Psychonomic Bulletin & Review
2007, 14 (1), 107-111

Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning

Alan D. Castel
University of California, Los Angeles, California

David P. McCabe
Colorado State University, Fort Collins, Colorado

and

Henry L. Roediger III
Washington University, St. Louis, Missouri

The relation between subjects' predicted and actual memory performance is a central issue in the domain of metacognition. In the present study, we examined the influence of item similarity and associative strength on judgments of learning (JOLs) in a cued recall task. We hypothesized that encoding fluency would cause a foresight bias, so that subjects would overestimate recall of identical pairs (*scale–scale*), as compared with strong associates (*weight–scale*) or unrelated pairs (*mask–scale*). In Experiment 1, JOLs for identical word pairs were higher than those for related and unrelated pairs, but later recall of identical pairs was lower than recall of related pairs. In Experiment 2, the effect of encoding fluency (inferred from self-paced study time) was examined, and a similar pattern of results was obtained, with subjects spending the least amount of time studying identical pairs. We conclude that overconfidence for identical pairs reflects an assessment of item similarity when JOLs are made, despite associative strength being a better predictor of later retrieval.

The accurate assessment of one's own memory performance is a crucial ability that has important applied and theoretical implications. In order to determine the relation between predicted and actual memory performance, experimental investigations of metacognitive judgments have compared judgments of learning (JOLs) with actual memory performance. Although previous research on JOLs has shown that these judgments often accurately predict future memory performance, important differences have been found between predicted and actual performance (e.g., Benjamin, Bjork, & Schwartz, 1998; Kelley & Jacoby, 1996; Koriat & Bjork, 2005). These illustrations of discrepancies between predicted and actual performance provide important clues regarding how one generates JOLs and, more generally, the insight people have about memory performance.

The seminal work on how people assess learning in a paired-associates task was conducted by Arbuckle and Cuddy (1969), who reported strong correlations between a pair's associative strength and subjects' predicted and later recall. Although associative strength is a strong predictor of later recall, Begg, Duft, Lalonde, Melnick, and Sanvito (1989) have demonstrated that memory predictions are typically based on ease of processing, so that semantically related pairs lead to higher JOL ratings, as do other situations that are conducive to ease of processing (see also Dunlosky & Matvey, 2001). Dunlosky and Nelson (1994) have shown that JOLs are best calibrated with actual performance when subjects can make *delayed* JOLs, which involve providing a JOL when asked later to retrieve an item that had previously been studied. This suggests that when subjects can incorporate both encoding and later retrieval dynamics to make metacognitive judgments, JOLs are more accurate, although Kimball and Metcalfe (2003) have recently shown that this may occur because JOLs are based on actual memory performance (i.e., retrieval), as opposed to metamemory. This brief review suggests that although JOLs are often accurate, under certain conditions, metacognitive performance does not reflect actual performance, probably because subjects are using easily accessible features or dimensions of the to-be-remembered stimulus as a basis for judgments, instead of relying on both the encoding and the retrieval conditions.

Recently, Koriat and Bjork (2005) demonstrated an illusion of competence in which subjects overestimated memory performance for certain types of paired associates, depending on the associative strength and directionality of association within the word pair. Specifically, Koriat and Bjork (2005) showed that recall was more sensitive to subtle differences in associative strength, relative to predicted

A. D. Castel, castel@psych.ucla.edu

recall, and that subjects' predictions were not sensitive to the associative direction in a cued recall task. Thus, subjects would give equally high JOLs for strong associates (e.g., *lamp–light*) and for highly related but weaker associates (e.g., *beautiful–nice*), despite recall being much better for the strong associates. This illusion of competence illustrates an intriguing situation in which subjects' JOLs are relatively insensitive to the associative strength of a word pair. This may occur because these predictions are made on the basis of the semantic relatedness of the pair or its ease of encoding (e.g., Begg et al., 1989), with little regard to retrieval conditions in which the second word must be recalled when the first word is presented. Thus, it is important to better understand what cues or properties of the to-be-remembered information are incorporated in the metacognitive process.

Although previous work has shown that subjects overestimated memory performance for weakly related semantic associates, it remains unclear how semantic relatedness influences metacognitive judgments. In many cases, semantic relatedness is given greater weight as a cue for JOLs than are extrinsic factors, such as the circumstances of learning, encoding strategies, or anticipated conditions at retrieval (e.g., Koriat, 1997). Consistent with previous research, we suggest that subjects make general assessments of semantic relatedness when making JOLs, as opposed to making assessments that take into account both encoding factors and retrieval conditions. Thus, high levels of semantic relatedness would likely lead to higher JOLs, regardless of how well these factors facilitate later recall of the items. To test this idea, we used a design similar to that in Koriat and Bjork (2005), in which subjects were presented with semantically related word pairs with high forward associative strength (e.g., *loaf–bread*), semantically related word pairs with low forward associative strength (e.g., *note–card*), and unrelated pairs (e.g., *scalp–lunch*). These conditions replicated those in Koriat and Bjork (2005), although we further ensured that certain materials or pairings did not influence the results by holding the target word constant for the various conditions across subjects, thus varying only the cues used across conditions. In addition, however, we included identical pairs, which consisted of two identical words paired together (e.g., *water–water*). Subjects might predict that this type of pair would be fairly well remembered due to seemingly high perceived semantic relatedness (consistent with the foresight bias described by Koriat and Bjork, 2005). However, Tulving (1974) demonstrated that identical pairs actually resulted in poorer subsequent cued recall performance in the recognition failure of recallable words paradigm, relative to other types of pairs. We predicted that if subjects use item similarity (which can be defined as both the perceptual and the semantic match between the two words in a word pair) as a cue for making JOLs, these identical pairs should receive very high JOLs, despite actual recall being relatively lower. Identical pairs may elicit high JOLs because these pairs contain highly accessible information that enhances ease of processing. If subjects give higher JOL ratings but exhibit relatively poorer recall for identical pairs, this outcome would suggest that subjects rely on item similarity at encoding when making JOLs. Thus, including identical pairs in the design provides a particularly strong test that perceived similarity drives JOLs.

EXPERIMENT 1

In Experiment 1, we followed the general design of Koriat and Bjork (2005) but also included identical pairs in order to examine how JOLs and recall would be influenced by both semantic relatedness and item similarity. Specifically, we were interested in whether the identical words would elicit inflated JOLs, relative to actual recall performance and relative to strongly associated pairs. If the subjects displayed overconfidence with identical pairs, this outcome might reflect a strong reliance on encoding fluency (i.e., the ease of associating the two words in a pair) when making JOLs. Thus, although JOL ratings may minimize subtle fluctuations in semantic relatedness, they may be overly sensitive to item similarity. Although item similarity can, in theory, vary on a continuum, we used identical pairs in order to test the extreme boundary of item similarity, to determine how JOLs and recall would be influenced by this factor.

Method

Subjects. The subjects were 24 undergraduate students from Washington University, who participated for course credit.

Procedure. The subjects were told that they would study 48 word pairs and that they should try to remember the pairs for a later cued recall test. They were told that the pairs could be related, unrelated, or identical and were given examples of each pair type. They were instructed that following the presentation of each pair, they would be asked to make a JOL regarding how well they would remember the second word when presented with the first word on a later memory test.

During the study phase, word pairs were presented in the center of the computer screen for 4 sec. Following a 500-msec delay, the subjects were presented with a question mark indicating that they should provide a JOL rating. The subjects were told that they should use a scale of 0 to 100, with 0 meaning that *they would definitely not remember* and 100 meaning that *they would definitely remember* the second word on a later cued recall test. The subjects were instructed to use the entire range from 0 to 100, and they gave their responses orally. The experimenter recorded each response. The order of presentation of the pairs was randomized for each subject.

Following the study phase, the subjects engaged in a 3-min distractor task that involved rating the pleasantness of two digit numbers. They were then told that they would be presented with the first word from each studied pair and that they should attempt to recall the second word. They were told to make their response aloud for the experimenter to record. The subjects were also told that they could guess if necessary but that they should try to be as accurate as possible. For each trial, the subjects had up to 10 sec to make a response, and the order of presentation of the words was randomized.

Materials and Apparatus. At study, each subject was exposed to the same 48 targets, but the cue associated with each target differed across four counterbalancing conditions. Thus, across subjects, each target was presented an equal number of times with a strongly related high-associate cue (e.g., *clever–smart*), a weakly related low-associate cue (e.g., *learn–smart*), an unrelated cue (e.g., *vine–smart*), or an identical cue (e.g., *smart–smart*). All items were medium- to high-frequency words, and the stimulus pairings were chosen on the basis of the likelihood that the target would be given as a response to the cue words according to the Nelson, McEvoy, and Schreiber (1999) free association norms. These values were between .41 and .75 (M = .55) for the high associates, between .01 and .04 (M = .02) for the low associates, and .00 for the unrelated and identical pairs. Of course, this value of .00 might actually be undefined for the identical pairs, because subjects in norming studies probably

believe that they are supposed to produce a word that is different from the cue word.

Results and Discussion

The mean predicted recall (JOL) and actual mean recall percentages for each word pair type are shown in Figure 1. The data show that although the subjects were fairly well calibrated for the various pair types overall, important differences between perceived and actual performance were present. In order to determine how JOLs and recall varied as a function of word pair type, these data were entered into a 2 (measure: JOL or recall) × 4 (word pair type) repeated measures ANOVA. There was a main effect of word pair type [$F(3,69) = 131.17, p < .0001$], but the main effect of measure did not reach conventional levels of significance [$F(1,23) = 2.64, p = .13$]. More important, there was a significant word pair × measure interaction [$F(3,69) = 13.65$, $p < .0001$]. As can be seen in Figure 1, this interaction was partly driven by higher JOLs provided for the identical pairs, relative to their poorer recall, and the fact that this pattern was reversed for the semantically related pairs. Paired sample *t* tests revealed that JOLs were significantly lower than recall for the weakly and strongly related pairs [both ts$(23) > 2.13, p < .05$], whereas JOLs and recall did not differ significantly for the unrelated pairs [$t(23) = 0.01$, $p > .99$]; however, JOLs were significantly higher than recall for the identical pairs [$t(23) = 2.56, p < .05$].

Errors in recall typically were ones of omission in which subjects failed to recall a target word when presented with the cue. This was especially the case for identical pairs: The subjects gave blank responses, as opposed to guessing with a related or an unrelated word. The subjects provided an incorrect nonidentical word on fewer than 8% of all the identical trials and provided a blank on 24% of all the identical trials. In general, this trend existed for the other three conditions, with blank responses outnumbering all other types of error responses.

These findings are consistent with the prediction that identical pairs would receive high JOL ratings, despite their recall performance being comparable to that of weakly related pairs. The general findings replicate those of Koriat and Bjork (2005) but extend the illusion of competency to identical pairs that have high perceptual and semantic similarity. The results demonstrate an important interaction between JOLs and recall for the strong associates and identical pairs. JOLs are strongly affected by the exact overlap between two identical items, despite the fact that recall is lower for identical pairs, relative to strongly related pairs (Tulving, 1974). Unlike for highly associated pairs, the subjects probably did not use elaborative processing or imagery to link the two identical items. In order to further investigate what properties are used when JOLs are made for semantically related and identical pairs, we examined how encoding fluency would influence JOL ratings in Experiment 2.

EXPERIMENT 2

One reason that identical pairs received high JOLs may be due to the fluency or ease of processing of the pair. Encoding fluency has been identified in a number of previous studies as an important basis for JOLs (e.g., Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat & Ma'ayan, 2005). One way to measure encoding fluency is to let subjects pace the study time for each pair during the encoding session. Previous research on study time allocation (e.g., Nelson, 1993; Thiede & Dunlosky, 1999) has shown that people typically direct more study time to items that are perceived as more difficult to learn, and Koriat and Ma'ayan have suggested that study time can be used as an index of encoding fluency. In Experiment 2, we attempted to replicate the main findings in Experiment 1 while leaving study time for each pair under control of the subject. We hypothesized that the subjects would spend the least time studying the identical pairs, a reflection of perceived ease of processing or fluency, but that these pairs would again elicit higher JOLs relative to actual recall.

Method

Subjects. The subjects were 44 undergraduate students from Washington University in St. Louis, who participated for course credit and had not participated in Experiment 1.

Procedure, Materials, and Apparatus. The procedure, materials, and apparatus were identical to those in Experiment 1, with one modification. The subjects were told that they could study each pair for as long as they wanted and could advance to the JOL rating (and then the subsequent pair) by pressing the space bar on the keyboard. The rest of the procedure was identical to that in Experiment 1.

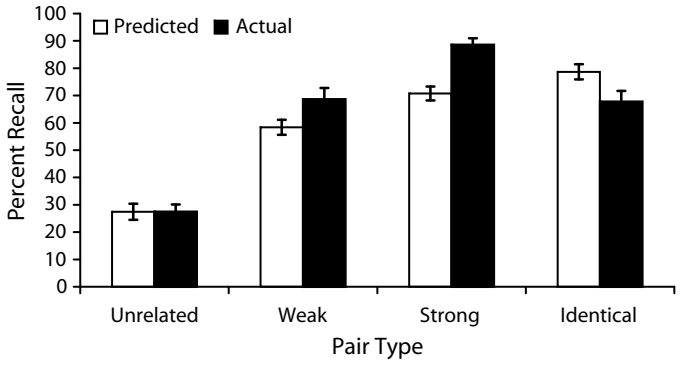

Figure 1. Mean predicted recall (judgments of learning) and actual recall as a function of the four types of word pairs in Experiment 1. Error bars represent standard errors of the mean.

Results and Discussion

The mean predicted recall (JOL) and actual mean recall percentages for each word pair type are shown in Figure 2, and the results are strikingly similar to those obtained in Experiment 1. As in Experiment 1, these data were entered into a 2 × 4 repeated measures ANOVA. There were main effects of word pair type [$F(3,129) = 214.45$, $p < .0001$] and measure [$F(1,43) = 13.63$, $p = .0001$] and a significant measure × word pair interaction [$F(3,129) = 16.95$, $p < .0001$]. In general, the findings replicated those in Experiment 1, despite the subjects being able to regulate their study time. JOL predictions were significantly higher than recall for the strong and weakly related pairs [$t(43) > 4.79$, $p < .0001$], whereas JOLs and recall did not differ significantly for the unrelated pairs [$t(43) = 1.22$, $p = .43$]. As in Experiment 1, JOLs were significantly higher than recall performance for the identical pairs [$t(43) = 2.04$, $p < .05$]. Also as in Experiment 1, errors were typically ones of omission, especially for the identical pairs.

The mean self-paced study times are presented in Table 1 and are also expressed in terms of normalized values (z-scores), to take into account differences in overall study time across individuals. For both measures, there was a main effect of word pair type [$F(3,129) > 16.34$, $p < .0001$], with follow-up pairwise comparisons indicating significant differences between all means ($p < .05$). Study time was shortest for the identical pairs and greatest for the unrelated pairs, in line with the idea that identical pairs are fluently processed.

The JOL and recall results from Experiment 2 replicate the findings from the first experiment, and the self-paced study time results reveal that the identical pairs received less study time than did the other types of word pairs. Koriat and Ma'ayan (2005) found that JOLs, as well as recall, decreased as a function of increasing self-paced study time, suggesting that items that require more study time are presumably less likely to be recalled. Although most of the findings from the present experiment are consistent with this line of reasoning, the data for the identical pairs represent an important exception.

In summary, and consistent with the results in Experiment 1, the subjects spent the least amount of time studying the identical pairs, but later recall was similar to that for the weakly related pairs. Although calibration was still relatively good for the identical pairs, in general, the subjects overestimated performance for these pairs. This outcome suggests that although identical pairs are thought of as quite easy to remember (as evidenced by high JOLs and low self-paced study time), subjects do not take into account that the lack of elaborative processing for these pairs might be detrimental to later recall.

Table 1
Mean Self-Paced Study Time (in Milliseconds and z-Score Units, With Standard Errors) for the Four Types of Word Pairs in Experiment 2

	Study Time			
	(in Milliseconds)		(in z-Score Units)	
Word Pair Type	*M*	*SE*	*M*	*SE*
Unrelated	6,880	547	0.29	0.04
Weakly related	6,427	501	0.08	0.04
Strongly related	5,932	419	−0.09	0.04
Identical	5,517	387	−0.26	0.05

GENERAL DISCUSSION

The goal of the present study was to better understand the cues that are used when metacognitive judgments are made. Previous findings had indicated that JOLs are relatively insensitive to differences in forward associative strength for weakly and strongly related word pairs (Koriat & Bjork, 2005). Following work by Tulving (1974), who showed that identical pairs are often poorly remembered, we showed that identical pairs are processed in a highly fluent manner and that subjects typically overestimate recall for identical pairs. Koriat's (1997) cue utilization framework suggests that intrinsic cues influence JOLs, and in the present study, the intrinsic features of identical pairs likely led the subjects to assign high JOL ratings, even though the subjects had probably done little elaborative processing to remember the pair. Dunlosky and Matvey (2001) also identified item relatedness as a critical factor for JOLs, and in the present study, an extreme example of item similarity was used to examine how JOLs are made in the presence of strong cue–target overlap. It should be

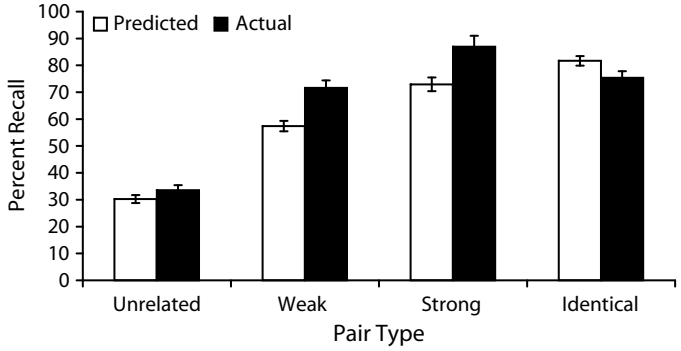

Figure 2. Mean predicted recall (judgments of learning) and actual recall as a function of the four types of word pairs in Experiment 2. Error bars represent standard errors of the mean.

noted that other nonidentical word pairs might also contain item similarity (in the semantic sense), but it appears that the perceptual and semantic similarity in the identical pairs captures subjects, so that they assign high JOLs and engage in short amounts of study time for these pairs. Also, Koriat (1997) stressed that JOLs are comparative in nature, and in the present study, list composition probably played an important role in JOL ratings and memory performance. Relative to the other pairs, the subjects likely inferred that identical pairs would be easiest to recall and, perhaps, did not use elaborative processing to remember the pair. Of course, if an entire study block consisted only of identical pairs, recall would be at ceiling; thus, list composition and study strategies engaged in mixed lists were important factors in the present experiments.

The results from the self-paced study session suggest that fluency plays a critical role, since subjects overestimate recallability of identical pairs. Identical pairs may be perceived as fluent because reading the first word facilitates reading speed of the second word, despite the fact that such fluency may not always be the best predictor of later recall. We argue that subjects use whatever features are most salient when making JOLs, and identical pairs provide fluency that leads to a global assessment of later memory performance. Consistent with the results in Koriat and Bjork (2005), subjects typically do not or cannot assess retrieval conditions when making JOLs (e.g., Kelley & Jacoby, 1996), and it may be the case that with practice (e.g., Koriat & Bjork, 2006), subjects will learn that identical pairs pose a challenge unless sufficient elaborative processing is utilized. This overestimation for identical pairs may be related to the finding that massed practice leads to higher JOL ratings, relative to spaced practice, despite the opposite being true for delayed recall (e.g., Zechmeister & Shaughnessy, 1980). The calibration of JOLs under these conditions is especially important for applied issues, such as studying for tests or examinations, and the present work suggests that the reliance on fluency and repetition when making metacognitive judgments can later lead to erroneous predictions of memory performance.

An illusion of competency was evident in the present study for identical words, consistent with a foresight bias (e.g., Fischhoff, 1975). This foresight bias occurs when predictions about one's success in recalling the correct answer are made in the presence of that answer, as in the present experiments. Koriat and Bjork (2005) also emphasized a "curse of knowledge," in that subjects make JOLs during encoding while in the presence of the answer for the cued recall test. This biases one's judgment to look for relations among the two words in the pairs and then use this as the basis for JOLs. Identical pairs provide the strongest perceived relation between two items, and for this reason, we argue that subjects are somewhat misled by this relation when developing JOLs, despite the fact that associative strength and elaboration are better predictors of later recall. Although the reliance on a global similarity heuristic when JOLs are made often leads to good synchrony between perceived and actual recall, the present findings reveal that under certain circumstances, metacognitive judgments must incorporate more specific properties and processes that influence long-term memory.

AUTHOR NOTE

We thank Endel Tulving for his valuable suggestions regarding the design of the experiments and David Balota, John Dunlosky, Matthew Rhodes, and anonymous reviewers for many useful comments. Andrew Addessi, Rachel Gartner, and Sarah Moynan were instrumental in data collection. Please address all correspondence to A. D. Castel, Department of Psychology, University of California, Los Angeles, Franz Hall, Box 951563, Los Angeles, CA 90095-1563 (e-mail: castel@psych.ucla.edu).

REFERENCES

ARBUCKLE, T. Y., & CUDDY, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, **81**, 126-131.

BEGG, I., DUFT, S., LALONDE, P., MELNICK, R., & SANVITO, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory & Language*, **28**, 610-632.

BENJAMIN, A. S., BJORK, R. A., & SCHWARTZ, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, **127**, 55-68.

DUNLOSKY, J., & MATVEY, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1180-1191.

DUNLOSKY, J., & NELSON, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory & Language*, **33**, 545-565.

FISCHHOFF, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception & Performance*, **1**, 288-299.

HERTZOG, C., DUNLOSKY, J., ROBINSON, A. E., & KIDDER, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 22-34.

KELLEY, C. M., & JACOBY, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory & Language*, **35**, 157-175.

KIMBALL, D. R., & METCALFE, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition*, **31**, 918-929.

KORIAT, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, **126**, 349-370.

KORIAT, A., & BJORK, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 187-194.

KORIAT, A., & BJORK, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, **34**, 959-972.

KORIAT, A., & MA'AYAN, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory & Language*, **52**, 478-492.

NELSON, D. L., MCEVOY, C. L., & SCHREIBER, T. A. (1999). *The University of South Florida word association, rhyme, and word fragment norms*. Available at www.usf.edu/FreeAssociation.

NELSON, T. O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology: General*, **122**, 269-273.

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.

TULVING, E. (1974). Recall and recognition of semantically encoded words. *Journal of Experimental Psychology*, **102**, 778-787.

ZECHMEISTER, E. B., & SHAUGHNESSY, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, **15**, 41-44.

(Manuscript received September 12, 2005;
revision accepted for publication May 15, 2006.)