Metacognition Learning
DOI 10.1007/s11409-015-9144-4

Prior experience shapes metacognitive judgments at the category level: the role of testing and category difficulty

Ruthann C. Thomas[1] · **Bridgid Finn**[2] · **Larry L. Jacoby**[3]

Received: 17 June 2014 / Accepted: 9 June 2015

Abstract Most metacognition research has focused on aggregate judgments of overall performance or item-level judgments about performance on particular questions. However, metacognitive judgments at the category level, which have not been as extensively explored, also play a role in students' study strategies, for example, when students determine what topics to study for an exam. We investigated whether category learning judgments (CLJs) were sensitive to differences in the difficulty of general knowledge categories. After either studying or being tested on facts from several categories (e.g., Shakespeare, Astronomy), participants estimated the likelihood that they could correctly answer new questions from those categories on a later test (i.e., they made CLJs). Results of two studies showed that CLJs were sensitive to differences in category difficulty. Further, participants gave lower or more conservative CLJs when they took an initial test as compared to studying questions from the categories. Results are discussed in terms of the value and relevance of CLJs both in educational settings and in theories of metacognition.

Keywords Metacognition · Category learning judgments · Testing effects · Underconfidence · Judgments of learning

✉ Ruthann C. Thomas
ThomasR@Hendrix.edu

Bridgid Finn
bfinn@ets.org

Larry L. Jacoby
lljacoby@artsci.wustl.edu

[1] Department of Psychology, Hendrix College, 1600 Washington Avenue, Conway, AR 72034, USA

[2] Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA

[3] Department of Psychology, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA

Published online: 04 July 2015

When studying for an exam, students appraise their knowledge of particular topics, such as the formation of the sun or composition of the solar system in an Astronomy class. Students judge their knowledge of a given topic and then use that judgment to choose which topics to emphasize during study. In most cases, students predict their future performance for new (as opposed to previously seen) questions about studied topics; we refer to these predictions as category learning judgments (CLJs; Jacoby et al. 2010). The accuracy of CLJs may have important consequences in educational settings. Misevaluating one's ability to answer new questions about a given topic before an exam could mean the difference between passing and failing. The primary goal of our experiments was to investigate whether students' CLJs were sensitive to differences in the difficulty of educationally relevant semantic categories. An additional goal of our experiments was to explore how prior experience with a category influences the accuracy of category-level predictions. Students often judge their learning of course material after studying, taking a practice quiz at the end of a chapter, or attending lectures. Our question was how particular types of experiences with categories influence CLJs. We compared the effects of prior study and test of a subset of questions from a category on students' predictions of how they would performance on new questions from the same category.

Despite the utility of metacognitive judgments at the category level, most research on metacognitive evaluations during learning involves item-level predictions about future memory performance on particular questions (i.e., judgments of learning, such as the likelihood of recalling that Dover is the capital of Delaware; e.g., Arbuckle and Cuddy 1969; Koriat 1997). People use judgments of learning (i.e., JOLs) to choose which items to study (Metcalfe and Finn 2008) as well as how long to spend on particular items (Son and Metcalfe 2000) and this strategic approach to studying results in effective learning (Kornell and Metcalfe 2006; Thiede et al. 2003). In sum, a large body of research has established the utility of JOLs both for predicting future performance on specific questions and for controlling study behavior (see Metcalfe 2009 for a review).

Current theories that explain the basis of JOLs suggest that predictions of future memory performance involves monitoring various cues that are available during study of an item, such as the ease of learning or the apparent difficulty of an item. According to Koriat's (1997) cue-utilization hypothesis, judgments of learning involve consideration of both intrinsic cues (i.e., characteristics of the study item, such as item difficulty or ease of learning) and extrinsic cues (i.e., conditions of learning or encoding strategies, such as the number of study opportunities or the depth of processing during study). Further, the accuracy of predictions of future memory performance depend on the extent to which these intrinsic and extrinsic cues used to make JOLs also influence later memory performance. The current studies aim to expand our understanding of the multiple bases of metacognitive judgments by considering judgments made at the category level.

Item-level JOLs are a useful metacognitive index about a single piece of information. However, when students learn educational concepts they typically need to learn interconnected facts that have an underlying similarity structure. In an Astronomy class, for example, students may study different concepts within a unit on the solar system, such as the formation and evolution of the Sun, the structure and composition of the solar system, as well as the distinction between the inner and outer solar systems. Accordingly, knowledge of one fact (e.g., formation of sun) may inform a student's understanding of a related fact (e.g., the distinction between the terrestrial planets of the inner solar system and the gas giants of the outer solar system). Metacognitive predictions at the category level expand on JOLs by

capturing the similarity and potential interdependence of memory for educational concepts. Further, CLJs also differ from JOLs in that they involve predicting the likelihood of answering *new* questions about a particular topic (i.e., any question about the solar system) rather than the likelihood of answering the identical question about a particular fact. Predictive CLJs also differ from metacomprehension judgments, which ask students to evaluate how much they have understood about a specific passage they have just read. Similar to CLJs, metacomprehension judgments ask people to evaluate their knowledge at a larger grain size than item-level judgments. However, in contrast to CLJs, they are evaluations about a specific previously studied passage rather than assessments about performance on new, related conceptual information. Metacognitive judgments made at the category level capture both the typical organization of educational concepts and the likely inclusion of new questions about familiar topics on exams. Thus, research on CLJs broadens and advances theories of metacognition by providing a framework for judgments about category level understanding, by tapping into students' predictions of their future performance on new questions made at the category level.

Recently, Jacoby, Wahlheim, and colleagues (Jacoby et al. 2010; Wahlheim et al. 2011; Wahlheim et al. 2012) initiated investigations of CLJs in the context of learning to classify perceptual categories. After studying specific exemplars of birds along with their family names, participants predicted the likelihood that they would correctly classify new exemplars of studied bird families on a later test (i.e., participants made CLJs). Results from these studies demonstrated that CLJs showed sensitivity to the benefits of some study conditions that promote effective learning: Participants predicted and achieved better classification of new birds following repeated testing compared to repeated studying (Jacoby et al. 2010) and following spaced as compared to massed studying (Wahlheim et al. 2011) of other exemplars from the same families. In contrast, Wahlheim et al. (2012) found that CLJs did not differ as a function of the number of unique exemplars of a family that were studied, even though variability in studied exemplars improved later classification of new birds over studying the same repeated exemplars (see e.g., Dukes and Bevan 1967). Repetition of the same exemplars may be associated with increased fluency or ease of processing, which participants then misattribute to better learning. Thus, similar to other metacognitive judgments (e.g., Begg et al. 1989; Metcalfe et al. 1993; Rhodes and Castel 2008), CLJs are also susceptible to fluency biases (Wahlheim et al. 2012). The research to date on CLJs suggests that CLJs for natural concepts (i.e., bird families) are sensitive to differences in category difficulty. Further, they are sensitive to the benefits of some effective study behaviors (Jacoby et al. 2010; Wahlheim et al. 2011) but not all (Wahlheim et al. 2012).

Overview of current experiments

In the current experiments, we extended research on CLJs to verbal materials similar to those learned in classroom settings. Prior research with CLJs (Jacoby et al. 2010; Wahlheim et al. 2011; Wahlheim et al. 2012) involved learning to classify exemplars of bird families based on perceptual details (e.g., length and shape of beak). In the current experiments, we investigated CLJs with various categories of general knowledge questions (e.g., Shakespeare, Astronomy, Modern Art). After either studying or being tested on facts from different categories, participants were asked to estimate the likelihood that they could correctly answer new trivia questions about a particular category on a later test.

A primary goal of this research is to better understand the basis for CLJs by investigating both intrinsic cues (here, category difficulty) as well as extrinsic cues (here, study and test experiences). Although pre-existing knowledge is likely to guide CLJs, recent learning experiences are also likely to provide a basis for these judgments. An additional goal of the current experiments was to explore how specific learning experiences with a subset of questions about a category influence CLJs about new questions from the same categories. In particular, we investigated how experiences studying as compared to testing influenced judgments of category level knowledge.

Before taking an exam, students often restudy or test themselves on course material either through textbook review sections, online quizzing portals, or questions provided by the instructor. For example, in research on students' study habits, most students report restudying their notes or their textbook as their primary study technique (Karpicke et al. 2009; Kornell and Bjork 2007). College students who report using self-testing as a common study strategy largely use testing to diagnose learning rather than to improve it (Karpicke et al. 2009; Kornell and Bjork 2007). That is, students tend to test themselves to assess what they know but not to intentionally bolster their learning. Recent research has demonstrated that although students show better long-term retention if they are tested on course material rather than simply restudying it (i.e., the testing effect; see Karpicke 2012, for a review), their immediate JOLs do not typically reflect this benefit (Karpicke et al. 2009; Kornell and Bjork 2007). Indeed, after taking a test, students are less confident in their learning than when they study without an intervening test (Finn and Metcalfe 2007, 2008; Koriat 1997; Koriat et al. 2002; Koriat and Bjork 2006; Meeter and Nelson 2003). Finn and Metcalfe (2007, 2008) found that students use their prior test performance to moderate their learning judgments, and suggested that lower confidence following a test may serve the adaptive purpose of highlighting items that could benefit from an additional study. In the current experiments, we investigated whether testing may similarly result in lower category confidence (i.e., lower CLJs) than when students studied without being tested.

In Experiment 1, participants either studied or were tested with feedback on a blocked subset of facts from each of six different categories (e.g., Shakespeare, Lions, The Human Body, Astronomy). After each item, participants estimated how likely they would be to correctly answer the question on a later test (i.e., immediate JOLs). After completing the initial study or test phase with questions from all six categories, participants then estimated the likelihood that they would be able to correctly answer new questions from each of the six categories (i.e., they made CLJs). A primary interest was in the sensitivity of CLJs to differences in category difficulty. Our index of category difficulty was based on the average proportion correct on questions from particular categories (e.g., average on questions about Astronomy) from data collected in several pilot studies[1], which we refer to as *normative difficulty*. Thus, sensitivity to category difficulty was evidenced in the correspondence between CLJs and the normative difficulty of the categories. Prior research with perceptual categories showed that CLJs are sensitive to category difficulty (Jacoby et al. 2010; Wahlheim et al. 2011; Wahlheim et al. 2012). We expected to extend that finding to verbal materials. Another goal was to explore whether studying as compared to testing influenced CLJs. We expected that

[1] Normative data was calculated using mean performance from an average of 290 participants who answered these same questions across a series of pilot experiments using trivia categories. In the pilot studies, the questions were always new to the experiment (i.e., participants has not encountered the question or answer previously) and presented in the same session as other questions from the same categories.

initial testing might lead students to be less confident in their category knowledge than students who studied without a test.

An additional question about prior experience with a category concerns how the difficulty of previewed questions guides CLJs. In Experiment 2, participants either studied or took a test on all easy or all difficult questions from a variety of categories. In so doing we manipulated the difficulty of the questions seen before participants made their CLJs to determine if experienced ease would influence CLJs. Further, in Experiment 2 students were given a final test allowing us to explore how CLJs predicted participants' own ability to correctly answer new questions. Together, these manipulations allowed us to further characterize the factors that influence metacognitive judgments made at the category level.

Experiment 1

Methods

Participants Eighty-four participants (30 male, 54 female) were recruited from the Washington University student participant pool. Participants were given course credit or monetary compensation ($10/hour) for their participation. Participants were randomly assigned to initial study or test groups. There were 40 participants in the initial study group and 44 participants in the initial test group.

Design Experience with Category (initial study vs. initial test) was manipulated between participants. The primary dependent measures were the magnitude of JOLs and CLJs as well as the correspondence between CLJs and normative difficulty of the categories (i.e., absolute and relative accuracy of CLJs).

Materials The materials were 144 general knowledge questions with 16 questions from each of nine categories (Astronomy, U.S. Civil War, Diseases, Movies, Musical Instruments, Lions, Modern Art, The Human Body, and Shakespeare). Some were selected from Nelson and Narens (1980) set and others were created based on facts found in books and on the Internet. Based on data from pilot studies[1] using the same questions and similar testing procedures, normative performance on questions from these categories ranged from 0.19 to 0.52 proportion correct ($M = 0.37$, $SD = 0.25$). In this experiment, the initial study/test and CLJs phases included only six of the nine categories, with each of the nine total categories counterbalanced to occur equally often across participants.

Procedure Participants were tested in groups of one to six people, with each in front of their own computer in a cubicle. The experiment consisted of two phases: an initial study/test phase followed by a CLJs phase. In the initial study/test phase, participants either studied questions with their correct answers (Initial Study group) or were prompted to provide an answer to each question (Initial Test group) for six categories (i.e., the old categories). Participants were presented with eight questions from each of the six categories in blocks. Before each category block, participants were told they were about to see trivia questions about that category (e.g., Shakespeare). The category name was also presented with each individual question. The orders of category blocks and of questions within each category block were randomized for each participant. In the Initial Study group, the category name along with each trivia question

and answer appeared onscreen to study for 5.5 s. In the Initial Test group[2], the category name along with a trivia question appeared above a response box where participants' typed response appeared onscreen. The question remained onscreen for 10 s followed by correct answer feedback for 3 s. Immediately after each question, participants in the Initial Study and Initial Test groups made a cue-only JOL. The target or the response given by the participant was cleared from the screen leaving only the cue question and a JOL sliding scale that ranged from 0 (low confidence) −100 (high confidence) in the bottom of the screen, with an arrow set at 50. Participants were instructed to use the mouse to drag the arrow to the point on the scale that represented their level of confidence with 0 indicating low confidence that they would be able to answer that question correctly on the follow up test and 100 indicating high confidence that they would be able to answer the question correctly on the follow up test. They were encouraged to use the full range of the scale. After using the sliding scale to make a confidence rating, participants clicked an Enter button and moved onto the next question.

After the initial study/test phase, participants made CLJs for each of the six categories. Participants were presented with six category names one at a time, and asked to predict their performance for new questions from this category. CLJs were also made on a slider scale ranging from 0 to 100 confidence using the same basic procedure used for JOLs. At the end of the study, participants were debriefed and thanked for their time.

Results and discussion

Overall, the average proportion correct on the initial test in the initial test group was 0.35 (SD= 0.16), which is similar to normative difficulty from pilot studies (M=0.37, SD=0.25). Our primary interest was in the sensitivity of CLJs to differences in normative category difficulty. Category learning judgments involve predictions of performance on new questions from the categories presented in the initial study/test phase. We were also interested in the JOLs made for questions presented in the initial study/test phase. We examined the effects of prior experience with the category on both the magnitudes of JOLs and CLJs as well as the correspondence between CLJs and normative difficulty using an Analysis of Variance (ANOVA). In Experiment 1, a final trivia test was not included; thus, we determined the relative accuracy of CLJs by comparing CLJs for each category to normative difficulty (i.e., proportion correct on questions from the same categories in pilot studies). In addition, we investigated the resolution of CLJs by calculating Pearson product-moment correlations between CLJs and normative difficulty. In each of the experiments, the significance level for all statistical tests was set at α=.05.

Judgments of learning The magnitude of JOLs was higher for the initial study compared to the initial test group (71.7 vs. 59.6), $F(1, 82)=15.74$, $\eta^2_p=0.16$. Participants predicted that they would correctly answer more questions on the later test if they initially studied the answers than if they were tested with feedback on the same questions. We report resolution between JOLs and normative performance on particular questions (based on pilot data) as a measure of the relative accuracy of JOLs for the sake of comprehensiveness; however, JOL resolution was

[2] Higher proportion correct on the final test may be due to prior exposure to different questions from the same categories that participants previewed in the initial study/test phase.

not a central focus. There was no difference in resolution between JOLs and normative performance between conditions, as measured by a gamma correlation (−0.02 vs. 0.01 for study and test groups, respectively), $F<1$. Gamma correlations for both conditions were not significantly different than zero, largest $t(39)=1.94$, $p>.05$. For some participants, a gamma correlation could not be computed because they got everything right or everything wrong, or had too many ties and the statistic could not be computed. Thus, degrees of freedom listed for gamma correlations may differ from the total number of participants used in the experiment). Next, we explore our main question of interest: whether category level metacognitive predictions about new questions also reflect prior experience with the categories.

Category learning judgments Similar to the pattern found with JOLs, the magnitude of CLJs was higher for the initial study compared to the initial test group (56.0 vs. 47.8), $F(1, 82)=5.47$, $\eta^2_p=.06$. Participants predicted that they would correctly answer more new questions on the later test if they initially studied answers than if they were tested with feedback on different questions from the same categories.

Our main question was whether CLJs were sensitive to differences in category difficulty, which would be evidenced by a high correspondence between CLJs and actual normative difficulty of the categories. Further, we were interested in whether the sensitivity of CLJs differed based on participants' prior experience with the categories (i.e., study or test).

As our primary measure of CLJ sensitivity, we examined the absolute accuracy of CLJs with a calibration measure in which predictions of performance on new questions from the category were compared to normative difficulty on questions from the categories. Calibration scores were computed by averaging the signed difference score between CLJs and normative difficulty across categories for each participant, with a score of zero representing perfect calibration. Although participants were overconfident in their performance on new questions from the categories in both groups (+15.0; all calibration scores were significantly different from zero; ts>5.00), participants were better calibrated (i.e., less overconfident) if they were initially tested rather than if they initially studied other questions from the categories (+11.2 vs. +19.2), $F(1, 82)=5.27$, $\eta^2_p=.06$. These results show that prior test experience with the categories improved the calibration of CLJs. Participants gave lower CLJs following an initial test compared to study experience, which brought their estimates of category difficulty closer to actual normative difficulty.

We also examined the relative accuracy of CLJs with a resolution measure that reflects the ability to discriminate between categories that are easier or more difficult than others. Resolution at the category level was measured by computing a Pearson product-moment correlation between average CLJs (i.e., averaged across participants) for each category and normative difficulty for each category. Figure 1 displays the normative difficulty as well as CLJs for initial study and test groups for each of the categories. Visual inspection of this figure suggests a strong correspondence between CLJs and normative difficulty of the categories. Category learning judgments from both the initial study and test groups increased as categories got easier. Consistent with this observation, there were strong positive correlations between mean CLJs and normative difficulty for new questions, averaged at the levels of category and participants, for both initial study ($r=0.90$, $p<.05$) and test groups ($r=0.87$, $p<.05$). Further, to show that individual participants were sensitive to differences in category difficulty, we computed Pearson product-moment correlations between each participant's CLJs and normative difficulty from each category, then calculated the average correlation across participants. As in the previous analysis, there were strong positive correlations between CLJs and

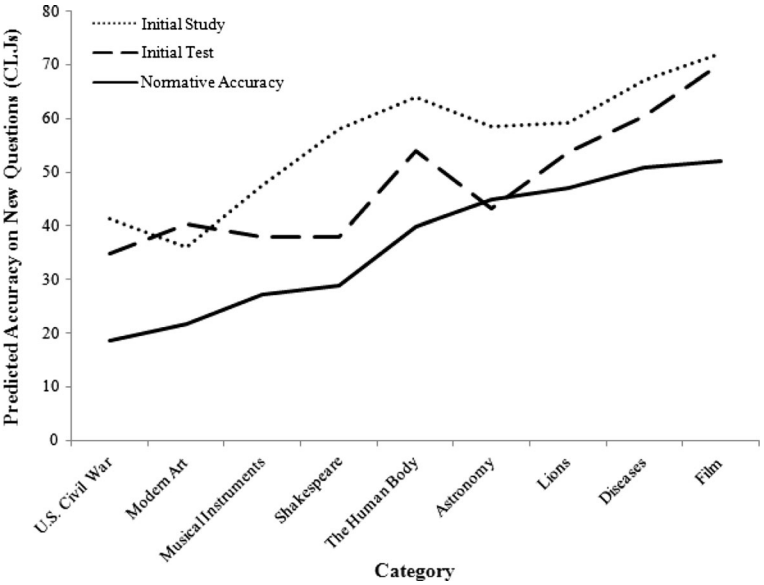

Fig. 1 Mean CLJs as a function of experience with category compared to normative difficulty on individual categories in Experiment 1

normative difficulty for both initial study ($M_r = 0.46$) and test groups ($M_r = 0.45$), but the resolution did not differ between the two groups, $F < 1$.

Summary In sum, results of Experiment 1 demonstrate that metacognitive judgments at the category level are sensitive to differences in normative category difficulty. Participants gave higher CLJs to easy categories and lower CLJs to more difficult categories. Further, when participants have prior test experience with a subset of questions from the category, they predict poorer performance on new questions from those categories compared to participants with prior study experience. The lower, or more modest, CLJs in the prior test group diminished their confidence to improve the absolute accuracy of CLJs for each category. However, both initial study and initial test groups were able to discriminate between relatively easy and difficult categories, as indicated by the good resolution of both groups' CLJs. In an educational setting, students may be less confident in their performance on all topics on an exam if they test themselves, which may lead students to spend to more time studying (Finn and Metcalfe 2007, 2008; Metcalfe and Finn 2008). However, the results here suggest that students would be able to differentiate between easy and hard topics regardless of whether they studied or tested themselves on the various topics.

Experiment 2

In Experiment 1, we found that CLJs were sensitive both to the type of prior experience (i.e., test versus study) as well as to differences in category difficulty. The goals of Experiment 2 were to replicate the results of Experiment 1 and to extend these findings by exploring how CLJs were influenced by an additional characteristic of prior category experience: the

difficulty of the previewed questions. In Experiment 1, the variability in category difficulty was inherent in the selected materials (e.g., students knew less about the U.S. Civil War than Modern Art prior to the experiment). In Experiment 2, we selected easy and difficult questions within each category to manipulate the difficulty of participants' initial experience with questions from the same categories. As in Experiment 1, participants either studied or took a test on questions from a variety of categories before making their CLJs. However, half of the participants were presented with extremely difficult questions whereas the other half were presented with easier questions. Our interest here was in an additional measure of sensitivity to category difficulty by investigating whether easier experiences would lead to higher CLJs as compared to more difficult initial experiences. Our prediction was that CLJs would track experience with the category such that higher CLJs would result after easier experiences and lower CLJs following more difficult experiences. In Experiment 1, we measured the sensitivity of CLJs by examining the relationship between CLJs and actual normative difficulty on new questions from the categories. A second extension of Experiment 2 was the inclusion of a final criterion test, which allowed us to investigate how CLJs predicted participants' own performance on new questions from the categories. Whereas normative difficulty was an appropriate index of typical category knowledge, the comparison of CLJs with actual performance would allow greater sensitivity to individual differences in category knowledge. As in Experiment 1, we investigated the absolute accuracy of CLJs using a calibration measure comparing predicted performance to both (a) normative difficulty, as in Experiment 1, as well as (b) final test performance.

Methods

Participants Ninety-five participants (37 male, 58 female) were recruited from the Washington University student participant pool. Participants were given course credit or monetary compensation ($10/hour) for their participation. Participants were randomly assigned to difficult or easy groups and initial study or test groups. Each of the four groups included 23 or 24 participants.

Design This study was 2 (Experience with Category) X 2 (Difficulty Group) X 2 (Question Type) mixed factorial design, with Experience with Category (initial study, initial test) and Question Difficulty (easy, hard) as between participants factors and Question Type on the final test (previewed question, new question) as a within participants factor. There were four independent groups (Initial Study / Easy Questions, Initial Study / Hard Questions, Initial Test / Easy Questions, Initial Test / Hard Questions). As in Experiment 1, the primary dependent measures were the magnitude of JOLs and CLJs as well as the correspondence between CLJs and actual difficulty of the categories (i.e., absolute and relative accuracy of CLJs). In addition, we also report final test performance along with the correspondence between CLJs and participants' own performance on the final criterion test (i.e., another measure of the absolute and relative accuracy of CLJs).

Materials The materials were 216 general knowledge questions with 24 questions from each of the same nine categories used in Experiment 1. For each category, half of the questions were easier (M=0.48, SD=0.09; range=0.36–0.58) and half of the questions were very difficult (M=0.13, SD=0.06; range=0.06–0.22) based on normative data from pilot experiments. Easy and difficult questions were divided into two sets matched on normative difficulty. Based on

Difficulty group, participants either saw easy or difficult questions during the initial study/test phase. Unlike Experiment 1, participants saw a subset of questions from all nine categories. In the final test phase, participants were presented with six previewed questions as well as six new questions matched in difficulty to the questions seen in the initial study/test phase.

Procedure Participants were tested in groups of one to six people, with each in front of their own computer in a cubicle. The experiment consisted of five phases: an initial study/test phase, a CLJs phase, a filled interval, a test phase, and a post-CLJs phase. In the initial study/test phase, participants either studied questions with their correct answers (Initial Study group) or provided answers to the previewed questions (Initial Test group) for all nine categories. Participants were presented with six questions from each of the nine categories in blocks. These questions were either easy or difficult, depending on the assigned Difficulty Group. Before each category block, participants were told they were about to see trivia questions about that category (e.g., Shakespeare). The category name was also presented with each individual question. In the Initial Study group, the category name along with each trivia question and answer appeared onscreen to study for 5.5 s. In the Initial Test group, the category name along with a trivia question appeared above a response box where participants' typed response appeared onscreen for 10 s followed by correct answer feedback for 3 s. After each question, participants in the Initial Study and Initial Test groups made a JOL. As in the prior experiment, JOLs were cue only and were made on a slider scale ranging from 0 to 100 %. There was no time limit to make the judgment.

After the initial study/test phase, participants made CLJs for each of the nine categories. Participants were presented with nine category names one at a time, and asked to predict their performance for new questions from this category. As in the prior experiment, CLJs were made on a slider scale ranging from 0 to 100 %. There was no time limit to make the judgment.

After all CLJs were made, participant moved onto the final test phase. During the final test phase 6 new questions and 6 previewed questions from each category were presented. The normative difficulty of the new questions in each category matched the difficulty level (i.e., easy vs. difficult) of the initial questions allowing assessment of CLJ accuracy. That is, participants in the Easy Group made predictions about and then received easy, new questions of the final test. Questions were blocked by category. A single question was presented and participants had as much time as they needed to provide a response. After entering a response, they clicked an enter button and moved onto the next question. After answering all 108 questions participants were shown a category label and asked to make two different types of retrospective CLJs for each category. Participants were first asked to indicate on a slider scale about how well they had performed overall on questions for a given category. Next they were asked to make a CLJ about their performance on only the new items from that particular category. After making the CLJs for each category the experiment ended. Participants were debriefed and thanked for their time.

Results and discussion

Overall, the average proportion correct on the initial test was 0.05 (SD=0.04) in the difficult group and 0.49 (SD=0.14) in the easy group, which is comparable to normative difficulty from pilot studies (M=0.13, SD=0.06 for easy, and M=0.48, SD=0.09 for difficult). As in

Experiment 1, our primary interest was in the sensitivity of CLJs to differences in the difficulty of categories. In addition, we also included a final test on previewed and new questions from the categories after participants made their CLJs to investigate the correspondence between their CLJs and their own future performance.

First, we report the effects of prior experience with the category and category difficulty on participants' performance on the final test and on the magnitudes of JOLs and predictive CLJs (henceforth referred to as CLJs) using 2×2 ANOVAs. Then, we examine accuracy of CLJs through their correspondence with two types of baseline measures of category difficulty: 1) normative difficulty from a separate study, and 2) participants' own final test performance. Finally, we report on the magnitudes of retrospective CLJs made by participants after the final test.

Final test performance First, we compared participants' final test performance for previewed and new questions as a function of prior experience with the category and question difficulty (see Table 1). Our main interest in this analysis was to investigate whether prior testing experience improved performance on previewed questions on the later test. Indeed, we found a significant Question Type X Experience with Category interaction, $F(1, 91)=4.58$, $\eta^2_p=.048$, suggesting that participants who took an initial test correctly answered more previewed questions on the final test than participants who initially studied the questions and answers (0.86 vs. 0.82). This difference did not quite reach significance however, $p>.05$. Participants in the initial test and initial study groups did not differ in their ability to correctly answer new questions on the final test (0.47 vs. 0.47). That is, the interaction was suggestive of a pattern of results consistent with a testing effect. In addition, a significant Question Type X Difficulty Group interaction, $F(1, 91)=300.56$, $\eta^2_p=.77$, suggested that participants in the difficult group learned more from initial exposure to the previewed questions than did participants in the easy group. In the easy group, participants were likely to know more of the answers to the questions before the experiment; thus, there was less new learning in the initial study/test phase.

Metacognitive predictions of performance: judgments of learning Based on prior research, we expected and found that JOLs were sensitive to the difficulty of the previewed questions (see Table 2). Participants predicted higher accuracy on the same questions on the final test when the previewed questions were easy compared to difficult (74.0 vs 54.1.), $F(1, 91)=49.28$, $\eta^2_p=.35$. In contrast to Experiment 1, the magnitude of JOLs did not differ

Table 1 Mean proportion correct on final test for previewed and new questions as a function of prior experience with the category and difficulty group in Experiment 2

	Difficult group		Easy group	
Prior experience with category	M	SD	M	SD
Initial study group				
Previewed questions	0.74	0.13	0.89	0.16
New questions	0.22	0.09	0.73	0.20
Initial test group				
Previewed questions	0.75	0.16	0.97	0.04
New questions	0.20	0.09	0.75	0.12

Table 2 Mean metacognitive predictions (JOLs and CLJs) and retrospective judgments of performance as a function of experience with category and difficulty group in Experiment 2

	Difficult group		Easy group	
Metacognitive judgment	M	SD	M	SD
JOLs (previewed questions)				
Initial study group	56.3	8.3	71.2	15.9
Initial test group	51.9	17.2	76.9	11.8
CLJs (new questions)				
Initial study group	43.1	8.8	57.9	14.0
Initial test group	29.6	17.0	50.5	14.1
Retrospective CLJs (all questions)				
Initial study group	54.9	13.2	73.1	15.9
Initial test group	44.2	19.6	74.9	11.7
Retrospective CLJs (new questions only)				
Initial study group	34.9	14.3	61.0	15.5
Initial test group	32.8	18.3	65.6	16.1

between initial study and test groups, $F<1$, which may be due to the fact that the performance difference between the study and test condition though significant was rather small. In addition, participants experienced all easy or all difficult items, which may have minimized the cue information arising from study or test. The Experience with Category X Difficulty interaction did not reach significance, F (1, 91)=3.12, p= .08, η^2_p=.03. Resolution between JOLs and the participant's own final test performance showed only a main effect of question difficulty, with higher resolution in the easy as compared to difficult group (0.66 vs. 0.41), F(1, 78)=7.83, η^2_p=.09. As in Experiment 1, gamma correlations could not be computed for some participants. This difference is difficult to interpret because the difficult group learned more than twice as much during the initial test phase than the easy group. Next, we explore whether category level metacognitive predictions about new questions also show sensitivity to category difficulty.

Category learning judgments First, we compared CLJs as a function of prior experience with the category and difficulty of the questions (see Table 2). As in Experiment 1, the magnitude of CLJs was higher for the initial study compared to the initial test group, F(1, 91)=13.53, η^2_p=.13. Participants predicted higher accuracy on new questions on a later test if they initially studied answers than if they were tested with feedback on different questions from the same categories.

Next, we examined whether CLJs were sensitive to category difficulty in two ways. First, we compared the magnitude of CLJs when the initial experience with the category consisted of either easy or difficult questions. When the initial questions were easy, participants predicted higher accuracy on new questions on a later test than when the initial questions were difficult, F(1, 91)=39.28, η^2_p=.30, and this effect did not depend on whether these questions were studied or tested (i.e., the Experience with Category X Difficulty Group interaction was not significant, $F<1.17$, $p>.10$). Results are consistent with the finding in Experiment 1 that CLJs track category difficulty, with higher CLJs to easier categories and lower CLJs to more difficult categories. Further, these results also extend findings from Experiment 1 to demonstrate that

the difficulty of questions seen during the initial study or test phase guides predictions of new questions from the same categories.

Second, as in Experiment 1, calibration scores were used as a measure of the absolute accuracy of CLJs. First, calibration scores were computed by taking the signed difference between CLJs and normative difficulty, with a score of zero representing perfect calibration. The top rows of Table 3 display calibration scores as a function of Experience with Category and Difficulty when normative difficulty was used as the baseline. All calibration scores were significantly different from zero, all *t*s>2.90. Although participants were overconfident in their predictions of performance on new questions from the categories overall (+14.7; t (95)=8.47), participants were better calibrated (i.e., less overconfident) if they were initially tested rather than if they initially studied other questions from the categories (+9.6 vs. +20.1), $F(1, 91)=13.53$, $\eta^2_p=.13$. Further, participants were also better calibrated (i.e., less overconfident) when initial questions were easy as compared to difficult (+6.5 vs. +23.3), $F(1, 91)=35.00$, $\eta^2_p=.28$. The Experience with Category X Category Difficulty was not significant, $F<1.17$, $p>.10$. As in Experiment 1, the lower, or more modest CLJs, in the prior test group diminished their confidence to improve the absolute accuracy of CLJs for each category. Thus, prior test experience with a subset of questions from a category brought predictions of performance on new questions in line with actual normative difficulty, compared to prior study experience.

To supplement our calibration measure with normative difficulty, we also calculated the correspondence between CLJs and participants' own final test performance on new questions the categories. The bottom rows of Table 3 display calibration scores as a function of Experience with Category and Difficulty when participants' own performance on the final test was used as the baseline. All calibration scores were significantly different from zero, all *t*s>2.97. There was a main effect of question difficulty group, $F(1, 91)=105.89$, $\eta^2_p=.54$. Overall, participants were overconfident in their CLJs when the initial questions were difficult (+15.2). In contrast, they were under confident in their CLJs when the initial questions were easy (−19.4). This pattern corresponds with the *hard–easy effect* typically observed in choice confidence (e.g., Lichtenstein and Fischhoff 1977) in which people tend to be overconfident in answers to difficult questions and under confident in answers to easy questions. These data suggest that the hard–easy effect extends to predictions of performance on new questions from studied categories. There was also a main effect of Experience with Category, $F(1, 91)=9.85$,

Table 3 Mean absolute accuracy (i.e., calibration bias) and relative accuracy (i.e., resolution) of CLJs as a function of experience with category and difficulty of questions in Experiment 2

	Difficult group		Easy group	
Baseline measure for calibration	M	SD	M	SD
Calibration (normative difficulty baseline)				
Initial study group	+30.1	8.8	+10.1	14.0
Initial test group	+16.5	17.0	+2.8	14.1
Calibration (own performance baseline)				
Initial study group	+21.2	11.8	−14.8	21.5
Initial test group	+9.3	15.3	−24.1	15.5
Resolution (within participants)				
Initial study group	0.06	0.24	0.20	0.28
Initial test group	0.07	0.37	0.34	0.25

η^2_p=.10, but the Experience with Category x Difficulty interaction was not significant, F< 1. Participants were overconfident in their CLJs when they initially studied the questions and answers (+3.19), but they were under confident when they took an initial test (−7.39).

When normative difficulty was used as the baseline for calibration, the lower CLJs in the initial test group were more accurate (i.e., in line with normative difficulty) than CLJs in the initial study group. However, when participants' own performance on the final test was used as the baseline for calibration, these lower CLJs resulted in underconfidence that did not improve calibration relative to the initial study group. This apparent discrepancy in results is likely due to differences in baseline levels of performance from normative difficulty and participants' own performance on the final test. Although there was a high correlation between participants' final test performance on new questions and normative difficulty on the same questions from the categories (r= 0.95 for easy, r= 0.95 for difficult, ps<.05), participants correctly answered more questions on the final test compared to normative data on the same questions in both difficult and easy groups[2]. In sum, across both measures of calibration, participants gave lower, or more conservative, CLJs when they had an initial test compared to an initial study experience.

In addition, we examined the relative accuracy of CLJs with a resolution measure that reflects the ability to discriminate between categories that are easier or more difficult than others. Table 2 displays both resolution measures as a function of Experience with Category and Difficulty. Resolution at the category level was measured by computing a Pearson product-moment correlation between CLJs, averaged across participants for each category, and normative difficulty for each category. Importantly, our manipulation of category difficulty restricted the range of normative difficulty across categories for both easy and difficult groups. Given that restricted range may limit our ability to detect meaningful correlations, we exercise caution in interpreting nonsigificant correlations. As in Experiment 1, there was a significant positive correlation between mean CLJs and normative difficulty for new questions, averaged at the levels of category and participants; r=.74, p<.05. Further, to show that individual participants were sensitive to differences in category difficulty, we computed Pearson product-moment correlations between each participant's CLJs and normative difficulty from each category, then calculated the average correlation across participants. As in the previous analysis, there was a significant positive correlation between CLJs and normative difficulty overall across participants (M_r= 0.17; t (94)=5.37). The resolution of CLJs did not differ between initial study and test groups (0.13 vs. 0.20, respectively) F<1.37, p>.10. The resolution of CLJs was higher when the initial questions were easy compared to difficult (0.27 vs. 0.07), F(1, 91)=11.32, η^2_p=.11. However, this difference should be interpreted with caution. The manipulation of category difficulty restricted the range of normative difficulty across categories, thus limiting our ability to detect a meaningful relationship between CLJs and normative difficulty. Accordingly, the low correlation for difficult questions may be due to restricted range of normative difficulty across categories when the questions were very difficult. Consistent with results of Experiment 1, the ability to discriminate between easier and more difficult categories was not influenced by initial study or test experiences with the categories.

Retrospective CLJs In addition to the CLJs given before the final test, we also examined retrospective CLJs about final test performance to get a better understanding of whether initial study/test experiences influenced evaluation of past performance on the final test. We compared mean retrospective CLJs for overall performance as well

as performance on new questions as a function of prior experience with the categories as well as difficulty group (see Table 2). First, participants evaluated how well they had performed overall on questions for a given category. Not surprisingly, retrospective CLJs about all questions were sensitive to question difficulty, $F(1, 91)=59.82$, $\eta^2_p=.40$, with higher CLJs when the initial and final test questions were easy than when they were difficult (74.0 vs. 49.5). Further, the Experience with Category X Difficulty Group interaction approached significance, $F(1, 91)=3.92$, $\eta^2_p=.04$, $p=.051$. When the questions were difficult, participants reported that they performed worse on all questions from the categories when they took an initial test than when they initially studied a subset of questions. However, when the questions were easy, prior study/test experience did not influence participants' judgments of how they had performed on the final test. Thus, even though participants who took an initial test performed just at least as well on the final test as participants who initially studied items, they judged their final performance on all questions more harshly.

Second, participants evaluated how well they had performed on *new* questions only for a given category. Not surprisingly, retrospective CLJs about new questions were also sensitive to question difficulty, $F(1, 91)=78.65$, $\eta^2_p=.46$, with higher CLJs when the initial and final test questions were easy than when they were difficult (63.3 vs. 33.8). However, prior/study test experience with previewed questions did not influence retrospective CLJs about new questions from the categories, all $Fs<1.10$. Thus, in contrast to predictive CLJs and retrospective CLJs about all questions, retrospective CLJs about new questions were not lower or more conservative for participants who took an initial test compared to participants who initially studied items.

General discussion

Results of two experiments support our key hypothesis that CLJs track category difficulty (Experiments 1 and 2) and predict participants' future performance on new questions from the categories (Experiment 2). Category learning judgments were sensitive to category difficulty, as revealed by participants' ability to discriminate between easier and more difficult categories in both experiments (i.e., high relative accuracy of CLJs) as well as higher magnitude of CLJs for categories with easy as compared to difficult questions (Experiment 2). Further, category difficulty influenced both CLJs and actual performance on trivia questions from the categories, consistent with prior research on the influence of intrinsic cues on JOLs (Koriat 1997). That is, participants gave higher CLJs to easier categories compared to more difficult categories in both experiments, and they correctly answered more questions from easier categories compared to more difficulty categories.

Another focus of our studies was how initial learning experiences (study versus test) influenced CLJs. Across both experiments, participants gave lower or more conservative CLJs when they took an initial test as compared to studying questions from the categories. This pattern –less confident predictions following a test as compared to study– has been demonstrated with other metacognitive judgments (e.g., JOLs; Finn and Metcalfe 2007, 2008; Metcalfe and Finn 2008) and suggests that students may rely on their prior test performance as a cue to moderate their CLJs. In addition, CLJs were better calibrated (i.e., CLJs were more closely aligned with

normative difficulty) when participants took an initial test rather than studying previewed questions from the categories. The observed metacognitive caution following the initial test may serve an important function in learning by helping students to identify when category knowledge needs further elaboration.

The finding of lower CLJs following testing as compared to study is also in line with previous research on CLJs for natural concepts (Wahlheim et al. 2012) as well as a large body of research on other metacognitions (e.g., Kelley and Jacoby 1996; Koriat et al. 2004) that demonstrate that processing fluency can influence assessments of knowledge. Students are typically less confident in their knowledge after testing (e.g., Karpicke 2009). Retrieving an answer to a test question may feel less fluent than restudying the answer. As a result, students may come to believe that they knew less about a topic after testing as compared to restudying (Kornell et al. 2011). Although prior experience with the category influenced absolute accuracy of CLJs, it did not differentially impact the relative accuracy of CLJs. This is a distinct pattern of results from that seen with item-level JOLs in which testing leads to improvements in relative accuracy as compared to a study only group (Finn and Metcalfe 2008). The absolute and relative accuracy of CLJs may have different consequences on study strategies. As a student prepares for their upcoming psychology exam, they must first choose a category or topic to start with and determine how long they plan to spend and which subcategories to cover. Relative accuracy across categories, as determined by the relationships among CLJs for different categories, may inform the category a student chooses to study, based on their assessment of which topic they know better than others. However, absolute accuracy, as determined by the magnitude of a given CLJ, may inform the amount of study time assigned to a given topic. Students who study and students who test themselves to prepare for a class exam may be equally aware of which categories are difficult or not. Thus, a student's prior study or test experience with the category may not inform which category they choose. Indeed, initial study and test groups did not differ in the relative accuracy of their CLJs. However, students may decide to dedicate more time to a topic if they took a prior test than if they only studied the answers. Lower confidence in category knowledge following testing may push students to allocate additional study time and develop their category understanding further. Indeed, Metcalfe and Finn (2008) found that students' metacognitive judgments directly influence their study choices. Though we did not investigate the outcome of lower CLJs on study choices, it is a topic that deserves further investigation.

In sum, the current experiments attempted to establish CLJs as a diagnostic metacognitive assessment of category knowledge. Results of the current studies join recent research (Jacoby et al. 2010; Wahlheim et al. 2011; Wahlheim et al. 2012) to establish the value of metacognitive judgments at the category level. The findings reveal several types of cues that people use to make their category level judgments. When people make predictions about their category level knowledge, they are sensitive to category difficulty. Further, results suggest people use their prior experiences with the category as a basis for CLJs. Research on CLJs advances metacognition research by going beyond item-level judgments to capture the organization and interrelationships among educational concepts. The current experiments provide a promising first step towards establishing the value of CLJs for predicting future performance using educationally relevant verbal materials.

Acknowledgments This research was supported by Grant 22020166, Applying Cognitive Psychology to Enhance Educational Practice, from the James S. McDonnell Foundation, awarded to Larry L. Jacoby. We are grateful to Chris Wahlheim for his input the procedure as well as Rachel Teune and Kara Chung for their assistance with data collection and data scoring.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81*, 126–131. doi:10.1037/h0027455.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610–632. doi:10.1016/0749-596X(89)90016-8.

Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology, 74*, 178–181. doi:10.1037/h0024575.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 238–244. doi:10.1037/0278-7393.33.1.238.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19–34. doi:10.1016/j.jml.2007.03.006.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1441–1451. doi:10.1037/a0020636.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486.

Karpicke, J. D. (2012). Retrieval-based learning: active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21*, 157–163. doi:10.1177/0963721412443552.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory, 17*, 471–479. doi:10.1080/09658210802647009.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175.

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. (2006). Mending metacognitive illusions: a comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1133–1145. doi:10.1037/0278-7393.32.5.1133.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162. doi:10.1037/0096-3445.131.2.147.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: the role of experience-based and theory based processes. *Journal of Experimental Psychology: General, 133*, 643–656.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224. doi:10.3758/BF03194055.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 609–622. doi:10.1037/0278-7393.32.3.609.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*, 787–794.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159–183. doi:10.1016/0030-5073(77)90001-0.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113*, 123–132. doi:10.1016/S0001-6918(03)00023-4.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159–163. doi:10.1111/j.1467-8721.2009.01628.x.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174–179. doi:10.3758/PBR.15.1.174.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 851–864. doi:10.1037/0278-7393.19.4.851.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19*, 338–368. doi:10.1016/S0022-5371(80)90266-2.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615–625. doi:10.1037/a0013684.

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

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

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: an investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*, 750–763. doi:10.3758/s13421-010-0063-y.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect. *Memory & Cognition, 40*, 703–716. doi:10.3758/s13421-011-0180-2.