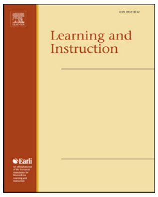
Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



Overconfidence in children's multi-trial judgments of learning

Bridgid Finn [a,*], Janet Metcalfe [b]

[a] Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08540, USA
[b] Department of Psychology, Columbia University, 1190 Amsterdam Avenue, New York, NY 10027, USA

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form
30 December 2013
Accepted 2 January 2014

Keywords:
Metacognition
Children
Overconfidence
Memory for past test

ABSTRACT

The underconfidence with practice effect (UWP) refers to the finding that people's judgments of learning shift from overconfidence to underconfidence on and after a first study-test trial (Koriat, Ma'ayan, & Sheffer, 2002). Finn and Metcalfe (2007, 2008) proposed that people show UWP because they use their memory of prior test performance as a cue to make subsequent judgments of learning and inadequately account for new learning (i.e. the Memory for Past Test (MPT) heuristic). In contrast to adults, 3rd and 5th graders' judgments showed persistent overconfidence on and after a first study-test trial. A second experiment tested children's ability to remember their prior test performance. Children's prior performance discriminations were accurate for items that they answered correctly on the prior trial, but were overconfident for items they had answered incorrectly indicating that their continued overconfidence was a result of faulty memory, rather than a failure to use the MPT heuristic.

1. Introduction

One of the guiding assumptions of theories of self-regulated learning is that people monitor past performance and make use of this information to regulate their future learning (e.g. Baker & Brown, 1984; Pressley, Borkowski, & Schneider, 1987; Pressley & Ghatala, 1989). Knowledge of how well one has previously performed should increase the likelihood of adopting effective and adaptive self-regulatory behaviors in the future. According to Moynahan (1973), "The ability to evaluate one's recall performance most likely is essential for effective memory monitoring, as it would seem difficult, if not impossible, for the subject to assess the effectiveness of a given recall strategy…if he did not know how well he had performed while using that strategy." (p. 246).

Adults are quite accurate in monitoring their past test performance, and can correctly discriminate previously incorrect from previously correct items (Finn & Metcalfe, 2008; Gardiner & Klee, 1976; Robinson & Kulp, 1970). In addition, adults are able to draw on their prior test performance to modify their predictive judgments and encoding strategies on subsequent learning trials (Finn & Metcalfe, 2007, 2008; Gardiner, Passmore, Herriot, & Klee, 1977; Halff, 1977; King, Zechmeister, & Shaughnessy, 1980; LaPorte & Voss, 1974). Children however, are not always able to make good use of their prior test performance to adjust their judgments and study strategies during successive learning trials (e.g. Bisanz, Vesonder, & Voss, 1978; Stipek & Hoffman, 1980; Stipek, Roberts, & Sanborn, 1984). There is some indication that children may remain overly optimistic even after they have had experience and feedback with a task (e.g. Lipko, Dunlosky, & Merriman, 2009; Shin, Bjorklund, & Beck, 2007; but see Lipko, Dunlosky, Lipowski, & Merriman, 2012) a pattern that may be protective against a loss of motivation (Bjorklund, 1997). Lipko et al. (2009) for example, asked preschool children to study 10 pictures, predict how many they would recall and then attempt to recall them. This entire cycle was repeated three times, with new pictures each time. Results showed that the preschooler's recall predictions were overconfident across all three trials, that is, they appeared not to use their recent experience to regulate their confidence about how they would perform on the new list.

Adults, in contrast, do use their prior experiences to regulate their metacognitions. While adults are typically overconfident on a first learning trial, they then shift to underconfidence on subsequent learning trials (e.g. Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007, 2008; Koriat, Sheffer, & Ma'ayan, 2002; Serra & Dunlosky, 2005), a well-studied phenomenon known as underconfidence with practice. In the standard paradigm showing the underconfidence with practice effect, participants study cue target pairs, make judgments of learning (JOLs, predictive judgments about performance on an upcoming test), and then take a cued recall test. The study–judge–test cycle is repeated multiple times with the same list of items, and underconfidence is shown after the first study-test trial. Finn and Metcalfe (2007, 2008) showed that underconfidence with practice results because people rely on their

[*] Corresponding author. Tel.: +1 646 942 0363.
E-mail addresses: bfinn@ets.org (B. Finn), jm348@columbia.edu (J. Metcalfe).

0959-4752/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.learninstruc.2014.01.001

memory of their performance on the prior test to make the subsequent judgments of learning (i.e. the Memory for Past Test (MPT) heuristic, and see also Ariel & Dunlosky, 2011; Cosentino, Metcalfe, Butterfield, & Stern, 2007; England & Serra, 2012 and Touron, Hertzog, & Speagle, 2010) and do not adjust for the new learning that has occurred during the subsequent trial.

In the current study, we tested children using the same paradigm as the one in which adults show a shift from overconfidence to underconfidence: In a first experiment, children studied, made item by item JOLs about their upcoming test performance, and were tested on the same list of vocabulary words for three trials. A second experiment tested children's ability to remember their prior test performance. A major objective was to explore children's JOLs over multiple trials to determine if their judgments, like those of adults, would demonstrate underconfidence with practice, or whether, as other research suggests (see e.g. Shin et al. 2007; for a review), they would stay persistently overconfident. The multi-trial paradigm has been extensively explored with adults, with results demonstrating that adults use their prior test experience to regulate their confidence. Thus, the paradigm allowed us to investigate the cues the children use to make their metacognitive judgments, to assess why children might fail to show the shift to underconfidence, and to isolate the locus of the expected overconfidence.

Understanding the source of children's persistent overconfidence even after multiple test trials, is critical if we are to inform educators and students how to evaluate learning more effectively and, consequently, optimize self-directed study. Because overconfidence has been shown to have critical consequences for the choices that students make when they self direct their own learning (Metcalfe & Finn, 2009), persistent overconfidence on the part of children could have adverse consequences. For example, inflated confidence could mislead students into overlooking items and concepts that could benefit from additional study time. This is not a trivial problem, as elementary school children are frequently given deskwork and homework during which time they are expected to effectively regulate their own study processes (Hofferth & Sandberg, 2001, and see also Metcalfe & Finn, 2013).

1.1. Metacognitive markers of the underconfidence with practice effect

The underconfidence with practice effect, repeatedly found with adults in multi-study-test trial experiments (e.g. Finn & Metcalfe, 2007, 2008; Koriat et al., 2002; Serra & Dunlosky, 2005), has been theorized (Finn & Metcalfe, 2007, 2008) to be due to use of the MPT heuristic after the first study-test trial to make JOLs about performance on the current trial. After the first study-test trial, JOLs incorporate information about item specific performance on the prior test. If the participant remembers that they failed to recall a particular item on the immediately past test, they give that item a lower JOL rating than if they remember that they recalled that item on the prior test. Underconfidence occurs if participants incorporate prior performance into their judgments and do not adjust appropriately for the new learning that has occurred in the study trial following the test. JOLs do increase over trials, but not enough to account for the new learning.

Underconfidence with practice is also characterized by changes in both the absolute and the relative accuracy of the metacognitive judgments over trials. The absolute accuracy, or calibration of the judgments, measures how well the mean item-by-item JOLs correspond to mean final test performance and provides an indication of whether a person can estimate their overall recall performance accurately (cf. Gigerenzer, Hoffrage, & Kleinböting, 1991). With repeated study-test trials JOLs display a calibration bias shift from an overestimation to an underestimation of performance. Whether a student's metacognitions are calibrated to their performance has important consequences for learning outcomes, since students study behaviors are closely tied to their metacognitions (Dunlosky & Thiede, 2013; Finn, 2008; Hacker, Bol, & Keener, 2008; Metcalfe & Finn, 2008): When students are overconfident they choose to study fewer items (Metcalfe & Finn, 2008) and their performance suffers (Dunlosky & Rawson, 2012).

Underconfidence with practice is also characterized by an increase in relative accuracy, or resolution, over trials. Resolution is an assessment of how well people can discriminate which items will be remembered and which will be forgotten. Resolution is high when people give low JOLs to items that they will get wrong on the test and high JOLs to items that they will answer correctly.

Along with the MPT heuristic, an anchoring-and-adjustment hypothesis has also been put forward as a companion explanation of the underconfidence component of the underconfidence with practice effect (England & Serra, 2012; Scheck & Nelson, 2005). It should be noted that the account does not attempt to account for the increase in relative accuracy over trials that also characterizes the underconfidence with practice effect. According to an anchoring explanation, people adjust their mean JOLs away from an anchor point. Underconfidence results when people adjust up from a psychological anchor point on the JOL scale but memory performance remains higher than the adjustment (Connor, Dunlosky, & Hertzog, 1997; England & Serra, 2012; Richards & Nelson, 2004; Scheck & Nelson, 2005). Generally, the anchoring explanation posits that people will be overconfident when performance is low (and below the anchor), and under confident when performance is high (and above the anchor). The explanation says that the underconfidence with practice effect is found because on the first learning trial performance is generally low (an overconfidence situation) but over learning trials becomes high (a classic underconfidence situation). Scheck and Nelson (2005) applied this logic to explain the underconfidence with practice effect and England and Serra (2012) have shown that when people are given instructions that the list will be easy as compared to difficult, overall judgments are different, providing evidence that anchoring affects JOLs.

2. The current studies

In Experiment 1 we tested whether Grade 3 and Grade 5 children's multi-trial judgments would show persistent overconfidence or show underconfidence with practice. Our hypothesis, given the large body of work demonstrating children's tendency toward overconfidence (e.g. Lipko et al., 2009; Shin et al., 2007), was that the children would not show underconfidence with practice (hypothesis 1). In contrast to our predictions, however, in a related experiment Lipko et al. (2012) concluded that by the 3rd grade children do show the underconfidence with practice effect. In their study kindergarten, 1st grade and 3rd grade participants were presented with a set of pictures of basic objects (e.g. clock, bug; 10 pictures were used for the kindergarteners and 1st graders, and 16 pictures were used for the 3rd graders). After the study phase, students were asked to make a global judgment about how many pictures they would remember after the pictures were covered. Then the pictures were covered and the students were asked to free recall the names of the pictures. After the recall phase the experimenter told the students how many they had recalled. The same procedure was then repeated immediately with the same set of pictures, in the same order, for a total of three study-test trials. Results showed that the Grade 3 children showed underconfidence on the second trial, but were not significantly underconfident on the third trial. Kindergarten and 1st grade children never showed underconfidence.

Although the results are suggestive, the paradigm used in the Lipko et al. study did not use a standard underconfidence with practice paradigm. It used pictures, not words; it used free recall rather than cued recall (as is normally the case in demonstrations of underconfidence with practice), predictions were made in the aggregate rather than item by item, and, perhaps most importantly, the children were explicitly told, at the end of each trial, exactly how many items they had just gotten right and wrong moments before they had to make judgments about how many they would get the next time. Hence, the judgments of the Grade 3 children may not have been based on their *memory* for their own item specific past test performance (as it is in adults) but rather on the explicit information that the experimenter had given them about how well or how poorly they had done overall. To make their next prediction the child could just remember what the experimenter just told him about his recall, rather than remembering how they had performed on the prior test, a cue that has been shown to be central in showing underconfidence with practice (Finn & Metcalfe, 2007, 2008). Although we do think that the results are interesting and informative, we do not think that they are conclusive with respect to whether children, when not explicitly told about their own immediately past performance, will or will not show the underconfidence with practice effect.

We highlight the study by Lipko et al. (2012) because, to our knowledge, it is the only one showing that children in the 3rd grade demonstrate anything like underconfidence with practice. We sought to investigate, in a standard paradigm, whether the children would show persistent overconfidence, or whether, as the Lipko et al. study suggests, they might show underconfidence with practice. We used cued recall with words,[1] and we tracked memory for individual items, as we had done in adults. Critically, we did not tell the children explicitly how many items they had just remembered on the just past test and we did not ask them to make only a global judgment of how many they would remember.

3. Experiment 1

3.1. Method

3.1.1. Participants, design, and materials

Twenty-four students from the Emily Dickinson School, PS-75, in New York City participated in total. Twelve participants were 3rd grade students (Mage = 7.75 years), and 12 were 5th grade students (Mage = 10.00 years). Approximately half the children in each group were girls. Only children who spoke fluent English were included in our sample. The children participated in an after-school program at Barnard College once a week. The small sample size reflects the total number of students that were available to participate as part of the after school program. There was no financial compensation for participation but the children were always given snacks before each day's activities, took part in interesting group activities when they were not participating in the experiment, and participated in a 'Columbia–Barnard party' at the end of the year. In this and in the studies that follow, the treatment of participants was in accordance with APA ethical standards. The sample was racially diverse (13% White, 28% Black, 51% Hispanic and 7% Asian) and was in line with New York City 3rd and 5th grade student demographics (3rd Grade: Asian: 14.8% Hispanic: 40.4% Black: 28.4% White: 15.4%; 5th Grade: Asian: 15.2%, Hispanic: 40.1%, Black: 29.3%, White: 14.6% (IBO, 2011)). Although precise information about the sample's SES was not available, the overall poverty level of the PS-75 school population, as measured by the percent of students eligible for free/reduced lunch, was 67%.

The design was a 2 x 3 mixed design: with grade (3rd vs. 5th) as a between participants factor and trials (1, 2 and 3) as a within participants repeated measure. Materials were definitions taken from the students' science and social studies textbooks and were the same for the 3rd and 5th grade participants and had been normed with other participants in prior years. The teachers in each grade reviewed the definitions and indicated that the materials were appropriate for the students. The general concepts that the questions focused on were fairly familiar to the students, as the materials had come from their own textbooks. The materials for our experiments are available on request.

3.1.2. Procedure

The experiment was conducted on iBook computers. Each child was tested individually, in a sound buffered room, with an experimenter/coach. At the beginning of the experiment participants were told that they were going to learn definitions. They were told to study so that later when they were given the definition they would be able to come up with the correct word on a test coming up in about 10 min. Participants were also informed that after each item they would be asked to indicate whether they thought they would remember the word that was paired with the definition in about 10 min. Before the experiment started, the experimenter made sure that the student understood what JOL was and how to make one using a slider scale in the program. They were given time to practice making JOLs using the slider scale. The study session consisted of 24 definitions presented for 3 s each. An example of one of the studied definition was, "The _____ is the fastest animal in the world. The answer (cheetah) was presented above the definition in red. After 3 s, the target answer (cheetah) disappeared while the definition remained. With the definition, but not the answer, present participants made a JOL on a slider scale ranging from 0 (not sure at all) to 100 (really sure) (Previous research and performance on the cued recall test established that students could learn approximately 40% of the definitions after one presentation). After they finished studying and making JOLs the definitions were shuffled and presented for a cued recall test. During the test, the definition appeared on the screen and was also read out aloud and the participant had to type in the correct word. The program had a built in formula that calculated a score for the response based on the way it was spelled (and which, in effect, ignored spelling errors). A score of 75 and above was treated as correct (and corresponds to what adult scorers would say are just spelling errors or typos, and hence are the correct answer). This study–judge–test cycle occurred three times with the same items. After all three cycles were completed the definitions were presented for a recognition test. Participants were asked to select the correct definition from a set of nine options.

3.2. Results

3.2.1. Cued recall test performance

A combined measures ANOVA showed a main effect of trial $F(2, 44) = 60.14$, $MSE = .01$, $p = .0001$, $\eta^2 = .73$. Performance increased over trials: Trial 1 ($M = .40$, $SE = .04$), Trial 2 ($M = .54$, $SE = .04$) and Trial 3 ($M = .64$, $SE = .05$). There was also a main effect of grade, $F(1, 22) = 4.94$, $MSE = .13$, $p = .03$, $\eta^2 = .18$. Grade 3 recall performance ($M = .44$, $SE = .06$) was lower than Grade 5 performance ($M = .62$, $SE = .06$). There was not a significant interaction between trial and grade, $F(2, 44) = 1.26$, $p > .05$. The range of performance on the test confirmed that the study materials and the allotted study time were appropriate for both age groups.

[1] Adults have typically shown underconfidence with practice with cue-target pairs such as (apple-chair) or foreign language vocabulary pairs (gato-cat). The vocabulary materials used with the children are a comparable cued recall task.

3.2.2. *Recognition test performance*

A univariate ANOVA showed a main effect of grade on the final recognition test, $F(1, 22) = 11.43$, $MSE = .02$, $p = .003$, $\eta^2 = .34$. Fifth grade participant's performance was better ($M = .95$, $SE = .04$) than 3rd grade participant's performance ($M = .77$, $SE = .04$).

3.2.3. *JOLs*

There was a significant main effect of trial, $F(2, 44) = 4.61$, $MSE = .03$, $p = .02$, $\eta^2 = .17$. JOLs increased over Trial 1 ($M = .68$, $SE = .03$), Trial 2 ($M = .70$, $SE = .04$) and Trial 3 ($M = .80$, $SE = .04$). There was not a significant main effect of grade, $F(1, 22) = 2.86$, $p > .05$. Nor was there a significant trial by grade interaction, $F(1, 22) = 1.13$, $p > .05$.

3.2.4. *Calibration*

Calibration is a measure of absolute accuracy and is measured by subtracting the mean recall performance subtracted from the mean JOL (both measures use a 0–100 point scale). When the calibration value is negative it reflects underconfidence while a positive value reflects overconfidence. Fig. 1 shows the difference between predicted and actual recall performance on all three trials. There was a significant main effect of trial, $F(2, 44) = 5.41$, $MSE = .02$, $p = .008$, $\eta^2 = .20$. On Trial 1 there was a large overconfidence bias ($M = .27$, $SE = .05$). Numerically the magnitude of the bias decreased from Trial 1 to Trial 2 ($M = .16$, $SE = .03$), although pairwise comparisons (all pairwise comparisons that follow were Bonferroni corrected for multiple comparisons) showed that the decrease was not statistically significant, $p > .05$. There was no difference in the overconfidence shown on Trial 2 and Trial 3 ($M = .15$, $SE = .04$), $p > .05$. There was no main effect of grade $F(1, 22) = 1.47$, $p > .05$. Nor was there a trial by grade interaction, $F < 1$, $p > .05$. While numerically the magnitude of overconfidence did decrease (but not significantly so) from Trial 1 to Trial 2, the children's judgments, nevertheless, did not come close to shifting to become underconfident. In line with our hypothesis, in contrast to adults, both the Grade 3 and Grade 5 children continued to be overconfident in their memories over all three trials (hypothesis 1).

3.2.5. *Gamma correlations*

We assessed resolution by computing gamma correlations, an index of predictive accuracy, for each participant, between JOLs and recall performance on each trial. Another marker of the underconfidence with practice effect, improved resolution over trials, was not obtained with the children, ($F < 1$). The average gamma was ($G = .57$, $SE = .10$) on Trial 1, ($G = .62$, $SE = .08$) on Trial 2, and ($G = .62$, $SE = .14$) on Trial 3. No other main effects or interactions were significant.

We computed a correlation between the current trial JOLs and performance on the preceding trial, allowing us to explore for evidence of use of the MPT heuristic. A stronger correlation between JOLs and the prior test (a correlation we will henceforth call a backward gamma), than between JOLs and the predicted current test (what we will henceforth call a forward gamma), would suggest use of past test performance. Supporting the claim that adults use MPT to make their JOLs on trials following a test, Finn and Metcalfe (2007) found significant differences between the magnitude of the backward and the standard forward gamma correlations with adults, with backward correlations more strongly positive than forward correlations. In line with our hypothesis, in contrast, the children's backward gammas (Trial 2 JOL – Trial 1 test performance, $G = .74$, $SE = .05$) were not significantly different from their forward gammas (Trial 2 JOL-Trial 2 performance, $G = .65$, $SE = .07$), $t(23) = 1.81$, $p > .05$, $d = .41$.

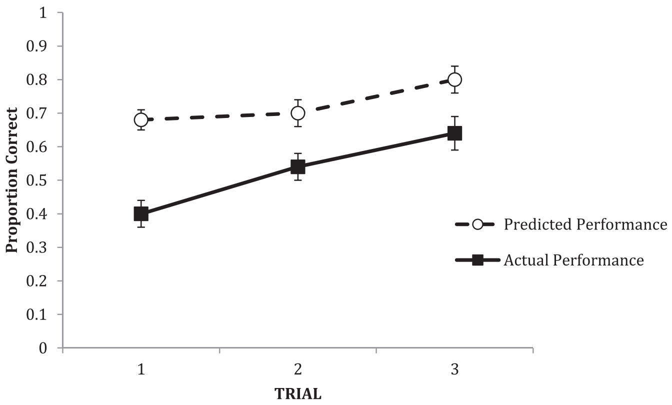


Fig. 1. Predicted and actual performance, collapsed over grade, for Trials 1, 2 and 3, Experiment 1.

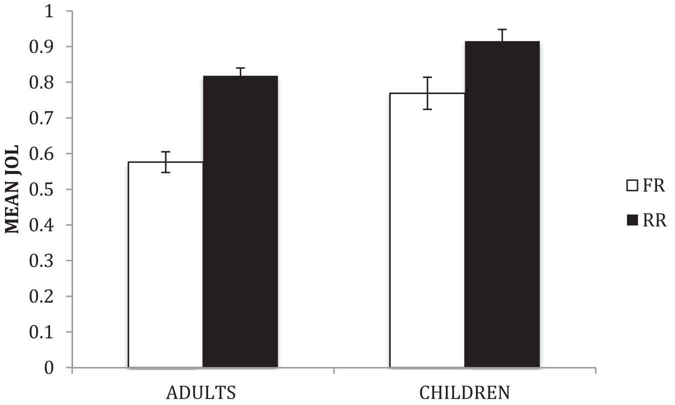


Fig. 2. Trial 2 JOLs conditionalized by recall performance on Trial 1 and Trial 2. Data depict findings from adult participants in Finn and Metcalfe (2007) and those from the children in Experiment 1. FR items were forgotten on Trial 1 and recalled on Trial 2, RR items were recalled on both Trial 1 and Trial 2. The task and materials were similar, but not identical for adult and child participants.

3.2.6. *Regression analyses*

Multiple regression analysis was used to examine the variation in Trial 2 JOLs in terms of Trial 1 and Trial 2 test performance, and allowed us to more clearly address how predictive each test was of Trial 2 JOLs. The data were analyzed using the prior trial and the current trial cued recall accuracy means for each participant as regressors and current trial JOLs as the dependant variable. We calculated separate regression models for Trial 2 and Trial 3 JOLs. The first model used Trial 2 JOLs as the dependent measure and Trial 1 and Trial 2 recall performance as regressors. The second model used Trial 3 JOLs as the dependent measure and Trial 2 and Trial 3 recall performance as regressors. In a previous study with adult participants, Finn and Metcalfe (2007) found that Trial 1 performance was a better predictor of Trial 2 JOLs than was Trial 2 performance. It should be noted that the task and the materials in the experiment with adults, although similar, were not identical to the task and materials in the experiment with children. With the children in the current study, a 2 (Trial: 2 or 3) × 2 (Predictor: prior trial or current trial) × 2 (Grade: 3rd or 5th) mixed measures ANOVA was computed using standardized beta scores from the regression models as the dependent measure. Results revealed no main effects or significant interactions ($Fs < 1$), and provided another indication that the prior trial did not appear to be weighted heavily in the children's judgments following the test.

3.2.7. JOLs conditionalized by recall performance on current and prior trials

To examine the relationship between the current trial JOLs and performance on the current trial and performance on the prior trial in more detail we broke down JOLs based on performance on the preceding trials. If children were relying on their prior test performance to make their judgments then items that would be correctly recalled on that trial should receive lower JOLs when they had been incorrect on the previous test as compared to when they had been correct on the previous test. Adults show this pattern, and the pattern predicts the underconfidence with practice effect. We conducted a 2 (Grade: 3rd or 5th) × 2 (Conditionalized set: FR: forgotten on Trial 1, remembered on Trial 2 versus RR: remembered on both trials) with the Trial 2 JOLs. Results showed that there was a main effect of conditionalized set $F(1, 21) = 15.54$, $MSE = 117.47$, $p = .001$, $\eta^2 = .49$, such that JOLs were lower overall for FR items. There was not a significant effect of grade, $F < 1$. Fig. 2 presents a contrast of the mean JOL for FR and RR items from the children and from adult participants in Finn and Metcalfe (2008). As can be seen in the figure, adults gave much lower JOLs to items that had been forgotten on the previous trial than the children did (.58 vs. .77 respectively). Indeed while 91% of the adult participants gave higher JOLs to RR items than to FR items only 71% of children showed this pattern.

We further examined the children's JOLs across all three trials with a 2 (Grade: 3rd or 5th) × 4 (conditionalized set: FFF: forgotten on all three trials, FFR: forgotten on Trials 1 and 2 and remembered on Trial 3, FRR: forgotten on Trial 1 and remembered on Trials 2 and 3, and RRR: remembered on all three trials) mixed measures ANOVA. The other conditionalized sets were excluded because instances of these categories were rare, (e.g., RRF: Remembered on Trial 1 and 2, but forgotten on Trial 3). Results revealed a main effect of conditionalized set $F(3, 54) = 13.88$, $MSE = .03$, $p = .0001$, $\eta^2 = .44$. As can be seen in Fig. 3, items remembered on all three trials (RRR) were given the highest JOLs, ($M = .92$, $SE = .03$), all $p < .05$. Notably, the items that had been forgotten on all three trials (FFF) showed a massive overconfidence bias, $M = .57$, $SE = .04$, and the JOLs for FFF items were not different than the JOLs for FFR items, $p > .05$. Even when an item had been incorrect on every trial, the children remained optimistic that they would remember it on the upcoming test. There was neither a main effect of grade ($F < 1$), nor was there a significant interaction between grade and conditionalized set ($F < 1$).

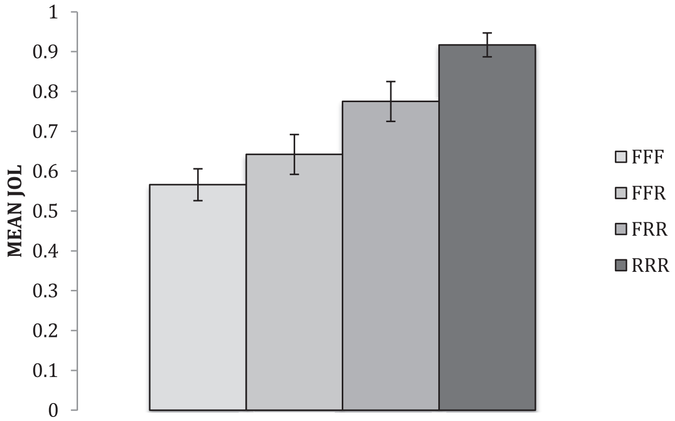


Fig. 3. JOLs conditionalized by recall performance on current and prior trials, collapsed over grade, for Trials 1, 2 and 3, Experiment 1. FFF items were forgotten on all three trials, FFR items were forgotten on Trials 1 and 2 and recalled on Trial 3, FRR items were forgotten on Trial 1 and recalled on Trials 2 and 3, and RRR items were remembered on all three trials.

By the time of the Trial 2 judgments, the children had received a second study presentation of the correct response (and a third presentation by Trial 3), thus ruling out the possibility that the positive bias in the judgments was due simply to lack of feedback as to what was the correct response. However, we assessed whether the overconfidence that the children displayed was due only to misremembering commission errors as having been correct. An error was judged as a commission error if the participant had entered an incorrect word or phrase in place of the correct definition. For example, a commission error was logged when a participant answered 'wiggle' in response to the definition, 'Because they have such short legs, instead of walking, ducks ——.' The correct answer was 'waddle'. If the children were using MPT in their judgments but showed a positivity bias for those items that they had previously generated an incorrect response for, then Trial 2 JOLs should be high for items that had been commission errors on Trial 1 and should approach zero for items that had been omission errors on Trial 1. The mean Trial 2 JOL for items that had been Trial 1 commission errors was .61 ($SE = .06$), which was significantly higher than the mean Trial 2 JOL for items that had been Trial 1 omission errors, ($M = .48$, $SE = .06$), $t(20) = 2.08$, $p = .05$, $d = .51$. Both means were significantly greater than zero, (smallest $t(20) = 7.56$, $ps < .001$). Even when the Trial 1 omission error would not be remembered on Trial 2, the children's JOLs were extremely overconfident, ($M = .47$, SE .07, $t(19) = 6.70$, $p < .001$, $CI_{95\%}$: .32–.62).

Finally, we conducted an analysis of JOLs across trials split by Trial 1 recall performance. We computed a 2 (Trial 1 recall status: correct vs. incorrect) × Trial (1 vs. 2) ANOVA using the mean JOL as the dependent measure (see England & Serra, 2012). There was not an interaction between Trial 1 recall status and Trial, $F < 1$, that is, the children's JOLs increased similarly over trials for both correct (Trial 1: .84, Trial 2: .89) and incorrect (Trial 1: .54, Trial 2: .60) Trial 1 items. Again, the pattern suggested that the children did not appear to adjust their JOLs based on how they had actually performed on the previous trial.

3.3. Discussion

Experiment 1 showed that over three study-test trials, Grade 3 and Grade 5 children did not show underconfidence with practice. One marker of the effect, underconfidence over repeated trials, was not obtained on either Trial 2 or Trial 3. Instead the children's JOLs showed overconfidence on all three trials. A second marker of underconfidence with practice, improved resolution over trials, was also absent. The lack of either underconfidence or improved gamma correlations over trials provided evidence that the children were not broadly relying on the MPT heuristic when making their JOLs, since use of MPT is accompanied by such a pattern. Further correlational and conditional analyses targeting use of MPT revealed that the difference in magnitude between the backward and forward correlations between the children and the adults (a group that has consistently shown use of MPT), revealed that the strength of the backward correlations as compared to the forward correlations, were not as strong for children as they were for adults. Children showed considerable overconfidence even for items that they would never recall correctly, and this was true for both errors of commission and omission. Finally, the differences found between the adults and children did not appear to be due to an overall higher level of competence for adults. Adults use the MPT heuristic when performance is fairly low (e.g. Finn & Metcalfe, 2007; Trial 1: $M = .21$, Trial 2: $M = .40$ and when it is higher, Trial 1: $M = .50$, Trial 2: $M = .82$). Thus, Adults use MPT appropriately when they have many incorrect trials, and when they have fewer incorrect trials. In contrast, children's whose performance was average (.44 and .62 for Grade 3 and Grade 5 respectively) did not.

4. Experiment 2

One implication of the MPT explanation of the underconfidence with practice effect is that accurate memory for prior test performance is used to make JOLs on the subsequent trial. The results of Experiment 1 suggested that children's judgments following a test do not reflect performance on that test. It was possible however that the children were using MPT to make their judgments but not accurately remembering how they had performed. Thus, the considerable overconfidence that the children demonstrated in Experiment 1 may have been due either to a failure to use the MPT heuristic *or* to the use of a positively biased MPT heuristic.

In Experiment 2, the children studied, made JOLs and took a test on a first trial as in Experiment 1. On Trial 2 instead of asking for JOLs, we asked if the children could explicitly remember their performance on the previous trial. If children were positively biased in their memory of their past test performance then the items that they had answered incorrectly would be misremembered as having been correct (hypothesis 1). Alternatively, if the children's memory of which items they had answered correctly and which items they had answered incorrectly was accurate it would suggest that they were not incorporating that information into their JOLs on the subsequent trial (hypothesis 2).

4.1. Method

4.1.1. Participants, design, and materials

Twenty-five students from New York City PS 75 participated in total. Fifteen participants were 3rd grade students, and 10 were 5th grade students. Participant recruitment, the sample's gender, mean age, ethnicity and socioeconomic backgrounds were the same as in Experiment 1. The design and materials of Experiment 2 was identical to Experiment 1 except for two differences. The first difference was that participants repeated the study-test trials twice instead of three times as in Experiment 1. The second difference was that on Trial 2, instead of making a JOL for each item, participants made a memory for past test judgment. At the beginning of Trial 2 the children were told that as they studied the words, they would be asked to indicate whether they had answered it correctly on the test that they just had. If they remembered answering correctly, they hit a "Got it!" button. If they remembered answering incorrectly, they hit a "Missed it" button. A cued recall test followed the study and judgment phases.

4.2. Results

4.2.1. Cued recall test performance

A combined measures ANOVA revealed a main effect of trial $F(1, 23) = 85.44$, $MSE = .002$, $p < .001, \eta^2 = .79$. Performance increased over trials: Trial 1 ($M = .50$, $SE = .05$), Trial 2 ($M = .63$, $SE = .05$). There was not a main effect of grade, $F(1, 23) = 1.11$, $p > .05$. The interaction between trial and grade was significant, $F(1, 23) = 11.39$, $MSE = .002$, $p < .01$, $\eta^2 = .33$. Performance showed a larger increase from Trial 1 to Trial 2 for 5th graders (Trial 1: $M = .53$, $SE = .08$, Trial 2: $M = .71$, $SE = .08$) than for 3rd graders (Trial 1: $M = .47$, $SE = .06$, Trial 2: $M = .56$, $SE = .07$).

4.2.2. Trial 1 calibration

JOLs were overconfident on Trial 1, $M = .16$, $SE = .04$. There was not a significant difference in calibration between the 3rd and 5th graders, $t < 1$, $p > .05$.

4.2.3. Memory for past test accuracy

Memory for past test judgments were coded as 1 when they were correct and 0 when they were incorrect. A separate analysis revealed no effect of grade for the children's memory for past test judgments, and so the following analyses collapsed over this variable. To assess the accuracy of the MPT judgments we compared Trial 1 test performance to the average MPT judgment. A comparison of MPT judgments and Trial 1 test performance found that MPT judgments were higher on average ($M = .57$, $SE = .05$) than was Trial 1 recall performance ($M = .49$, $SE = .05$), $t(24) = 2.04$, $p = .05$, $d = .50$, which indicated that the children were overconfident about how much they had remembered on the prior test.

To explore discrimination accuracy in more detail we examined MPT judgments for items answered incorrectly and correctly. If the overconfidence seen in Experiment 1 was due to an overestimation of correctness of incorrect items on the prior trial, then we should find that the children's memory for incorrect items was positively biased. For a given item answered correctly on the first trial, the children said that they had answered it correctly .97 ($SE = .01$) of the time. Thus, their judgments for correct items were very accurate. In contrast, for a given item answered incorrectly on the first trial, the children said that they had answered it correctly .19 ($SE = .06$) of the time, which was significantly greater than zero, $t(23) = 3.45$, $p < .01$, $\text{CI}_{95\%}$: .08–.31. In contrast, adults show extremely good accuracy both for items they remembered and items they did not. For example, for adult participants in Finn and Metcalfe (2008, Experiment 3) the probability of saying that they had answered an item correctly when they had actually answered it correctly was .94, and the probability of falsely calling an item remembered when it had not been answered correctly was only .03.

We analyzed what proportion of the erroneous MPT judgments were made for commission errors in comparison to omission errors. It was possible that children were only remembering as correct those errors that they had generated a mistaken response to on the test. On average, .61 ($SE = .07$) of the items that were answered incorrectly on the first test were commission errors. The children misremembered having answered .28 ($SE = .08$) of the commission errors correctly and answering .12 ($SE = .06$) of the omission errors correctly. These proportions were marginally significantly different from each other, $t(18) = 2.02$, $p = .06$, $d = .48$. The proportion of commission errors that were remembered as correct was significantly different than zero, $t(22) = 3.51$, $p < .01$, $CI_{95\%}$: .10–.39. The proportion of omission errors remembered as correct was also significantly different from zero, $t(19) = 1.82$, $p(one\ tailed) < .05$,

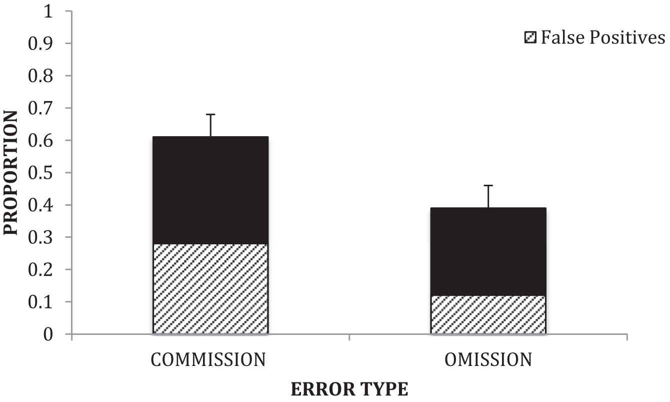


Fig. 4. Proportion of errors that were commission or omission errors in Experiment 2. Hatch marks represent the proportion of commission and omission errors that the children falsely remembered as having been correct, i.e. false positives.

$CI_{95\%}$: .01–.25. These data are depicted in Fig. 4. While the children remembered more of their commission errors as having been correct than omission errors, it is notable that the MPT judgments were overconfident even when no response had been entered. The results demonstrate that the children's memory of their prior test performance was inaccurate for items that they answered incorrectly—both when a mistaken response had been generated and when no response at all had been given.

4.2.4. Comparisons between JOLs and MPT judgments in Experiments 1 and 2

Would reliance on the positively biased MPT have resulted in overconfidence? If the children had relied MPT when making their Trial 2 JOLs in Experiment 1, then overconfidence should have resulted. To assess whether use of MPT would have resulted in Trial 2 overconfidence we compared the overconfidence of the Trial 2 JOLs in Experiment 1 with the overconfidence of MPT judgments in Experiment 2. Results revealed that the MPT judgments showed the same level of overconfidence ($M = .08$, $SE = .03$, $t(24) = 2.04$, $p = .05$, $CI_{95\%}$: .01–.15) as the Trial 2 JOLs in Experiment 1 ($M = .16$, $SE = .03$). This comparison indicated that relying on overly optimistic memory for past test performance would result in approximately the same magnitude of overconfidence as that seen with the JOLs.

4.3. Discussion

While the children's judgments were very accurate for items that were answered correctly on the prior trial, they were inaccurate in their discrimination of the incorrect items. The children misremembered approximately 20% of the incorrectly answered items as having been correct. This positivity bias wasn't shown only for commission errors, but was shown also with omission errors, albeit somewhat less pronounced. The results also demonstrated that a reliance on positively biased MPT would result in roughly the same level of overconfidence as that seen with the JOLs in Experiment 1.

5. General discussion

Experiment 1 used the same paradigm in which adults consistently show underconfidence with practice, but we found that, 3rd and 5th grade children did not. Rather than demonstrating a shift toward underconfidence after experience with the test, the children's judgments remained overconfident (hypothesis 1). Instead of demonstrating improved relative accuracy over trials, the accuracy of the children's gamma correlations also remained unchanged. Conditionalized analyses revealed that the children were especially overconfident with items that would never be remembered, those items that were forgotten on each of the three study test trials. While adults showed stronger backward correlations between JOLs and the prior test than they did forward correlations between JOLs and the upcoming test, the children did not. But, this did not necessarily mean that the children weren't using MPT in their subsequent judgments, as strong backward correlations indicate reliance on *accurate* memory for past test.

Experiment 2 tested children's discrimination of their prior test performance, since memory for past test is thought to be one of the primary cues used to make JOLs following a test. The children's memory for past test judgments were very accurate for items that they had answered correctly. However, the children were less accurate in discriminating incorrectly answered items. They made false positive errors on about 20% of the items that they had not answered correctly on the immediately preceding trial. Together these findings suggest that the children's memory for the prior test was incorrectly biased toward remembering more items as having been correct than actually were (hypothesis 1).

Taken together the results suggest that children were not using MPT, or at least not accurate MPT, to make their judgments. One possibility is that the children were using their positively biased MPT to make their judgments. In this scenario, when the children made their JOLs, they tried to recall their performance on the previous trial, but since they remembered getting more right than they actually did get right, the judgments were overconfident. Indeed, use of the positively biased information would predict the amount of overconfidence shown with the JOLs that followed the tests. Of course, another possibility is that the children did not use MPT at all to make their judgments. It may be that MPT heuristic is a more advanced metacognitive strategy that appears by young adulthood (e.g. Finn & Metcalfe, 2008). If not MPT, the children may have been using anchoring on all three trials, and setting an anchor point that was consistently higher than their performance.

Recently Tauber and Rhodes (2012) have suggested that use of the MPT heuristic could reflect a mixture of both indirect and direct memory effects. In their study they tested older and younger adults in a similar paradigm to the one presented here in which participants were given three study–judge–test cycles with the same set of items. They hypothesized that if use of the MPT heuristic only relies on an explicit search of memory, then older adults, who show age related episodic memory deficits (e.g. Jacoby, Jennings, & Hay, 1996; Zacks, Hasher, & Li, 2000), should not show the same pattern of underconfidence with practice as young adults. However, older adults in their study did show underconfidence with practice and the JOLs of young and older adult participants were similarly influenced by prior test performance. This pattern was in contrast to findings from Rast and Zimprich (2009) who found that older adults did not show underconfidence with practice over several study–judge–test trials. To complicate matters further, Cosentino et al. (2007) showed that among older adults with memory deficits the effect depended upon whether they were or were not aware of their memory deficit. Those who were unaware of the deficit failed to enlist the MPT heuristic and remained overconfident while those who were aware showed underconfidence with practice despite the memory deficit. Tauber and Rhodes concluded that because older adults typically show episodic memory deficits, yet demonstrated the same pattern of underconfidence as young adults, the MPT heuristic may be partially driven by an implicit memory mechanism. Unfortunately, none of these researchers asked the older adults to make explicit MPT judgments as we did here, and as Finn and Metcalfe (2008) did with young adult participants. Thus, although the Tauber and Rhodes (2012) conjecture is very plausible it has not yet been definitively demonstrated.

We believe our findings are complementary to other explanations that have been offered to account for children's overconfidence. Schneider and collaborators (e.g. Schneider, 1998; Schneider & Lockl, 2002; Visé & Schneider, 2000) suggest that overconfidence is not necessarily a marker of monitoring deficits or a problem of self-regulation. They appeal to research showing that there are conditions under which even kindergarten children can accurately monitor their memories (e.g. Yussen & Levy, 1975), and studies showing that children are, in general, able to make accurate predictions about another child's performance, even while showing overconfidence in their own (Schneider, 1998; Stipek, 1984). Instead Schneider and colleagues have suggested that overconfidence may reflect motivational biases such as wishful thinking. The wishful thinking hypotheses fits well with our finding that children incorrectly remembered errors on the prior test as having been answered correctly. Another explanation comes from Wellman (1985, and see also Heckhausen, 1984, for a related view) who suggests that children may over value the amount of effort that they have expended in attempting a particular task when making performance predictions, and undervalue how much information there is to be remembered. Young children in particular (e.g. preschool and kindergartener aged) do not consistently differentiate effort from ability (Nicholls, 1978; Stipek & MacIver, 1989). It is possible that children in our study remembered expending effort on a particular incorrect item, and took that to be a marker of correctness. Since the children do not assign lower confidence in learning to items on which they have previously been unsuccessful on a past test, it may be advantageous for teachers to know that children do not do this on their own, and to intervene to focus the children's attention and study on these particular 'not quite learned' items.

Overconfidence is not usually taken to be adaptive, however, there may be a motivational benefit in children's overestimations of their memories. Shin et al. (2007) examined the relationship between overconfidence in memory abilities and changes in memory performance over repeated trials with children. The children were classified into high and low overestimation groups based on how overconfident they were on the first study trial. The researcher's results showed that high estimators showed a greater increase in performance over trials than those who showed low levels of overconfidence. Bjorkland and colleagues (Bjorklund, 1997; Bjorklund & Green, 1992; Shin et al., 2007) have argued that children—perpetual novices—need to persist at tasks to increase their skills. Those children who believe their skills are better than they actual are may stick with a task longer than they would have if they were less confident in their abilities. A longer time on task means more practice, and ultimately improvements in criterion performance.

The experiments here provide insight into how children monitor their learning as it dynamically unfolds over the course of repeated study and test sessions. These findings have important educational implications. In the classroom students typically encounter the same material more than once making it critical to understand how students evaluate their knowledge as it changes over the course of repeated experiences with the material. By targeting the underconfidence with practice effect we were able to show developmental differences in the use of prior test performance–a cue that has been shown to be crucial for effectively updating learning strategies—between adults and children's subsequent predictions about their performance. In educational contexts students' overconfidence in their knowledge could lead to a false confidence in the strategies that they adopted to learn the information and an underestimation of how much time they should allocate to homework or to studying for upcoming tests.

Our findings show that children remember their prior errors as successes. The children's optimism is apparent in their judgments about how well they will do in the future and how well they did in the past. Our results indicate that their persistent overconfidence appears to be a result of faulty, positively biased memory for prior errors, rather than a failure to use the memory for past test heuristic.

Acknowledgments

This research was supported by Grant R305H060161 from the Institute of Educational Sciences, Department of Education, and Grant 220020166 from the James S. McDonnell Foundation. The authors are wholly responsible for the content of this article and thank the members of Metalab and the Roediger lab for their help and comments.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*, 171–184. http://dx.doi.org/10.3758/s13421-010-0002-y.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. David Pearson (Ed.), *Handbook of reading research*. New York: Longman.

Bisanz, G. L., Vesonder, G. T., & Voss, J. F. (1978). Knowledge of one's own responding and the relation of such knowledge to learning: a developmental study. *Journal of Experimental Child Psychology, 25*, 116–128. http://dx.doi.org/10.1016/0022-0965(78)90042-5.

Bjorklund, D. F. (1997). The role of immaturity in human development. *Psychological Bulletin, 122*, 153–169. http://dx.doi.org/10.1037/0033-2909.122.2.153.

Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist, 47*, 46–54. http://dx.doi.org/10.1037/0003-066X.47.1.46.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50–71. http://dx.doi.org/10.1037/0882-7974.12.1.50.

Cosentino, S. A., Metcalfe, J., Butterfield, B., & Stern, Y. (2007). Objective metamemory testing captures awareness of deficit in Alzheimer's disease. *Cortex, 43*, 1004–1019. http://dx.doi.org/10.1016/S0010-9452(08)70697-X.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280. http://dx.doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*, 58–61. http://dx.doi.org/10.1016/j.learninstruc.2012.05.002.

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review, 19*, 715–722. http://dx.doi.org/10.3758/s13423-012-0237-7.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36*, 813–821. http://dx.doi.org/10.3758/MC.36.4.813.

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. http://dx.doi.org/10.3758/PBR.15.1.174.

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

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: an assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 227–233. http://dx.doi.org/10.1016/0022-5371(76)90021-9.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16*, 45–54. http://dx.doi.org/10.1016/S0022-5371(77)80006-6.

Gigerenzer, G., Hoffrage, U., & Kleinbötting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*, 506–528. http://dx.doi.org/10.1006/obhd.1998.2807.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: a focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York: Taylor & Francis.

Halff, H. M. (1977). The role of recall opportunities in learning to retrieve. *American Journal of Psychology, 90*, 383–406. http://www.jstor.org/stable/1421871.

Heckhausen, H. (1984). Emergent achievement behavior: some early developments. In J. Nicholls (Ed.), *Vol. 3. The development of achievement motivation.* (pp. 1–32). Greenwich, CT: JAI Press.

Hofferth, S., & Sandberg, J. F. (2001). Changes in American children's time, 1981–1997. In S. Hofferth, & T. Owens (Eds.), *Children at the millennium: Where have we come from, where are we going? Advances in life course research series* (pp. 193–229). New York: Elsevier Science.

Jacoby, L. L., Jennings, J. M., & Hay, J. F. (1996). *Dissociating automatic and consciously-controlled processes: implications for diagnosis and rehabilitation of memory deficits*. In D. J. Herrmann, C. L. McEvoy, C. Hertzog, P. Hertel, & M. K. Johnson (Eds.), *Basic and applied memory research: Theory in context* (Vol. 1); (pp. 161–193). Hillsdale, NJ: Erlbaum.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: the influence of retrieval practice. *American Journal of Psychology, 93*, 329–343. http://www.jstor.org/stable/1422236.

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

LaPorte, R., & Voss, J. F. (1974). Paired-associate acquisition as a function of number of initial nontest trials. *Journal of Experimental Psychology, 103*, 117–123. http://dx.doi.org/10.1037/h0036836.

Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young children are not underconfident with practice: the benefit of ignoring a fallible memory heuristic. *Journal of Cognition and Development, 13*, 174–188. http://dx.doi.org/10.1080/15248372.2011.577760.

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: the role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology, 103*, 152–166. http://dx.doi.org/10.1016/j.jecp.2008.10.002.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174–179.

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*, 1–28. http://dx.doi.org/10.1007/s11409-013-9094-7.

Moynahan, E. D. (1973). The development of knowledge concerning the effect of categorization upon free recall. *Child Development, 44*, 238–246. http://www.jstor.org/stable/1128042.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of own attainment, and the understanding that difficult tasks require more ability. *Child Development, 49*, 800–814.

New York City Independent Budget Office. (2011). *New York City Public school indicators: Demographics, resources, outcomes*. Retrieved from http://www.ibo.nyc.ny.us/iboreports/2011edindicatorsreport.pdf.

Pressley, M., Borkowski, J. G., & Schneider, W. (1987). *Cognitive strategies: good strategy users coordinate metacognition and knowledge*. In R. Vasta, & G. Whilehurst (Eds.), *Annals of child development* (Vol. 5); (pp. 80–129). Greenwich, CT: JAI Press.

Pressley, M., & Ghatala, E. S. (1989). Metacognitive benefits of taking a test for children and young adolescents. *Journal of Experimental Child Psychology, 47*, 430–450. http://dx.doi.org/10.1016/0022-0965(89)90023-4.

Rast, P., & Zimprich, D. (2009). Individual differences and reliability of paired associates learning in younger and older adults. *Psychology and Aging, 24*, 1001. http://dx.doi.org/10.1037/a0016138.

Richards, R. M., & Nelson, T. O. (2004). Effect of the difficulty of prior items on the magnitude of judgment of learning for subsequent items. *American Journal of Psychology, 117*, 81–91.

Robinson, J. A., & Kulp, R. A. (1970). Knowledge of prior recall. *Journal of Verbal Learning and Verbal Behavior, 9*(1), 84–86.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with practice effect: boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*, 124–128. http://dx.doi.org/10.1037/0096-3445.134.1.124.

Schneider, W. (1998). Performance prediction in young children: effects of skill, metacognition and wishful thinking. *Developmental Science, 1*, 291–297.

Schneider, W., & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In T. Perfect, & B. Schwartz (Eds.), *Applied metacognition* (pp. 224–257). Cambridge, UK: Cambridge University Press.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence–with–practice effect? *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 1258–1266. http://dx.doi.org/10.1037/0278-7393.31.6.1258.

Shin, H., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's overestimation in a strategic memory task. *Cognitive Development, 22*, 197–212. http://dx.doi.org/10.1016/j.cogdev.2006.10.001.

Stipek, D. J. (1984). Young children's performance expectations: logical analysis or wishful thinking. *Advances in motivation and achievement, 3*, 33–56.

Stipek, D. J., & Hoffman, J. M. (1980). Development of children's performance–related judgments. *Child Development, 51*, 912–914. http://www.jstor.org/stable/1129485.

Stipek, D., & MacIver, D. (1989). Developmental changes in children's assessment of intellectual competence. *Child Development, 60*, 521–538.

Stipek, D. J., Roberts, T. A., & Sanborn, M. E. (1984). Preschool-age children's performance expectations for themselves and another child as a function of the incentive value of success and the salience of past performance. *Child Development, 55*, 1983–1989.

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27*, 474. http://dx.doi.org/10.1037/a0025246.

Touron, D. R., Hertzog, C., & Speagle, J. Z. (2010). Subjective learning discounts test type evidence. *Experimental Psychology, 57*, 327–337. http://dx.doi.org/10.1027/1618-3169/a000039.

Visé, M., & Schneider, W. (2000). Determinants of performance prediction in kindergarten and school children: the importance of metacognitive and motivational factors. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 32*, 51–58.

Wellman, H. M. (1985). The child's theory of mind: the development of conceptions of cognition. In S. R. Yussen (Ed.), *The growth of reflection in children.* (pp. 169–206). San Diego, CA: Academic Press.

Yussen, S. R., & Levy, V. M. (1975). Developmental changes in predicting one's own span of short-term memory. *Journal of Experimental Child Psychology, 19*(3), 502–508.

Zacks, R. T., Hasher, L., & Li, K. Z. (2000). Human memory. In F. I. M. Craik, & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed) (pp. 293–357). Mahwah, NJ: Lawrence Erlbaum Associates.