an indicator, but that there is no such ability if a relative accuracy score, e.g. the within-person correlation gamma, is used as a measure. Nevertheless, such a trait should be interpreted more as a motivational construct than a metacognitive ability, as bias is an indicator for overconfidence and underconfidence.

Furthermore, studies with larger sample sizes are needed, e.g. classroom studies, to investigate every single item in the structural equation modelling instead of computing three subscales. By means of an explorative analysis using Mplus software, we incorporated dichotomous items and calculated paths as logistic regressions, and indeed, such analyses allowed interpretations at the level of single items when applied to the data of the present study. But there are statistical problems with such kinds of analyses due to the huge amount of variables requiring a much bigger sample than in the present study.

To examine the influence of conditions on the validity of judgements more precisely, further research with a full matrix of conditions is needed: global vs. specific, and predictive vs. retrospective. To avoid the statistical problems of the present study, the answer “don’t know” in the comprehension test should be dropped. Further, to allow the computation of gammas, more than one item should be used for predictive judgements. Despite these problems, the presented results of the study showed that using structural equation modelling may be one way to enrich metacomprehension research by analysing the factors that influence the generation and the validity of judgements about knowledge more deeply.

References

Bannert, M. (2007). *Metakognition beim Lernen mit Hypermedia [Metacognition in learning with hypermedia]* (Vol. 61). Münster: Waxmann.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler [Statistics for social scientists]*. Berlin: Springer.

Brown, A. L. (1980). Metacognitive development and reading. In R. J. Spiro, B. C. Bruce & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 453-481). Hillsdale, N.J.: Erlbaum.

Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology, 22*(1), 39-52.

Desoete, A. (2007). Evaluating and improving the mathematics teaching-learning process through metacognition. *Electronic Journal of Research in Educational Psychology, 5*(3), 705-730.

Desoete, A., & Roeyers, H. (2006). Metacognitive macroevaluations in mathematical problem solving. *Learning and Instruction, 16*(1), 12-25.

Diehl, J. M., & Arbinger, R. (1992). *Einführung in die Inferenzstatistik [Introduction to statistical inference]* (2 ed.). Eschborn bei Frankfurt am Main: Klotz.

Dillon, A., & Gabbard, R. (1998). Hypermedia as an educational technology: A review of the quantitative research literature on learner comprehension, control, and style. *Review of Educational Research, 68*, 322-349.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: a componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15*(3), 462-474.

Dunlosky, J., Hertzog, C., Kennedy, M. R. T., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology, 9*(1), 4-11.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228-232.

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, UK: 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 and Language, 52*(4), 551-565.

Furnham, A., & Rawles, R. (1999). Correlation between self-estimated and psychometrically measured IQ. *Journal of Social Psychology, 139*, 405-410.

Gillström, A., & Rönnberg, J. (1995). Comprehension calibration and recall prediction accuracy of texts: Reading skill, reading strategies, and effort. *Journal of Educational Psychology, 87*(4), 545-558.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702-718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory and Cognition, 15*(1), 84-93.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*(2), 119-136.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160-170.

Hasselhorn, M. (2001). Metakognition [Metacognition]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (2 ed., pp. 466-471). Weinheim: Psychologie Verlags Union.

Holling, H., & Preckel, F. (2005). Self-estimates of intelligence - methodological approaches and gender differences. *Personality and Individual Differences, 38*(3), 503-517.

Jacobs, J. E., & Paris, S. G. (1987). Childrens metacognition about reading: issues in definition, measurement, and instruction. *Educational Psychologist, 22*(3&4), 255-278.

Jonsson, A.-C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*, 559-574.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*(3), 311-333.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289-325). Cambridge, NY: 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 memory and metamemory* (Vol. 1, pp. 117-135). New York, NY: Psychology Press.

Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345-391.

Lin, X., Hmelo, C., Kinzer, C. K., & Secules, T. (1999). Designing technology to support reflection. *Educational Technology Research & Development, 47*, 43-62.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: a review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280-296.

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 609-616.

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 68-92). Cambridge, UK: University Press.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*(3), 349-363.

Morris, C. C. (1995). Poor discourse comprehension monitoring is no methodological artifact. *Psychological Record, 45*(4), 655-668.

Nelson, T. O. (1984). A comparision of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109-133.

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*(3), 257-260.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (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*(1), 7-28.

Pallier, G., Wilkinson, R., Danthir, V., Kleitman, S., Knezevic, G., Stankov, L., et al. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology, 129*(3), 257-299.

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly, 22*(2), 219-236.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435-451.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Edaucational Psychology, 19*, 143-154.

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *Journal of Experimental Education, 65*(2), 135-146.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & DeBacker Roedel, T. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*(3), 433-444.

Schraw, G., & Nietfeld, J. L. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology, 90*(2), 236-248.

Schwartz, B. L., & Metcalfe, J. (1996). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: knowing about knowing* (2 ed., pp. 93-113). Cambridge, MA: MIT Press.

Simons, P. R. J. (1986). Metacognition. In E. De Corte & F. E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 436-444). Oxford, UK: Elsevier Science.

Tergan, S.-O. (2002). Hypertext und Hypermedia: Konzeption, Lernmöglichkeiten, Lernprobleme und Perspektiven [Hypertext and hypermedia: conceptions, learning, problems in learning, and perspectives]. In L. J. Issing & P. Klimsa (Eds.), *Information und Lernen mit Multimedia* (pp. 99-112). Weinheim: Beltz.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73.

Torrano Montalvo, F., & Gonzáles Torres, M. C. (2004). Self-regulated learning: current and future directions. *Electronic Journal of Research in Educational Psychology, 2*(1), 1-34.

Unz, D. C., & Hesse, F. W. (1999). The use of hypertext for learning. *Journal of Educational Computing Research, 20*, 279-295.

Veenman, M. V. J. (2005). The assessment of metacognitive skills: what can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und Metakognition. Implikationen für Forschung und Praxis* (pp. 77-99). Münster: Waxmann.

Weaver, C. A., III. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 214-222.

Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: the role of text difficulty in metamemory of narrative and expository texts. *Memory and Cognition, 23*(1), 12-22.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin and Review, 4*(3), 387-392.

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

Appendix

Item Example of Comprehension Test

Item No. 3:

Reinforcement is

a) only the presentation of a positive consequence, but not the disappearance of an aversive stimulus.
b) only the disappearance of an aversive stimulus, but not the presentation of a positive consequence.
c) both the presentation of a positive consequence and the disappearance of an aversive stimulus.
d) both the presentation of a positive consequence and the presentation of an aversive stimulus.
e) don't know

In case you have not answered "don't know": How sure are you that your answer is correct?

not at all sure ☐ ☐ ☐ ☐ ☐ ☐ very sure

Item Example of Transfer Test

Item No. 3:

Educational Situation

Willi doesn't want to do his homework. His parents feel helpless at first, but after a while they decide they will reward him for completing his homework. They are looking for adequate possibilities. Which procedure would a behaviorist recommend to them based on the theory of operant conditioning?

_____

_____

_____

_____

_____

In case you have written down an answer: How sure are you that your answer is correct?

not at all sure ☐ ☐ ☐ ☐ ☐ ☐ very sure