Why Does Excellent Monitoring Accuracy Not Always Produce Gains In Memory Performance?

John Dunlosky[1], Michael L. Mueller[1], Kayla Morehead[1], Sarah K. Tauber[2], Keith W. Thiede[3], & Janet Metcalfe[4]

[1]*Kent State University,* [2]*Texas Christian University*, [3]*Boise State University,* [4]*Columbia University*

Header: When Monitoring Matters

Send correspondence to: John Dunlosky

Kent State University

Department of Psychology

Kent, OH 44242

jdunlosk@kent.edu

330-672-2207 (ofc); 330-672-3786 (fax)

Abstract

Does excellent monitoring of learning support improvements in subsequent relearning? Although some studies answer this question affirmatively, others have suggested that excellent monitoring may not matter. Accordingly, we address the question, When will highly accurate monitoring judgments benefit restudy? According to the *contingent-efficacy* hypothesis, excellent monitoring accuracy will not benefit learning (a) when restudy itself produces only small learning gains for items that were restudied, (b) when few (or most) of the items have been learned prior to restudy, and (c) when learners use their accurate judgments inappropriately for making restudy selections. Under these circumstances, the *contingent-efficacy* hypothesis predicts that restudy will be suboptimal, whereas under more ideal conditions (e.g., learning gains are high during restudy), excellent monitoring is expected to enhance restudy efficacy. By confirming these predictions across three experiments, the current research reconciles the prior discrepancies and reveals when excellent monitoring will matter for effectively guiding restudy.

Words: 149

Characters with spaces: 58,994

Key words: monitoring; self-regulation; learning; monitoring accuracy; metamemory

Why Does Excellent Monitoring Accuracy Not Always Produce Gains in Performance?

Does knowing yourself really matter? If a musician knows when he is performing well versus when he is struggling, will he be able to use this self knowledge – or ability to monitor his performance – to effectively practice and enhance his performance overall? If a student in biology can accurately judge which introductory concepts she has learned well versus those she has not learned well, will she be able to effectively regulate her study and improve her learning? The answer to these questions, from both intuition and theory, is "yes." Concerning the latter, theories of metacognitive control tightly link the accuracy of monitoring with the effectiveness of control, because monitoring presumably is used in the service of control (e.g., Dunlosky & Ariel, 2011; Koriat & Goldsmith, 1996; Metcalfe, 2009; Metcalfe & Finn, 2008; Nelson & Narens, 1990; Winne & Hadwin, 1998). In particular, theories of self-regulated learning assume that people use on-going monitoring to make decisions on how to proceed to achieve a learning goal. Thus, if a learner's monitoring is used to decide how to regulate subsequent behavior to meet a goal, then regulation is bound to be ineffective if monitoring is poor. For instance, if a student inaccurately judges what she has learned, then using monitoring to regulate learning would not be expected to benefit the efficacy of restudy.

Even though self-regulation theory predicts that monitoring accuracy can benefit self-regulation and performance, it does not necessitate that even excellent monitoring accuracy will always benefit performance. Consider again the student who is studying biology concepts – she may be perfectly accurate at monitoring which concepts she knows best (and those she knows the least well), yet (a) she may not have the time to learn those concepts that she has accurately judged as not-yet learned, or (b) she may make poor decisions about how to use her monitoring

to regulate subsequent studying. These examples suggest that even highly accurate monitoring will not always matter and lead to questions with less intuitive answers, namely: When does knowing yourself matter? Or, what factors moderate the benefits of accurate monitoring on performance? To provide preliminary answers to these questions, we first compare two investigations that reported evidence relevant to estimating the impact of excellent monitoring accuracy on performance. The investigations (Kimball, Smith, & Muntean, 2012; Nelson, Dunlosky, Graf, & Narens, 1994) used nearly identical methods, but the outcomes from them led to contrasting conclusions. By reconciling their contrasting conclusions, a main goal of the present research is to reveal conditions in which highly accurate monitoring does (and does not) support effective restudy and overall performance gains.

Contrasting Conclusions about the Importance of Metacognitive Monitoring

Kimball et al. (2012) and Nelson et al. (1994) evaluated the importance of accurate monitoring to restudy efficacy using a method called the honor-dishonor (HD) method by Kornell and Metcalfe (2006). To illustrate the HD method, imagine learners who are studying paired associates (e.g., dog – spoon), and that they can already recall half of the items on a cued recall test (i.e., dog - ?) when they judge their learning. After judging their learning, a computer program selects half of the items for restudy. Suppose the learners are perfectly accurate at judging which items have been relatively well learned versus poorly learned (e.g., judging that all recalled items have been better learned than the previously unrecalled items). The learner's judgments are then used to select those items that were judged as having not yet been learned. This allocation strategy is used by many students (Metcalfe & Kornell, 2005), and in the context of learning paired associates during a single study session, it is an appropriate way to use the judgments when students' goal is to learn all the items (for an exception, see Kornell &

Flanagan, 2014).[1] After selection, the learners restudy either the selected items or the items that were not selected for restudy. In the former case, the selections were *honored*, and in the latter case, the selections were *dishonored*. After restudy, a final criterion test across all items (whether restudied or not) is administered (for variants of this method, see Kornell & Metcalfe, 2006; Metcalfe & Finn, 2013). If highly accurate monitoring supports effective restudy decisions, then a key prediction is that final test performance will be higher when the restudy selections are honored than when dishonored.

Now consider details from both investigations. First, Kimball et al. (2012) had participants study word pairs and then make delayed judgments of learning (JOLs) that were cued by the stimulus alone. These JOLs are highly accurate at discriminating between well-learned items and those that have been less-well learned (for a meta-analysis, see Rhodes & Tauber, 2011). After making these delayed JOLs, the participants chose items for restudy, and those decisions were either honored or dishonored. An *honor-dishonor* effect can then be computed as follows: mean final performance (across all items) for the honor group minus mean final performance (across all items) for the dishonor group. We refer to positive values as an honor-dishonor *increment* (because honoring decisions improves performance) and to negative values as an honor-dishonor *decrement* (because honoring decisions decreases performance). In contrast to the aforementioned prediction, however, when learners' used cue-only delayed JOLs to select items for restudy, honoring (vs. dishonoring) those decisions often resulted in a statistically insignificant honor-dishonor effect. Even when honoring decisions produced significant honor-dishonor increments, the increments were small. For instance, in Experiment

[1]A different decision rule may be more optimal in contexts where studying the easiest unlearned items first (vs. the more difficult unlearned items) is essential to learn the more difficult items (see the region-of-proximal learning heuristic, Metcalfe & Kornell, 2005). In the current context (where learning one item is not relevant to learning others), restudying all currently unlearned items is assumed to be effective (to foreshadow, the outcomes from the current experiments confirm this assumption).

1a, only a 10% improvement was observed in final performance when learners' use of cue-only delayed JOLs was honored (vs dishonored). The observed HD increments in the other relevant experiments (Experiments 1b-1c and 2a-2c) were even smaller and ranged from about 3% - 8%. As concluded by Kimball et al. (2012), "although a minority of our findings… appeared to provide empirical support for the prediction [that highly accurate monitoring can be used to make effective restudy decisions], effect sizes were quite small, much smaller than might be expected" (p. 942). Thus, a major conclusion of this paper was that excellent monitoring accuracy has a minimal impact on the efficacy of restudy, which was offered as a broad generalization.

Second, Nelson et al. (1994) used a version of the same method employed by Kimball et al. (2012). Participants studied paired associates, made highly accurate delayed JOLs, and then the computer selected items for restudy. For the honor condition, the computer selected half the items for restudy that had the lowest JOLs (and a group of participants who self selected also chose items given the lowest JOLs, as expected). For the dishonor condition, half of the items that had the highest JOLs were chosen for restudy. In contrast to Kimball et al. (2012), a substantial HD increment was found, with final performance after restudy being 22% greater for those who restudied the items given low JOLs than for those who restudied those given higher JOLs (estimated from Figure 1, Nelson et al., 1994). They concluded that "people's metacognitive monitoring of idiosyncratic knowledge has functional utility in causal chains for learning" (p. 207).

These outcomes and resulting conclusions seem to be inconsistent and mutually incompatible. In seeking a resolution, we propose the *contingent-efficacy hypothesis*, which

(described further below) predicts that several factors will moderate the benefits of excellent monitoring accuracy for supporting effective restudy and producing larger HD increments.

Research Overview

In the present experiments, participants studied paired associates during multiple study-test trials. They then attempted cued recall for each item and made a JOL immediately afterwards – a procedure akin to delayed JOLs that yields high levels of discrimination between well-learned versus poorly learned items (Nelson, Narens, & Dunlosky, 2004). Next, (a) the computer used each participant's JOLs to select either the items given the lowest JOLs (honor condition) or highest JOLs (dishonor condition) for restudy (Experiments 1, 2, & 3, as in Nelson et al., 1994) or (b) the learners themselves selected half the items for restudy and those choices were either honored or dishonored (Experiment 3, as in Kimball et al., 2012). Next, the participants restudied the target items, and then a final criterion test across all items was administered. To evaluate the contingent-efficacy hypothesis, we explore the degree to which a focal moderates the degree to which accurate monitoring will support HD increments. Prior to each experiment, we discuss the rationale for why each factor could partly moderate the size (and direction) of HD effects and in turn explain the apparent inconsistency between the prior outcomes. By empirically resolving the inconsistent conclusions, the present experiments seek to reveal when excellent monitoring accuracy will improve restudy efficacy and when it will not.

Experiment 1

In Experiment 1, we investigated whether learning gains during restudy moderate the benefits of monitoring accuracy. Learning gain here refers to the observed change in performance for restudied items that were not correctly recalled prior to restudy – that is, the degree to which restudy helps participants learn those items that they had not yet learned.

According to the contingent-efficacy hypothesis, why restudy efficacy (as measured by the HD increment) is expected to be contingent on learnings gains is straightforward. If learning gains during restudy are minimal, then even excellent monitoring may have a minor impact on restudy efficacy and result in little-to-no HD increment. That is, when the restudy trials are largely inert (e.g., if restudy time is too brief or restudy occurs during a single, massed trial), it may not matter whether the highly accurate judgments are used to isolate unlearned items for restudy (honor condition). In this case, restudy is not actually helping people learn what they accurately judged that they had not learned. By contrast, if learning gains are high after restudy and previously unlearned items are subsequently learned and recalled, then monitoring accuracy may matter a lot and lead to a significant increase in the HD increment.

To evaluate whether this factor – the degree to which restudy boosts performance for previously unlearned items – can partly explain the different outcomes from the aforementioned investigations, we estimated learning gains from them. In Kimball et al. (2012), estimated performance after restudy for those items that were initially not learned was between .25 and .35, whereas in Nelson et al. (1994), the learning gain was close to .95. Thus, in the former study, even when learner's decisions were honored, they gained relatively little from restudy, whereas in the latter, the gains from restudy were substantial. Of course, these studies differed in numerous ways, so the main goal of Experiment 1 was to experimentally evaluate whether differential learning gains during restudy moderates the size of the HD increment.

As noted above, all participants studied paired associates and made delayed JOLs. During the subsequent restudy phase, learning gains were manipulated as follows. For the low-gain group, each item slated for restudy was presented individually (for 1 s) for a single restudy-test trial. For the high-gain group, each item slated for restudy was presented individually (for 6

s) across 3 restudy-test trials. After the restudy phase, a final paired-associate test (e.g., dog - ?) was administered for all items. According to the contingent-efficacy hypothesis, a relatively large HD increment will occur for participants in the high-gain group (consistent with Nelson et al., 1994), whereas a relatively small (if any) HD increment will occur for participants in the low-gain group (consistent with Kimball et al., 2012).

Method

Design, Participants, and Materials

A 2 (Honor-dishonor group: honor vs. dishonor) × 2 (Learning gain: low vs. high) between-participants factorial design was used. One hundred and twenty undergraduates were recruited from Kent State University to fulfill a partial course requirement and were randomly assigned to one of four groups. The sample size (i.e., about 30 per group) in all experiments was chosen a priori based on prior research that has demonstrated significant HD effects; more participants were included in Experiment 2, because more students than needed had registered before access to registration had been closed (and none were turned away). In Experiment 1, nine participants had no variability in their JOLs during the initial study phase, so they were excluded from analyses because a lack of variability in JOLs would result in randomly selecting items for restudy. Thus, 26, 31, 26, and 28 participants ($N = 111$) were in the honor low, honor high, dishonor low, and dishonor high groups, respectively. Items consisted of forty unrelated pairs of concrete words (from Hertzog et al., 2002). Items were presented for each participant in a random order during each phase of the experiment (i.e., study, cued recall, and restudy). JOLs and cued recall were self-paced by the participants.

Procedure

The experiment consisted of three phases: initial study, restudy, and final recall. During the initial-study phase, which was the same for all four groups, items were presented one at a time for 2 s in the center of the computer screen. After studying all of the items, participants were given a cued recall test in which they were shown the first word of an item and were asked to recall the second (e.g., dog - ?). Following each recall attempt, participants made a JOL on a scale from 0 to 100% using the prompt "How likely are you to remember the correct response if you were shown the first word in about 30 seconds?" In the present context, JOLs needed to accurately discriminate between items that were more-well vs. less-well learned at the end of the initial-learning phase. Thus, although we used a JOL prompt that involved predicting future performance because such prompts were used by Kimball et al. (2012) and Nelson et al. (1994), we expected that the brief 30-sec interval would encourage participants to consider how well they had just learned the items (and not the degree to which they would forget them over a long retention interval). Consistent with this expectation, participants' JOLs highly discriminated between items that were well-learned versus less-well learned during the initial study phase (see Result, under Preliminary Analyses below). After attempting to recall and judging all of the items, the mean percentage of correct recall was computed (by the data-collection program), and participants advanced to the restudy portion of the experiment if and only if they had correctly recalled 40% or more of the items on that test trial. Participants who did not meet the recall criterion repeated the initial-study phase until they achieved 40% on a single test phase. This procedure was used to help ensure that all participants began the restudy phase being able to recall about half of the items.

During the restudy phase, both manipulations occurred. For the honor group, each participant restudied 20 of the items that he or she had assigned the 20 lowest JOLs during the

final trial of the initial-study phase. For the dishonor group, the 20 items with the highest JOLs were selected for restudy. Participants were not told how the words were selected for restudy. The other manipulation involved the duration of study and length of the restudy phase. For the low-gain group, participants were shown the 20 items for 1 s each (during a single trial), and for the high-gain group, participants were shown the items for 6 s each (during each of the three trials). Following the presentation of the restudy lists, participants were given another cued recall test over all 40 items. Moreover, to ensure different learning gains, participants in the low-gain group had only a single restudy-test trial, whereas participants in the high-gain group had three restudy-test trials. After the restudy phase, participants were presented with instructions for the final test (which remained on the screen for at least 30 s), and then a final cued recall test was administered for all 40 items.

Results

Preliminary Analyses

We first present analysis of measures relevant to interpreting any possible differences in the HD increments across groups. In Table 1 (see Electronic Supplementary Material 1, available from PsychArchives at {link to be placed here}), we present the number of trials that participants required to reach criterion during the initial-learning phase ("Trials to Criterion"), recall performance at the end of the initial-study phase ("Initial Performance"), mean JOLs, JOL accuracy, and learning gains, along with relevant inferential statistics. The Supplementary Material includes descriptive values and inferential analyses for these measures, so we focus just on the most relevant here. In particular, the JOLs were expected to demonstrate high levels of accuracy for discriminating between items that had (vs. had not) been recalled by the final test trial of the initial-study phase (Rhodes & Tauber, 2011). Note that the accuracy of JOLs is

typically estimated by the correspondence between JOLs and *final* performance (given JOLs are predictions of future recall), but the intervening restudy trials would inappropriately decrease the actual accuracy of these JOLs. However, given that recall fluctuations from delayed JOLs to final recall (without intervening restudy) would be minimal in the present context (see Nelson et al., 2004), the current analysis involving initial recall and the delayed JOLs should provide a close estimate to predictive accuracy. Even so, in the present context, the functional utility of monitoring judgments is to discriminate between what had initially been learned well versus less well prior to restudy, so the current analyses involving performance prior to JOLs in the calculation of JOL accuracy is arguably the most relevant to understanding the efficacy of subsequent restudy decisions. As shown in Table 1, as with the typical predictive accuracy of delayed JOLs (Rhodes & Tauber, 2011), the mean intra-individual gamma correlations between recall on the final trial of the initial phase and JOLs was high (Median for all groups was .97 or higher).

Finally, as a manipulation check, we computed learning gains for the two groups. Learning gain refers to the increase in performance from before to after restudy for those items that had been restudied but had not been correctly recalled after the initial-study phase. Thus, for each participant, we computed the probability of recall on the final test for those items that (a) had not been correctly recalled during the initial-study phase and (b) that were presented during the restudy phase. These values are presented in Table 1, and as expected, the learning gains were greater for the high-gain group (*M*s = .93 and .97 for the honor and dishonor groups, respectively) than for the low-gain group (*M*s = .22 and .35), which was critical for testing the aforementioned predictions. A main effect occurred that favored the dishonor over the honor group, which likely arose because the learning-gain measure is conditionalized on the restudy

selections. In particular, for the dishonor group, the items given the highest JOLs would be selected for restudy, so participants in this group would be restudying easier unlearned items (as compared to the larger set of more difficult unlearned items that had been given low JOLs that would be slated for restudy by the honor groups).

Honor-dishonor Effects

Final recall performance is presented in Figure 1. An HD increment is represented by an increase in performance for the honor group above the dishonor group, and the results are clear: the HD increment was larger for the high-gain group (24%, Cohen's $d = 2.35$) than for the low-gain group (3%, $d = 0.21$). Consistent with the aforementioned observation, a 2 (Honor-dishonor groups) × 2 (Learning-gain groups) ANOVA revealed a significant interaction, $F(1,106) = 24.27$, $MSE = .01$, $p < .01$, $\eta_p^2 = .19$). In this case, the HD increment for the high-gain group was significantly greater than zero, $t(57) = 9.0$, $p < .01$, whereas the HD increment for the low-gain group was not, $t(50) = 0.76$. Although less relevant to the present goals, the main effects for the honor-dishonor and the learning-gain manipulations were significant, $F(1,107)s = 38.8$ and 45.9, respectively, with both $MSEs = .01$, $ps < .01$, and $\eta_p^2 s > .25$, and do not qualify the aforementioned interaction.

Discussion

The outcomes presented in Figure 1 establish that excellent monitoring can matter a lot, but it does not always matter. More generally, even when people can accurately identify what they currently can recall (vs. what they cannot recall), if they do not have the opportunity to learn items that they had judged as less well known and had selected for restudy, their accurate monitoring will not help them (Kornell & Metcalfe, 2006). In the present case, the difficulties arose because students in the low-gain group were given limited time to restudy the items that

they (accurately) judged to be less well learned, whereas those in the high-gain group had the opportunity to learn previously unlearned items. These outcomes from Experiment 1 can at least partially explain the apparently conflicting conclusions from Kimball et al. (2011) and Nelson et al. (1994).

Experiment 2

Factors other than learning rate may be expected to limit the HD increment and to moderate the benefits of accurate monitoring. In particular, the size of the HD increment may be contingent upon the initial level of performance prior to restudy. It may be especially important when researchers adopt versions of the honor-dishonor method to explore the degree to which learners' choices lead to relatively effective self regulation (e.g., Hanczakowski, Zawadzka, & Cockcroft-McKay, 2014; Kimball et al., 2012; Kornell & Metcalfe, 2006; Metcalfe & Finn, 2013; Nelson et al., 1994; Son, 2010; Tullis & Benjamin, 2012). To understand this possible contingency, it is useful to consider extremes. When learners recall *none* of the items prior to selection, even perfectly accurate monitoring would be expected to yield a minimal HD increment, because regardless of whether choices are honored or dishonored, all selected items would need to be learned. At the other extreme, if all items had been correctly recalled prior to restudy, then honoring (vs. dishonoring) choices is not expected to benefit restudy, at least with respect to a relatively immediate test that would minimize the impact of forgetting. Based on this rationale, it seems reasonable that excellent monitoring accuracy will benefit restudy efficacy more when learners initially have learned some, but not all, of the items. When initial performance prior to making JOLs is nearer the middle of the scale, learners restudying the items given the higher judgments (i.e., the dishonor group in Experiment 1) would be largely restudying those items they could already recall and hence would not be expected to gain much

from restudy; by contrast, learners restudying the items given lower judgments (i.e., the honor group) would largely be restudying items that they could not yet recall and hence could benefit from restudy (i.e., assuming the relearning rate is sufficiently high). According to this rationale, the HD increment will be the largest when initial performance (prior to restudy) is near fifty percent (i.e., middle of the scale).

Differences in initial performance (prior to restudy) may also explain the apparently conflicting conclusions from Nelson et al. (1994) and Kimball et al. (2012). Namely, the initial level of performance prior to restudy was near .50 in the former investigation, but it was closer to .25 in the latter one. Perhaps the small (and sometimes non-significant) HD increments reported by Kimball et al. (2012) partly arose because learners could not benefit from excellent monitoring accuracy because the majority of items needed to be restudied. To evaluate this prediction, we manipulated the initial level of performance (prior to restudy) in Experiment 2. Some participants continued the initial-study phase until they could correctly recall at least 40% of the items (which was expected to produce around .50 recall after the initial-study phase – as in Experiment 1, see Table 1, "Initial Performance," Electronic Supplementary Materials). Other participants only received one trial during this phase, so that their initial level of learning would be significantly lower. All participants then studied those items slated for restudy (either honored or dishonored) multiple times, so as to obtain a high learning gain (which Experiment 1 established was necessary to produce an honor-dishonor increment). According to the contingent-efficacy hypothesis, the HD increment will be larger for the group who had an initial level of performance closer to 50%.

Method

Design, Participants, and Materials

A 2 (Honor-dishonor groups) × 2 (Initial learning: low vs. mid) between-participants factorial design was used. One hundred fifty-two undergraduates were recruited from Texas Christian University to fulfill a partial course requirement and were randomly assigned to one of four groups: Honor Low, Honor Mid, Dishonor Low, and Dishonor Mid. Seven participants had no variability in their JOLs during the initial-study phase, so they were excluded from analyses as in Experiment 1. Thus, 38, 35, 38, and 34 participants (N = 145) were in the honor low, honor mid, dishonor low, and dishonor mid groups, respectively. The same items and randomized order of presentation were used as in Experiment 1. JOLs and cued recall were self-paced by the participants.

Procedure

The procedure was identical to those used for the high-gain groups in Experiment 1 (i.e., during restudy, items were presented for 6 s each across 3 restudy trials), except with regard to the differences between groups in the initial learning phase. For participants in the mid-initial learning groups, study-test trials continued during the initial-study phase until 40% or more of the items were correctly recalled on a single trial (as in Experiment 1). For participants in the low-initial learning groups, the items were presented for only a single study-test trial (at a rate of 1 s per item).

Results

Preliminary Analyses

We begin by presenting analyses of measures relevant to interpreting any possible differences in the HD increments across groups (see Table 2, Electronic Supplementary Materials). First, performance after the initial-study phase was near the middle of the scale for the mid-initial learning groups (Ms = .58 and .57 for the honor and dishonor groups,

respectively) and significantly greater for these groups than for the low-initial learning groups (*M*s = .20 and .21, respectively). Second, JOL accuracy was again high (Median for all four groups = 1.0), indicating that participants' JOLs discriminated between items that had been learned well (vs. less well) during the initial-study phase.

Honor-dishonor Effects

Final recall performance is presented in Figure 2. Although both groups did demonstrate a significant HD increment, this effect was larger for the mid-initial learning group (27%, Cohen's $d = 1.77$) than for the low-initial learning group (13%, $d = 0.89$). Consistent with these observations, a 2 (Honor-dishonor group) × 2 (Initial learning groups) ANOVA revealed a significant interaction, $F(1,141) = 9.0$, $MSE = .02$, $p < .01$, $\eta_p^2 = .06$. The main effects for the honor-dishonor and the initial-learning manipulations were significant, $F(1,141)s = 66.5$ and 58.6, respectively, with both *MSE*s = .02, *p*s < .001, and $\eta_p^2 s > .25$.

Experiment 3

Experiments 1 and 2 empirically demonstrated that even when monitoring accuracy is excellent, restudy efficacy (as measured using the HD method) is contingent upon (a) the level of learning gains after restudy and (b) the level of initial performance prior to restudy. In these experiments, we used a computer algorithm to simulate effective restudy decisions (i.e., select the items given the lowest JOLs) and ineffective restudy decisions (select the items given the highest JOLs) in this context (i.e., single-session, paired-associate learning). We consider these as honor and dishonor conditions (respectively), because prior research has shown that college students typically choose items given the lowest JOLs for restudy (Metcalfe & Kornell, 2005).

Despite the prior evidence, however, at least three recent studies have shown that some college students do the opposite (Morehead & Dunlosky, 2020; Morehead, Dunlosky, & Foster,

2017; Tullis & Benjamin, 2011); that is, they are more likely to choose to select to restudy materials with which they had given higher than lower JOLs. According to the contingent-efficacy hypothesis, selection strategies will moderate restudy efficacy (even when monitoring is highly accurate). In particular, this selection strategy for restudying items (i.e., select items judged as more well learned) is likely ineffective in the present context, given that during a single learning session (which are typical for research in this domain), items that are recalled after an initial study trial will be recalled later without further restudy, whereas items that are not recalled after an initial study will not subsequently be recalled (i.e., reminiscence is near zero). Thus, if these learners' monitoring accuracy is near perfect, they may show a *reversed* HD effect; that is, for this subset of students who choose the items with higher JOLs, *dis*honoring their selections may boost later performance (because they would be studying those items that could benefit from further study), whereas honoring their selections may be relatively inert (because restudying already known items will not enhance learning gains).

Such individual differences in decisions about which items to select for restudy could also partly explain the apparent discrepancies from the prior studies. In particular, whereas Nelson et al. (1994) had a computer use an effective strategy in using JOLs to select items for restudy (i.e., choose items given the lowest JOLs), Kimball et al. (2012) had participants select which items to restudy. Perhaps some of these participants made ineffective restudy decisions, which would reduce the overall benefits of accurate monitoring for improving restudy efficacy and further constrain the HD increment.

To evaluate these predictions, we used the HD method from Experiments 1 and 2, with one addition: some participants selected items for restudy (and their selections were either honored or dishonored). Given the outcomes of Experiments 1 and 2, it is evident that for an

experiment to demonstrate an HD increment, the initial performance prior to restudy should be near the middle of the scale (about 50%) and the learning rate during restudy should be high. Accordingly, we used the procedures from Experiments 1 and 2 that had produced initial performance near 50% and a high learning rate. The contingent-efficacy hypothesis predicts that the HD increment will be significantly greater for those participants who select to restudy items given low JOLs than for those participants who select to restudy items given high JOLs. We also included groups (honor vs. dishonor) in which the computer selected items for restudy as in Experiments 1 and 2. If outcomes from these groups replicate the prior HD increments (Figures 1 & 2), then any differences in the self-selection groups could be attributed to their selections per se (and not to differences in the population of participant samples for Experiment 3).

Method

Design, Participants, and Materials

A 2 (Honor-dishonor groups) × 2 (Restudy selection: computer selection vs. self selection) between-participants factorial design was used. One hundred and twenty nine students from Kent State University participated to fulfill a partial course requirement and were randomly assigned to the groups. Sample sizes were 43, 43, 42, and 41 for the self-selection honor and dishonor groups and the computer-selection honor and dishonor groups, respective. Materials from Experiment 1 were used.

Procedure

The procedure was identical to Experiments 1 and 2 except with regard to the differences between groups. As per the mid-initial learning groups, study-test trials continued during the initial-study phase until 40% or more of the items were correctly recalled on a single trial (as in Experiment 1). After the initial learning trials, items were selected for restudy. For the

computer-selection group (replication of prior experiments), the computer used each participant's JOLs to select items for restudy. For the self-selection group, participants selected as many items for restudy as they wanted. To equate for the number of items restudied for those in the computer-selection group (in which 20 items were selected for restudy) and those in the self-selection group, we adopted the procedure from Nelson et al. (1994): If participants selected fewer than 20 (or more than 20) items, then enough items were randomly selected to either be added (from the unselected items if fewer than 20 items had been selected) or removed (from those selected if more than 20 were selected). During restudy, items were presented for 3 study trials at a 6 s presentation rate per item (as per the high-gain groups in Experiment 1) to ensure a high level of learning gain after restudy.

Results

Preliminary Analyses

A critical prediction concerns whether participants who use different selection strategies will demonstrate different magnitudes of an HD effect. But, did individual differences in selection strategies occur in this participant sample? To answer this question, we correlated each participant's JOLs with their restudy selections. The mean across intra-individual gamma correlations was -.02 (*SEM* = .14) for the honor group and .04 (*SEM* = .12) for the dishonor group, $t(72) = 0.34$. The frequency distribution (combined for both groups) of these correlations is presented in a figure in the Electronic Supplementary Materials, which highlights the variability in restudy selections: many participants did choose to restudy items given lower JOLs (resulting in a negative correlation), but a meaningful subset of participants used the opposite selection strategy. For subsequent analyses involving comparisons between participants who tended to select items with lower vs. higher JOLs, we split participants into two groups: those

with a negative correlation (demonstrating a tendency for selecting items with lower JOLs) and those with a positive correlation (demonstrating a tendency for selecting items with higher JOLs). For brevity, we will refer to the former as the low-JOL group (*n*s = 18 and 18, for the honor and dishonor groups, respectively) and the latter as the high-JOL group (*n*s = 17 and 20, for the honor and dishonor groups, respectively).[2]

As in Experiments 1 and 2, we also present analyses of secondary measures (see Table 3, Electronic Supplementary Materials). In general, outcomes were consistent with the design parameters chosen for this experiment: Performance after the initial-study phase was near the middle of the scale for all groups, JOL accuracy was again high (median accuracy was greater than .94 for all groups), and learning gains were also near 1.0 (*Median* = 1.0 for all groups).

Honor-dishonor Effects

To establish the size of the HD effects, we present final recall performance for the computer-selection and self-selection groups in Figure 3. The significant interaction, $F(1,165) = 29.8$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .15$, indicates that the HD increment is contingent on using a computer algorithm to choose items that were judged as less-well learned versus allowing participants to choose which items to restudy. The computer-selection groups demonstrated a large HD increment, $t(81) = 9.14$, $d = 2.0$, whereas the HD increment was minimal for those who selected items for restudy, $t(84) = 0.22$, $d = .05$. The main effects for the honor-dishonor manipulation was significant, $F(1,165) = 26.0$, $MSE = 0.02$, $p < .001$, and $\eta_p^2 = .14$, whereas the main effect of selection group was not, $F(1,165) = 0.01$, $MSE = 0.02$, $p = .99$, $\eta_p^2 = .00$.

More important, the non-significant HD increment for the self-selection group may be hiding an embedded interaction that is relevant to whether participants tended to select for restudy items with lower versus higher JOLs. To evaluate this possibility, we conditionalized

[2]Participants who had indeterminate correlations or had a correlation of zero were not included in these analyses.

final recall performance as a function of how participants selected items. As shown in Figure 4, the HD effects were dramatically different for the two groups. The main effect for participant selection (choose less well vs. more well learned items) was significant, F(1,69) = 4.7, MSE = 0.009, p = .03, whereas the main effect of honor-dishonor manipulation was not, F(1,69) = 0.52, MSE = 0.009, p = .47, η_p^2 = .01. The significant interaction, η_p^2 = .06, F(1,69) = 125.6, MSE = 0.009, η_p^2 = .65, supports the pattern evident in Figure 4: Whereas those who selected the less-well learned items for restudy (left two bars) showed a significant HD increment, t(34) = 9.62, d = 3.21, those who selected the more well-learned items showed a significant HD *decrement,* t(35) = 6.71, d = 2.21. For the latter participants, although using their highly accurate JOLs could have benefited study (as per the computer-selection group), their ineffective restudy selections undermined their learning. This HD *decrement* establishes that highly accurate monitoring is not sufficient for improving learners' restudy efficacy. Of course, this conclusion is relevant to the parameters of the current experiment – a short retention interval, a final cued recall test, and so forth – that were chosen to closely match those used by Kimball et al. (2012) and Nelson et al. (1994). Under other conditions, the selection rule to choose the *easiest* items (i.e., selecting items given the higher JOLs) may be an effective strategy; for instance, perhaps with a longer retention interval, final performance may be boosted if the more well-learned items are restudied (see Kornell & Flanagan, 2014). A challenge for future research will be to more fully explore the various parameters that could influence the size (and direction) of HD effects.

General Discussion

Does knowing yourself matter? In the present research, we answered this question empirically in the context of memory monitoring, which is arguably the most widely investigated process-oriented component of metacognition. In this case, intuition suggests that the answer is

"yes", but evidence from Kimball et al. (2012) put such intuition into doubt and supported the conclusion that excellent judgment accuracy does not improve the efficacy of restudy decisions – that knowing yourself in this domain does not matter much. The present evidence both replicated their findings, but as important, it confirmed the contingent-efficacy hypothesis. In doing so, the present research revealed why Kimball et al. (2012) concluded that excellent monitoring did not improve restudy efficacy and provided a deeper understanding about when accurate monitoring will matter. Namely, at least three factors – performance prior to restudy, the learning gains during restudy, and how people use JOLs to make restudy selections – will influence the degree to which accurate monitoring can lead to performance gains during restudy.

Understanding the Mixed Evidence About Monitoring-Performance Relations

To briefly review the mixed evidence, two investigations motivated the present research because they used the functionally identical HD method. In particular, Kimball et al. (2012) posed the question "Does delaying judgments of learning really improve the efficacy of study decisions?", and, based on their evidence, answered it with "Not so much." By contrast, Nelson et al.'s (1994) evidence led them to conclude that monitoring can play a functional role in learning. Although these conclusions can be viewed as contradictory, evidence from the present experiments indicate they can be reconciled. That is, in both cases, accurate monitoring could have improved the efficacy of restudy decisions. More specifically, the parameters of the experiments in Kimball et al. (2012) were set at levels so that learners' could benefit little (if any) from using their accurate monitoring to regulate their restudy. We estimated that the learning gains fell between .25 and .40 across their experiments and that initial performance prior to selection was somewhere between .28 and .40. As demonstrated in the present Experiments 1 and 2, even when participants' monitoring was highly accurate, using their judgments did not

benefit learning (much) under these conditions. Moreover, participants in Kimball et al. (2012) selected which items to restudy, and the mean correlations between JOLs and restudy decisions (for the delayed JOL groups relevant to the current research) ranged from -.44 to -.90. These mean values are less than -1.0 and suggest that some participants were not using the most effective reselection strategy; that is, they appeared to be selecting some items that were judged as more well learned. As shown in Experiment 3, participants who self selected showed a reduced HD increment, and the subset who tended to select items judged as more well learned showed an HD *decrement*.

A major conclusion from the current research is that such small (or even reversed) HD increments do not reflect the impotency of accurate monitoring for supporting learning gains. Instead, they reflect conditions under which the cognitive system could not benefit from the use of accurate monitoring because of limits in memory itself (Experiments 1 and 2) or in metacognitive control (Experiment 3). By contrast, under the present experimental conditions, when participants restudy those items given the lower JOLs and learning gains are high, excellent monitoring accuracy matters a lot, as demonstrated by significant HD increments.

The present analyses also provide alternative interpretations for outcomes from two other investigations, and in doing so, they indicate avenues for future research. First, consider Begg, Martin, and Needham (1992), who concluded that "memory monitoring does not make a valuable contribution to memory" (p. 212). Based on the present outcomes, an examination of their method indicates that memory monitoring could have at best made a minor contribution to learning. In particular, participants studied items and made cue-only delayed JOLs (as in the present experiments), but they were not allowed to use their judgments to select which items to restudy, and they were given only a single opportunity to restudy the items. The learning gain

after restudy was low (i.e., estimated at or below .20 in both of their experiments; Begg et al., 1992). Thus, based on the present rationale and outcomes, participants could not have benefited from their accurate monitoring, and this failure apparently arose from both deficits in the metacognitive system (because the task would constrain monitoring-guided control of learning) and in the memory system (because the massed restudy yielded little learning gain). Put differently, the method used by Begg et al. (1992) ensured that accurate memory monitoring could not make a valuable (or any) contribution to learning.

Second, Tullis and Benjamin (2012) explored the degree to which aging in adulthood influenced the effectiveness of self regulation. Their innovation was in estimating the degree to which ineffective control of study decisions could contribute to age-related differences in learning by using the honor-dishonor method (for another approach, see Krueger, 2012). Based on a significant HD effect for younger adults and a lack of an HD effect for older adults, they concluded that "this reveals a dramatic failure in metacognitive control, in the absence of any obvious monitoring deficit, in older adults" (p. 743). The current research suggests some alternative possibilities. For instance, older adults tend to perform worse than younger adults after a single massed study trial, so perhaps initial performance prior to restudy put them at a disadvantage (e.g., further from the middle of the scale, as per the low-initial performance group, Figure 2, Experiment 2). Likewise, older adults may have enjoyed a smaller learning gain after restudy, which would also limit the HD increment (low-gain group, Figure 1, Experiment 1). Importantly, these alternatives implicate limited cognitive performance and not a dramatic failure in metacognitive control. Tullis and Benjamin (2012) provide some indirect evidence that these alternatives may not entirely account for the age-related deficit in the HD increment, so our point here is not that their interpretation is necessarily incorrect. Instead, given that the current

analysis reveals how differences in the HD effect could arise from multiple causes (some due to poor metacognition and others due to poor cognition), these possibilities should be evaluated by using methods where initial performance and learning gains can be directly estimated.

Will Enhancing the Accuracy of Monitoring Judgments Improve Restudy Efficacy?

The main issue addressed in the present research was whether excellent monitoring accuracy could support effective restudy, as indicated by a significant HD increment. Concerning the question of whether improving monitoring accuracy will also improve learning, few researchers have experimentally manipulated monitoring accuracy so as to evaluate whether higher levels of accuracy yield higher levels of performance after restudy (but see Thiede, Anderson, & Therriault, 2003). Kimball et al. (2012) did manipulate accuracy by having participants make either delayed JOLs cued by the stimulus alone (which were the focus of the present research) or by immediate JOLs (which support significantly lower levels of accuracy). The HD increments were low in all cases and not significantly greater for delayed JOLs than immediate JOLs. Such outcomes may lead one to conclude that improving accuracy may not matter much, but the present research indicates that the answer to the question is still unresolved. The limiation is that Kimball et al. (2012) used a method that produced low learning gains and low levels of initial performance before restudy, which the current research demonstrated produce small-to-no HD effects. Put differently, the methods that Kimball et al. (2012) used to evaluate whether increases in judgment accuracy will result in larger HD increments were not sensitive to revealing those effects. Under conditions where highly accurate JOLs would produce a substantial HD increment (as demonstrated in the current research), perhaps the HD increment would be larger when restudy decisions are based on delayed than immediate JOLs. That is, the question becomes whether HD increments are higher for delayed and immediate

JOLs under conditions (a) that could support significant HD increments (i.e., learners make effective restudy decisions, learning gains are high, et cetera) and (b) where all else is equal (i.e., restudy decisions, initial level of learning, learning gains) across groups except for the different levels of judgment accuracy.

Thus, whether improving JOL accuracy improves the HD increment is still an open question. Perhaps improvements in JOL accuracy will monotonically increase the HD increment, but other relationships are possible. For instance, a criterion level of JOL accuracy may be required to obtain maximal gains (in terms of the HD increment) from using JOLs to guide restudy. In this case, the relationship between JOL accuracy and the HD increment will be a step function. One extreme possibility is that a relatively low level of JOL accuracy would be needed to obtain a maximal HD increment – if so, then the modest levels of accuracy often supported by immediate JOLs (for a meta-analysis, see Rhodes & Tauber, 2011) may be all that is required to effectively guide learning and yield maximal HD increments under supportive conditions. Most important, the current research has revealed the conditions needed to estimate the form of this function – that is, learning gains must be maximized, learners must begin the restudy trial with mid-level performance, and they must restudy those items judged as least well learned. If these conditions are not met, then HD effects may artifactually constrained by the methd and the experiment would not be sensitive for detecting the potential differences in restudy efficacy that could arise from group differences in monitoring accuracy.

Closing Remarks

Some prior evidence indicated that knowing oneself may not matter much in the domain of metamemory, given that accurate self monitoring did not yield subsequent benefits to performance (e.g., Begg et al., 1992; Kimball et al., 2012). By supporting a contingent-efficacy

hypothesis about when excellent monitoring will matter, the present evidence indicates that this prior research examined conditions in which one would not expect excellent monitoring to improve the efficacy of restudy. These boundary conditions are important to discover and to explore, but they do not indicate that monitoring accuracy cannot improve people's restudy efficacy. Instead, they are symptomatic of an interactive metacognitive-cognitive system in which both metacognition (e.g., high JOL accuracy and appropriate restudy decisions) and cognition (e.g., memory performance prior to restudy and subsequent relearning gains) need to operate effectively for one to reap the benefits of accurate monitoring. So, does knowing yourself really matter? We suspect it can matter a lot, but regardless of the domain, it does not have to matter. Even when a musician accurately identifies difficulties while playing and even when a student accurately judges what has not been learned well, they will not enjoy the benefits of their accurate monitoring if they do not have the time or skills needed to improve their performance.

Electronic Supplementary Material

ESM 1. Tables 1-3, Distribution Figure for Experiment 3

References

Begg, I. M., Martin, L. A., & Needham, D. R. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology, 4*, 195-218.

Dunlosky, J & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed), *Psychology of Learning and Motivation*, *54*, 103-140.

Dunlosky, J. & Thiede, K. W. (1998) What makes people study more? An evaluation of factors that affect people's self-paced study and yield "labor-and-gain" effects. *Acta Psychologica, 98*, 37-56.

Hanczakowski, M., Zawadzka, K., & Cockcroft-McKay, C. (2014). Feeling of knowing and restudy choices. *Psychonomic Bulletin & Review, 21*, 1617-1622.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208-216.

Hertzog, C. Kidder, D., Powell-Moman, A., & Dunlosky, J. (2002). Monitoring associative learning: What determines the accuracy of metacognitive judgments. Psychology & Aging, *17*, 209-225.

Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory & Cognition, 38*, 923-954.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

Kornell, N., & Flanagan, K. E. (2014). Is focusing on unknown pairs while studying a beneficial long-term strategy? *Journal of Cognitive Psychology, 26*, 928-942.

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

Krueger, L. E. (2012). Age-related effects of study time allocation on memory performance in a verbal and a spatial task. *Educational Gerontology, 38*, 604-615.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159-163.

Metcalfe, J. & Finn. B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174-179.

Metcalfe, J. & Finn. B. (2013). Metacognition and control of study choice in children. *Metacognition & Learning, 8*, 19-46.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52,* 463-477.

Morehead, K. & Dunlosky, J. (2020). Do students make effective decisions when regulating their learning of categories? *Translational Issues in Psychological Science, 5*, 43-52.

Morehead, K., Dunlosky, J., & Foster, N. L. (2017). Do people use category-learning judgments to regulate learning of natural categories? *Memory & Cognition, 45*, 1253-1269.

Nelson, T. O., Dunlosky, J., Graf, E. A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. Psychological Science, *5*, 207-213.

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

Nelson, T. O., Narens, L. & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods, 9*, 54-69.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137,* 131-148. doi: 10.1037/a0021705

Son, L. K. (2010). Metacognitive control and the spacing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 255-262.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multi-trial learning. *Psychonomic Bulletin & Review*, *6*, 662-667.

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

Thiede, K.W., Redford, J.S., Wiley, J., & Griffin, T.D. (2012). Elementary school experience with comprehension testing may influence metacomprehension accuracy among 7th and 8th graders. *Journal of Educational Psychology, 104,* 554-564.

Tullis, J. G., & Benjamin, A. B. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review, 19*, 743-749.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 277-304). Hillsdale, NJ: Erlbaum.

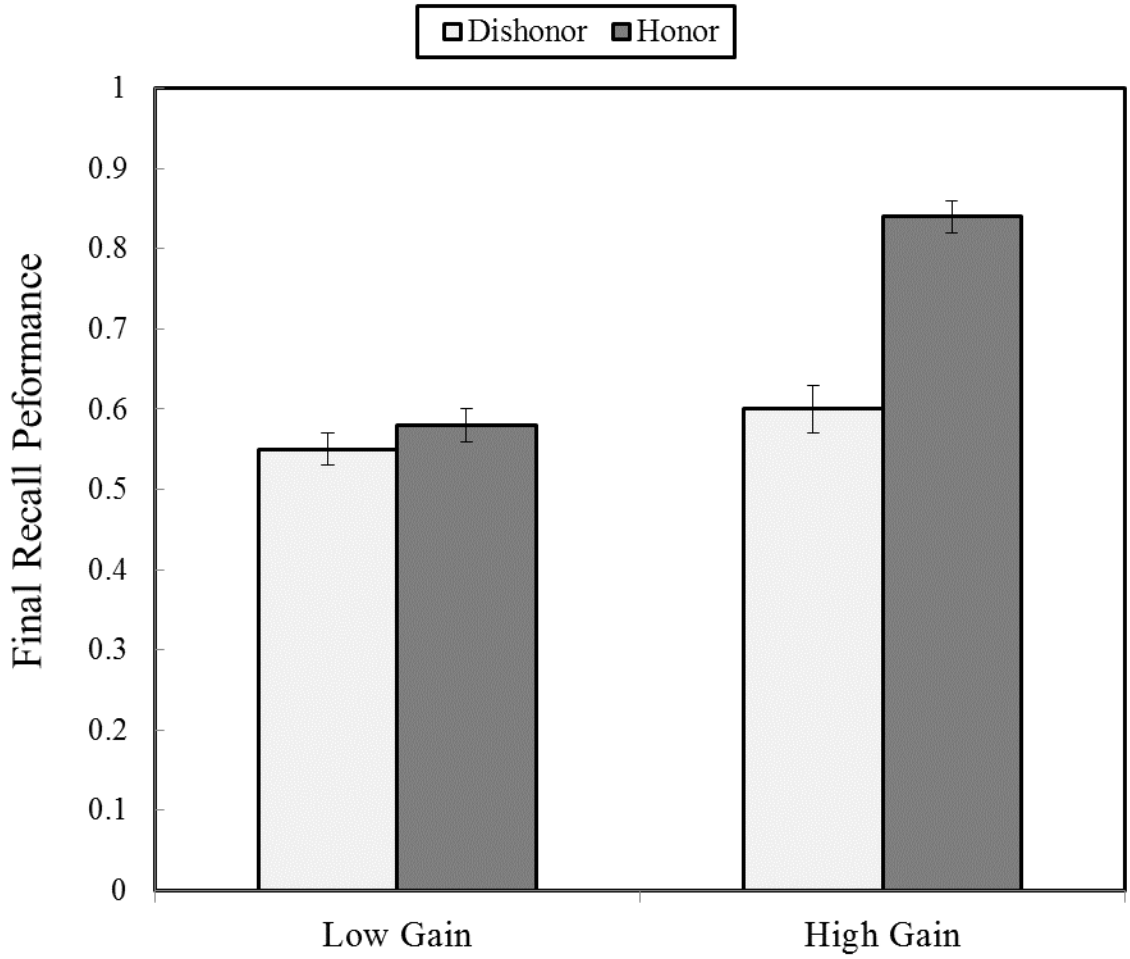

Figure 1. Final recall performance as a function of group for Experiment 1. Bars are standard errors of the mean.

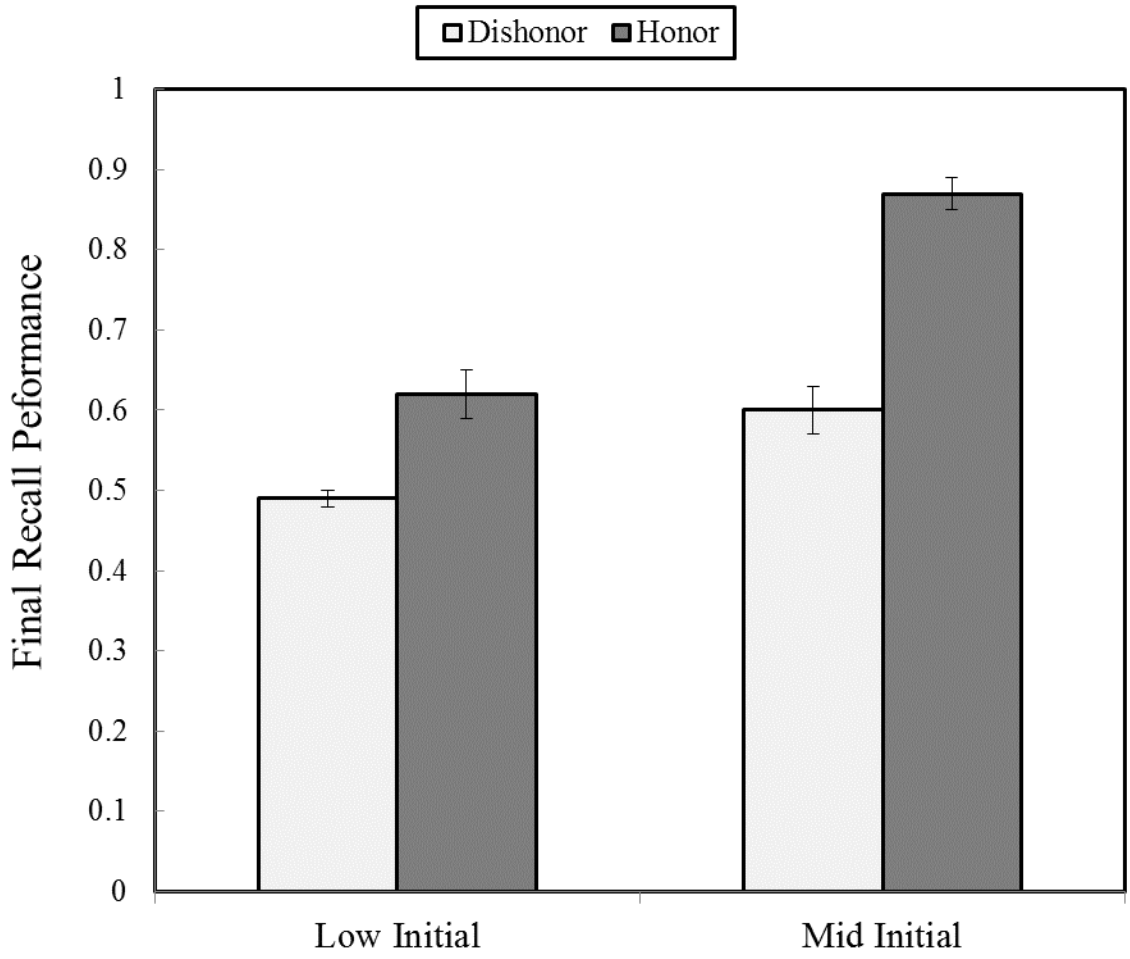

Figure 2. Final recall performance as a function of group for Experiment 2. Bars are standard errors of the mean.

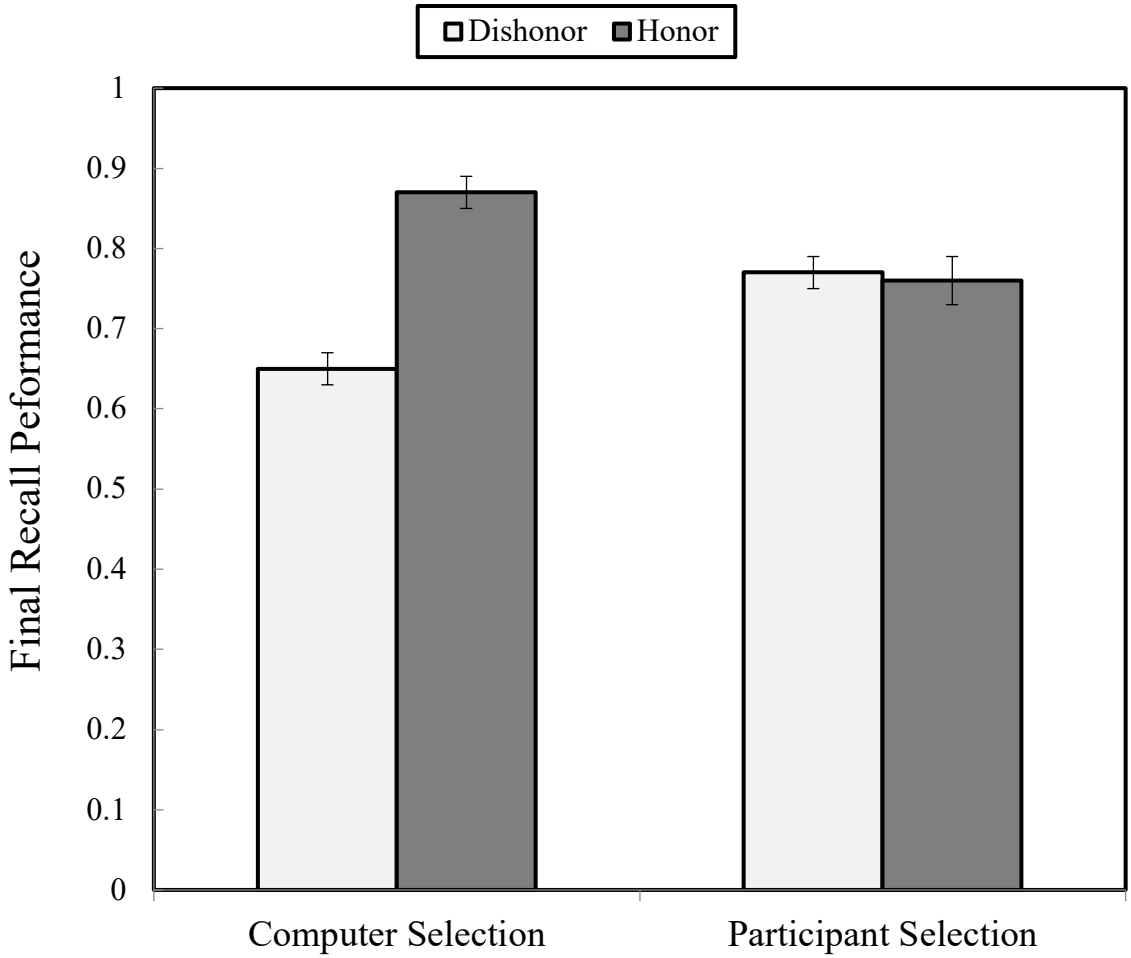

Figure 3. Final recall performance as a function of group for Experiment 3. Bars are standard errors of the mean.

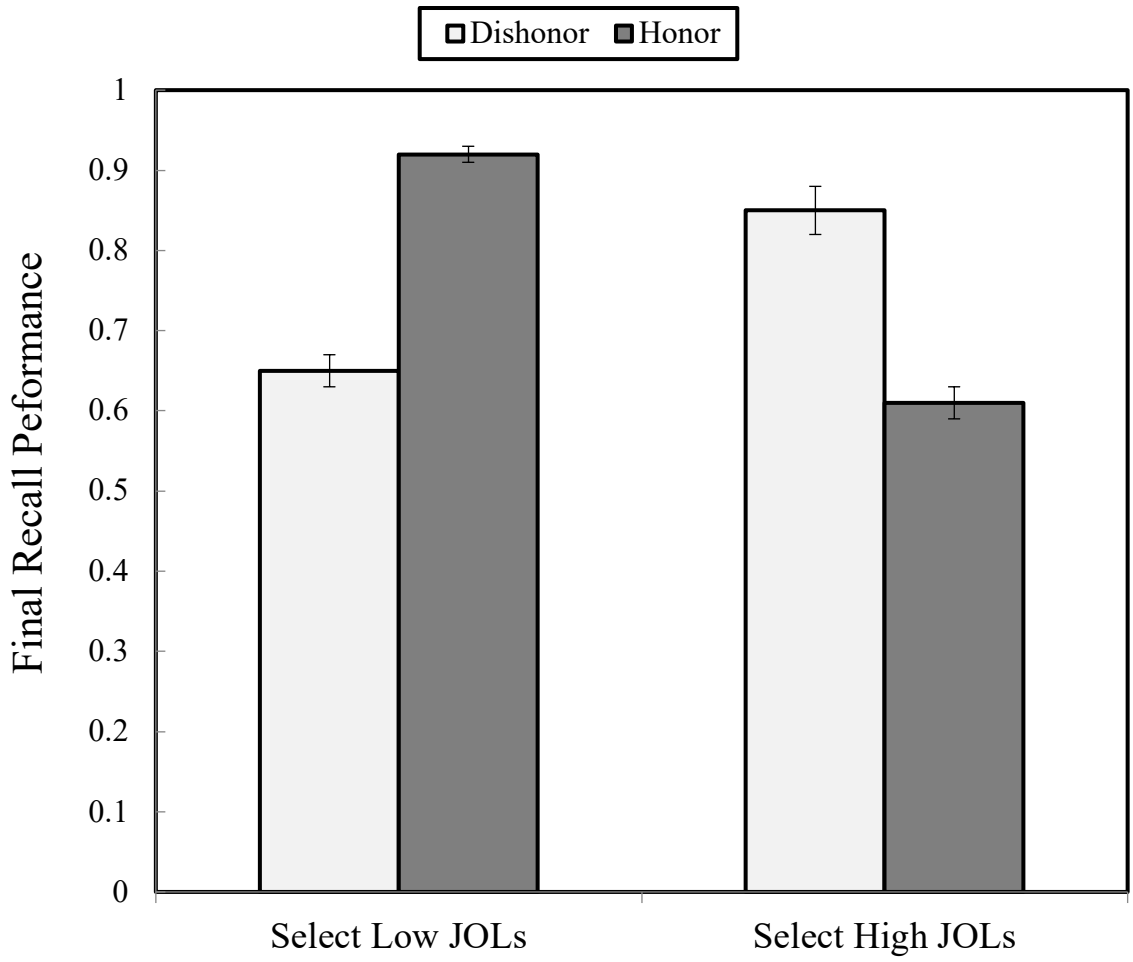

Figure 4. Final recall performance (Experiment 3) for the self-selection groups as a funciton of honor and dishonor groups and as a function of whether participants tended to select items given low judgments of learning (JOLs) for restudy or tended to select items given high JOLs for restudy. Bars are standard errors of the mean. See text for details.