Title

The Unintended Consequences of Checklists

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36)

ISSN

1069-7977

Authors

Stave, Elise A.
Muentener, Paul J.
Schulz, Laura E.

Publication Date

2014

Peer reviewed

The Unintended Consequences of Checklists

Elise A. Stave (estave@alum.mit.edu), Paul J. Muentener (pmuenten@mit.edu), & Laura E. Schulz (lschulz@mit.edu)
Department of Brain & Cognitive Sciences, MIT, Cambridge, MA 02138 USA

Abstract

Research suggests that checklists reduce errors in fields ranging from aviation to medicine. Checklists are effective in part because their content is not randomly selected from available information but strongly sampled from information experts believe is critical. This sampling process supports the inference that unlisted information is unlikely to be important. However, this predicts that checklists might leave learners selectively vulnerable to unlisted sources of error. In Experiment 1, we show that adults in an aviation class detect fewer unlisted sources of error given a checklist than at baseline. In Experiment 2, we show that this inductive bias does not require previous experience with checklists: given a checklist for organizing a room, children (mean: 62 months) selectively overlooked unlisted items relative to baseline, and did so even when told the list might be incomplete.

Keywords: checklist, inductive bias, pedagogy, cognitive development.

The widespread adoption of checklists has dramatically improved public health and safety. Use of surgical safety checklists globally is associated with a nearly two-fold decrease in surgical mortality and a significantly lower rate of complications (Haynes, et al., 2009). In emergency rooms, checklists are associated with a nearly four-fold reduction in medical errors (Arriaga, et al., 2013). Aviation checklists are routinely used to assure the completion of complex tasks before takeoff and during the flight (Rantz 2009) and the Civil Air Patrol trains pilots on the IM SAFE mnemonic checklist (for illness, medication, stress, alcohol, fatigue, and eating) to determine whether a pilot is fit to fly (Hales & Pronovost, 2006). Lately, checklists have gained prominence in the popular press as policy interventions offering a rare combination of simplicity, affordability, and efficacy (see Gawande, 2009 for discussion and review).

By limiting the range of hypotheses (e.g., about possible sources of error) that learners need to consider, checklists support rapid and efficient monitoring of the environment. However, we suggest that this benefit comes at a cost: precisely because checklists constrain the hypothesis space, checklist users should be less likely than users at baseline to attend to any unlisted information. If this unlisted information turns out to be important, checklists may have a paradoxical effect, leaving learners more vulnerable to error than they were at baseline.

In this respect, checklists inherit the advantages and disadvantages of pedagogical instruction more generally. Recent computational work has formalized pedagogical reasoning assuming a) that teachers sample evidence most likely to support a rational learner's belief in the target hypothesis (as specified by Bayes' law), and b) that learners update their beliefs assuming that the teacher is sampling data accordingly. The details of the computational account have been specified elsewhere (Shafto & Goodman, 2008; Shafto, Goodman, & Frank, 2012) and are not critical to the current work, so here we will discuss the inferences supported by the formalization intuitively.

Consider inferences about sources of error in an aviation scenario. We can imagine a default sampling process in which a learner assumes that errors are rare (Oaksford & Chater, 1994) and randomly inspects the environment for potential problems without knowing which problems are most important or most likely. Under this weak sampling assumption (Tenenbaum, 1999) the learner may detect some number (n) of potential errors but this discovery should not license (above the baseline assumption that errors are rare) the learner's belief that there are no other, as yet undiscovered, sources of error.

Now consider the same inferences in a pedagogical context. An expert constructing a checklist should also assume that problems are rare. She should list potential problems because listing only these raises the probability of the correct hypothesis more than adding things that are unlikely to be problems. A learner who assumes this is what the expert is doing is thus licensed to assume that things not on the list are unlikely to be problems.

Additionally, the expert should try to list *all* likely sources of error. If the learner believes the expert might list only a subset (n – 1) of the potential errors, the learner will have to consider more hypotheses overall and each hypothesis (including the correct one, n) will be treated as less probable. A related set of inferences governs the learner. Given a checklist listing n items, the hypothesis that there are *only* n potential problems is more probable than the hypothesis that there are n + 1 (or more) potential problems because it is more likely that the expert would have listed n sources of error if those were the *only* probable sources of error than if additional potential errors were present.

Thus in contrast to the default sampling process, the pedagogical sampling process not only provides grounds for discovering n sources of error, but also supports the inference that there are no other, as yet undiscovered, sources of error. Intuitively, the learner can infer that if there were other likely sources of error, a knowledgeable, helpful expert should have listed them. The strong sampling assumption licensed by the pedagogical context thus strengthens the inference that absence of evidence for a problem is evidence of its absence. Checklists then, may support an inductive bias that cuts both ways. Effective checklists constrain the information the learner needs to consider, promoting efficient identification of the target information. However, because the checklist is effective precisely *because* it constrains the information the learner considers, it may reduce the probability that learners consider non-target information.

Previously, researchers have investigated this double-edged sword of pedagogy in the context of children's exploratory play (Bonawitz et al., 2011). The study showed that children shown a single function of a toy in a pedagogical context (e.g., "This is my toy. Watch this!") explored less and discovered fewer additional functions of the toy than they did in a baseline condition, a condition when the function was demonstrated accidentally, or a condition when the instruction was interrupted to imply that the evidence might not be exhaustive (e.g., "This is my toy. Watch this! Oh, I have to take a phone call"). However, outside of experimental contexts, teachers are unlikely to willfully provide evidence that induces a false belief in the learner. Indeed, the pedagogical sampling assumption is based on the premise that teachers typically select evidence that helps learners converge on the correct hypothesis. The greater risk lies with what past Secretary of Defense Donald Rumsfeld infamously but pithily described as "unknown unknowns". If both the teacher and learner assume that the pedagogically communicated information is exhaustive, but other information turns out to be critical, then the instructed learner may be selectively disadvantaged. This concern becomes especially critical in the context of checklists intended to promote public safety. In helping learners avoid expected sources of error, checklists may induce greater vulnerability to unexpected sources of error. Thus here we extend the previous work on the potential costs of pedagogically transmitted information (Bonawitz et al., 2011) to see if effects observed in traditional instruction (a teacher talking to a learner) obtain for any kind of information transmitted from a presumably helpful, knowledgeable source to a naïve learner.

We predict that, relative to baseline, checklists will improve learners' performance on listed sources of error but impair their performance on unlisted sources of error. In Experiment 1, we test these predictions with adults in an aviation class with an aviation-based picture task. In Experiment 2, we test the hypothesis that the selective impairment results from a rational inference about how evidence is sampled and does not depend on previous experience with checklists, by looking at whether the same predictions hold for young children (four to six-year-olds). Additionally, we look at whether the unintended consequences of checklists can be eliminated simply by telling learners that the checklist may be incomplete and urging them to explore broadly.

Experiment 1

Methods

Participants

Eighteen adults (mean age: 26 years; range 18 – 48 years; 72% male) were recruited through the MIT Ground School, an aviation instruction class. Participants were told they were going to participate in a study of human factors in aviation. Participants were randomly assigned to a *Checklist* condition or a *Baseline* condition. One participant, randomized to the *Checklist* condition, did not begin the experiment, resulting in a final sample of eight participants in the *Checklist* condition and nine in the *Baseline* condition.

Materials

Participants were given a drawing of a plane taking off on a runway. (See Figure 1.) The task was constructed in conversation with the Ground School instructor, who ensured that all the relevant information was familiar to the students from previous classwork.

Procedure

Participants were tested at their desks in an MIT classroom at the beginning of a Ground School class. The experiment was conducted as a paper/pencil task.

All participants were given a sheet of paper with the picture of the plane and the runway and read the following instructions: "You are on your designated runway awaiting clearance for takeoff. Please **circle and label** the aspects of this image that need to be considered in order to ensure a safe and successful takeoff." In the *Checklist* condition, participants were also given a written checklist with the following instructions: "Use this checklist to guide you: Type of runway, Fuel, Center of gravity, Runway slope, Wind direction". In addition to the five listed items, there were five other target items on the picture: Length of runway, Obstructions, Cloud cover, Density Altitude, Airport Elevation. The number of these ten target items circled was used as the dependent measure. No other features of the picture were coded. Participants were allowed as much time as they liked to complete the task, after which they were asked to hand their papers in.

Results

Throughout, we will refer to the items listed on the checklist as checklist items and the items not listed as unlisted items, although for participants in the baseline condition no checklist was available. Participant responses were coded by an individual unaware of the hypothesis. 100% of the data was also recoded by the first author. Intercoder reliability was high (Cohen's Kappa > .9). Results are displayed in Figure 1.

Figure 1: Picture used in Experiment 1

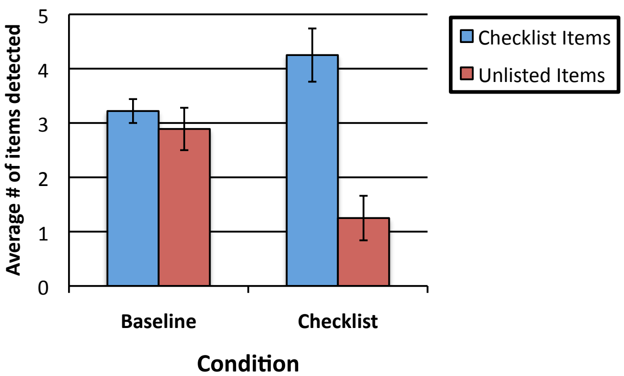

Figure 2: Results from Experiment 1

There was no significant difference between conditions in the total number of target items identified (*Checklist M* = 5.50, *SD* = 2.27; *Baseline M* = 6.11, *SD* = 1.45, t(15) = 0.74, *p* = *ns*).

We conducted a 2 x 2 ANOVA with item type (Checklist items vs Unlisted items) as the repeated measure and condition (*Baseline* vs. *Checklist*) as the between subjects measure. There was a main effect of item type (F(1, 15) = 22.06, $p < .001$), and no effect of condition (F(1, 15) = .55, *p* = ns). As predicted, there was an interaction between item type and condition (F(1, 15) = 14.12, $p < .01$): participants in the *Checklist* condition were less likely to identify unlisted items than those at *Baseline* (*Checklist M* = 1.25, *SD* = 1.16; *Baseline M* = 2.89, *SD* = 1.17, t(15) = 2.89, $p < .05$). Participants were also more likely to detect checklist items in the *Checklist* condition than the *Baseline* condition (*Checklist M* = 4.25, *SD* = 1.39; *Baseline M* = 3.22, *SD* = 0.67, t(15) = 1.98, $p < .05$, one-tailed).

Arguably, participants in the *Checklist* condition might have noticed the unlisted target items but inferred that they were only supposed to circle the listed items. However, the results suggest that this was not the case as 5 of the 8 participants (50%) in the *Checklist* condition listed at least one item not on the checklist.

Discussion

The results of Experiment 1 suggest that although checklists confer an advantage relative to baseline for listed items, they impair participants' ability to detect unlisted items. We suggest that these results follow from rational inferences about sampling processes. Arguably however, the results instead reflect learners' past successful experience with checklists. By the time adults attend a MIT aviation class, they presumably have successfully used checklists in tasks ranging from grocery shopping to assembling merchandise. Learners' experience with checklists, rather than their inferences about how evidence was sampled, may have led them to focus only on the checklist items.

To test whether the inferences induced by checklists depend on prior experience with checklists, we developed a pictorial checklist task appropriate for young children. Sensitivity to sampling processes has been demonstrated in very young children, indeed even in infants (Bonawitz et al., 2011; Gweon, Tenenbaum, & Schulz, 2009; Xu & Garcia, 2008; Kushnir, Xu, & Wellman, 2010). If the pattern of results found in Experiment 1 is due to rational assumptions about the sampling process, rather than past success with checklists, then even children with no, or very limited exposure to checklists, should be impaired on unlisted items relative to baseline. We also asked whether any negative consequences of checklists could be eliminated if learners were explicitly told that the checklist might be incomplete and encouraged to explore broadly.

Experiment 2

Participants Fifty-one children (mean: 62 months, range: 48-82 months; 51% male) were recruited at the Boston Children's Museum. Three children were excluded from analysis due to: choosing not to participate (n = 2) or experimenter error (n =1), leaving 48 children. Children were assigned to a *Checklist* condition, a *Baseline* condition, and an *Enhanced Checklist* condition, n = 16/condition.

Materials Children in the *Baseline* condition were given a picture of a messy room as they would find it when they walked in (Figure 3a). Children in the *Checklist* condition were given a checklist with pictures of five target items to clean up (Figure 3b.) Children in the *Enhanced Checklist* condition were given the same checklist as above, with eyes at the bottom of the page to remind them of a verbal prompt to look for additional items to fix (Figure 3c).

Procedure

Children were tested individually in a private testing room at the Children's Museum. Children were told that they would see a messy room and their job was to fix everything in the room. Children were then introduced to a room looking exactly like the room depicted in the picture in the *Baseline* condition. Children in the *Baseline* condition were told, "Fix everything in the room and use this picture to guide you." Children in the *Checklist* condition were told: "Fix everything in the room and use this checklist to guide you." In the *Enhanced Checklist* condition children were told "Fix everything in the room and use this checklist to guide you. There may be more items to fix than just the items on your checklist, so make sure to look around the room to check that you have done everything."

The child was left to explore the room freely and their behavior was videotaped. The experiment was terminated when children returned to the experimenter and said they were all done. If the child did not return to the experimenter but paused for ten consecutive seconds, the experimenter asked, "Are you all finished? Is there anything else you'd like to fix?" Children who said they were not finished were allowed to continue; if they said they were done or paused for another ten consecutive seconds, the experiment was terminated. The number of the ten target items that children fixed in the room was used as the dependent measure.

Results

The results are displayed in Figure 4. Again, we will refer to the items listed on the checklist as checklist items and the items not listed as unlisted items, although for participants in the *Baseline* condition, no checklist was

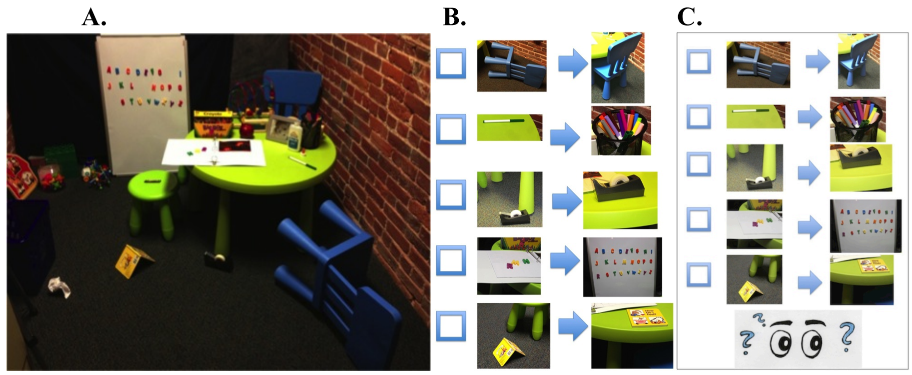

Figure 3. Stimuli used in Experiment 2: A. global picture used in *Baseline* condition; B. pictorial checklist used in *Checklist* condition; C. enhanced checklist in *Enhanced Checklist* condition

introduced. Participant responses were coded by the first author; 40% of the data was recoded from videotape by a coder blind to conditions. Intercoder reliability was high (Cohen's Kappa > 0.9).

We conducted a 2 x 3 ANOVA on participants' responses with item type (Checklist items vs Unlisted items) as the repeated measure and condition (*Baseline, Checklist*, *Enhanced Checklist*) as the between subjects measure. There was a main effect of item type ($F(1, 45) = 141.98$, $p < .001$) and a main effect of condition ($F(2, 45) = 4.13$, $p < .05$). As predicted, there was an interaction between item type and condition ($F(2, 45) = 23.66$, $p < .001$). To explore this interaction we conducted separate repeated-measures ANOVAs comparing the *Checklist* and *Enhanced Checklist* conditions to the *Baseline* condition.

A 2 (Item type: Checklist item vs. Unlisted item) X 2 (Condition: *Checklist* vs. *Baseline*) ANOVA yielded a main effect of item ($F(1, 30) = 64.50$, $p < .001$), a main effect of condition ($F(1, 30) = 6.08$, $p < .05$), and an interaction between item type and condition ($F(1, 30) = 38.66$, $p < .001$). This interaction reflected the fact that participants were more likely to fix unlisted items in the *Baseline* condition (M=3.56, SD=1.36) than the *Checklist* condition (M=1.06, SD=1.53) ($t(30) = 4.88$, $p < .001$); there was no difference for listed items (*Checklist:* M=4.5, SD=1.03; *Baseline*: M=4.0, SD=1.37; $t(30) = 1.17$, p = ns). This interaction also reflected the fact that participants tended to fix more checklist items than unlisted items in both conditions (*Checklist*: $t(15) = 8.223$, $p < .001$; *Baseline*: ($t(15) = 1.82$, $p = .089$).

Similar results found comparing the *Enhanced Checklist* condition to the *Baseline* condition. A 2 (Item type: Checklist item vs. Unlisted item) X 2 (Condition: *Enhanced Checklist* vs. *Baseline*) ANOVA yielded a main effect of item ($F(1, 30) = 74.43$, $p < .001$), a main effect of condition ($F(1, 30) = 4.19$, $p = .05$), and an interaction between item type and condition ($F(1, 30) = 42.83$, $p < .001$). This interaction reflected the fact that participants were more likely to fix unlisted items in the *Baseline* condition than the *Enhanced Checklist* condition (M=1.44, SD=1.21) ($t(30) = 4.66$, $p < .001$); there was no difference for listed items (*Enhanced Checklist:* M=4.63, SD=.72; $t(30)=1.62$, p = ns). This interaction also reflected the fact that participants tended to fix more checklist items than unlisted items in both conditions (*Enhanced Checklist*: $t(15) = 9.26$, $p < .001$, *Baseline*: ($t(15) = 1.82$, $p = .089$). The *Checklist* and *Enhanced Checklist* conditions did not differ.

As in Experiment 1, these results are unlikely to be due to participants in the Checklist conditions misinterpreting the task as one in which they were only supposed to correct the items on the Checklist: 7 of the 16 participants (44%) in the *Checklist* condition and 12 of the 16 participants (75%) in the *Enhanced Checklist* condition fixed at least one item that was not on the checklist.

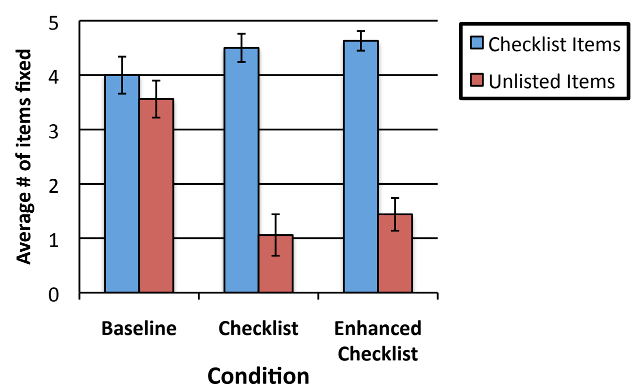

Figure 4: Results from Experiment 2

Discussion

The results of Experiment 2 suggest that prior experience with checklists is not necessary for learners to infer that only the listed items are important. Even very young children are susceptible to an inductive bias in which the presence of a checklist impairs their detection of unlisted items relative to baseline. Moreover, this bias is robust. Telling the children that there might be more items to fix than were listed on the checklist and prompting the children to explore more broadly did not ameliorate the unintended consequences of checklists. Children continued to perform poorly on unlisted items relative to baseline. Indeed, in Experiment 2, we failed to show any advantage for checklists: children were as successful in the *Baseline* condition as in the *Checklist* conditions. There are many possible explanations for our failure to replicate the benefits of checklist here. The experimenter attempted to match the checklist and non-checklist items for salience; nonetheless, it is possible that the checklist items were easier to detect overall and led all children to perform near ceiling. Alternatively, children's chance to pre-view a picture of the whole room in the *Baseline* condition might have offset any disadvantage they would otherwise have had with respect to the listed items. However, even though the children may or may not have experienced a concurrent or past history of benefitting from checklists, they seemed to assume the checklists were comprehensive and were impaired in detecting unlisted items. This suggests, consistent with previous work (Gweon et al., 2009) that the inferences of even very young children are sensitive to how information is sampled.

General Discussion

Many studies have focused on the positive consequences of checklists (Haynes 2009, Arriaga et al., 2012). However, the current work suggests that checklists also have unintended consequences. The very constraints on the hypothesis space that make checklists effective tools for efficient learning may support the inference that information not on the checklist is less likely to be important. Although rational, this inference may not always be accurate. Checklist designers are not omniscient, and our results suggest that checklists may leave learners especially vulnerable to unanticipated errors.

Our study also suggests that the inferential biases induced by checklists manifest at an early age, even before children have much experience with checklists or other kinds of formal instruction. Moreover, even in its earliest manifestation, the bias may be challenging to eradicate. Explicitly telling children that a checklist might not be exhaustive and encouraging them to look around failed to bring children's performance back to baseline.

The current results raise both theoretical and practical concerns. From a theoretical perspective, we have argued that the trade-offs induced by checklists result from assumptions about how evidence is sampled. Our findings are consistent with that account. Arguably however, providing a checklist might have simply diverted learners' attention toward the listed items and away from other aspects of the task. Many studies have shown that selective attention to some features of a task impairs attention to or recall for other features: this manifests in studies of change blindness (Simons & Levin, 1997; Simons & Rensink, 2005), in interference effects in memory (Anderson & Spellman, 1995; Healey, Campbell, Hasher, & Ossher, 2010; Postman & Underwood, 1973; Schooler & Engstler-Schooler, 1990), in cases of functional fixedness (e.g., German & Defeyter, 2000), and in negative priming (where priming one item impairs retrieval of others; May, Kane, & Hasher, 1995; Tipper, 2001). To distinguish a broader attentional account from our account based on sampling, future work might look at whether impaired detection of unlisted items is less likely to occur if the learner believes the items on the checklist are generated randomly, by a naïve learner, or by an interrupted instructor. Although these control conditions with checklists are still in progress, previous studies have run comparable conditions in both adults and children, and show that children selectively constrain their inferences to target items only when these are generated by knowledgeable agents and not when generated by non-pedagogical processes (e.g., Shafto & Goodman, 2008; Bonawitz et al., 2011; Gweon et al., 2009).

From a practical perspective, the current results raise questions about how to preserve the benefits of checklists without incurring the costs. We emphatically do not want to throw out the baby with the bathwater: checklists and other forms of pedagogical instruction are invaluable to learning. This point was made forcibly in the original research formalizing pedagogical sampling assumptions (Shafto & Goodman, 2008). In the experiment, learners were shown two dots sampled from inside a rectangle and asked to identify the rectangle from which the dots were sampled. If the dots are sampled randomly the task is impossible: an infinite number of rectangles contain the dots (e.g., rectangles that just barely contain the dots and rectangles the size of the page, the room, the state, etc.). However, when asked to sample two dots for a learner, participants reliably choose two dots that were not only consistent with the true hypothesis but distinguished the true hypothesis from all others: the two dots delineating the endpoints of the rectangle's diagonal. Learners, for their part, selected the rectangle on the diagonal, assuming a helpful partner had sampled the data (Shafto & Goodman, 2008). Pedagogical instruction thus converted an intractable search problem into a trivial one. The fact that pedagogical instruction promotes efficient learning by constraining the hypothesis space necessarily means that learners will be less likely to discover information not in the hypothesis space; however, insofar as the instructor is indeed knowledgeable, this will be more often a feature than a bug.

Consequently, we are very far from proposing that pilots should fly or that surgeons should operate without checklists. As noted, abundant research suggests that in the contexts in which checklists are normally used (where the environment is familiar, tasks are well rehearsed, the users are themselves experts, the listed items may otherwise be missed, and non-checklist errors are rare) the use of checklists reduces errors well below baseline (see Gawande, 2009, for a popular review). Nonetheless, the theoretical trade-off and empirical results discussed here reveal the potential vulnerability to unexpected sources of error. Indeed, learners' tendency to miss unlisted items could be magnified in contexts where a checklist is typically used without incident.

The practical challenge lies in how to retain the advantages of checklists while reducing exposure to their potential disadvantages. Unfortunately, the manipulation we attempted here was ineffectual. Merely warning learners that the checklist might be incomplete and encouraging additional exploration did not improve performance. Of course, we only failed to improve the performance of very young learners; the manipulation might be more effective in contexts that matter more. However, when possible, or when safety concerns are paramount, our account suggests that one optimal solution might be to send in two independent observers: one with a checklist and one without. The checklist user should be best equipped to detect anticipated sources of error; the other observer may be more likely to spot everything else.

Acknowledgments
Thanks to Sathya Silva of the MIT Ground School, the Boston Children's Museum and participating families. We also thank Jessica Wass and Aviana Polsky for coding.

References

Anderson, M., & Spellman, B. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102*, 68-100.

Arriaga, A.F., Bader, A.M., Wong, J.M., Lipsitz, S.R., Berry, W.R., Ziewacz, J.E., Hepner, D.L., Boorman, D.J., Pozner, C.N., Smink, D.S., Gawande, A.A. (2013). Simulation-Based Trial of Surgical-Crisis Checklists. *New England Journal of Medicine*, 368:246-253.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N.D., Spelke, E., Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*, 322–330.

Gawande, A. (2009). *The Checklist Manifesto: How to Get Things Right*. New York: Henry Holt and Company.

German, T., & Defeyter, M. (2000). Immunity to functional fixedness in young children. *Psychonomic Bulletin & Review, 7*, 707-712.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066-9071.

Hales, B., Pronovost, P. (2006). The checklist—a tool for error management and performance improvement. *Journal of Critical Care, 21*, 231-235.

Haynes, A.B., Weiser, T.G., Berry, W.R., Lipsitz, S.R., Breizat, A.S., Dellinger, E.P., Herbosa, T., Joseph, S., Kibatala, P.L., Lapitan, M.M., Merry, A.F., Moorthy, K., Reznick, R.K., Taylor, B., and Gawande, A.A. (2009). A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population. *New England Journal of Medicine*, *360*, 491-499.

Healey, M., Campbell, K., Hasher, L., & Ossher, L. (2010). Direct evidence for the role of inhibition in resolving interference in memory. *Psychological Science, 21*, 1464-1470.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134-1140.

May, C., Kane, M. & Hasher, L. (1995). Determinants of negative priming. *Psychological Bulletin, 118*, 35-54.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Postman, L., & Underwood, B. (1973). Critical issues in interference theory. *Memory and Cognition, 1*, 19-40.

Rantz, W. G., Dickinson, A. M., Sinclair, G. A., & Houten, R. V. (2009). The effect of feedback on the accuracy of checklist completion during instrument flight training. *Journal of applied behavior analysis*, *42*(3), 497-509.

Schooler, J. & Engsteler-Schooler, T. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36-71.

Shafto, P., Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. Proceedings of the *30th annual conference of the Cognitive Science Society*.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning From Others The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341-351.

Simons, D., & Levin, D. (1997). Change blindness. *Trends in Cognitive Sciences, 1*, 261-267.

Simons, D., & Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9*, 16-20.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning* (Doctoral dissertation, Massachusetts Institute of Technology).

Tipper, S. (2001). Does negative priming reflect inhibitory mechanisms? A review and integration of conflicting views. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 54*, 321-343.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.