CHAPTER X

IMITATION LEARNING IN PROBLEM SOLVING TASKS: MEMORIZING OR UNDERSTANDING?

Frédéric Dandurand†, Simcha Samuel† and Thomas R. Shultz‡

†Department of Psychology, McGill University (Canada)

‡ Department of Psychology and School of Computer Science, McGill University (Canada)

Many activities in life involve planning and solving problems. People learn to solve problems by trial-and-error, by explicit teaching, and by observation of skilled people. Observational learning, also called imitation learning, is a powerful, intrinsically social mechanism (Bandura 1986).

In cognitive psychology, problem solving is a classic research area. The dominant approach is information processing theory (Newell and Simon 1972) where problems are described using states, transitions, operators and constraints (Holyoak 1995). Traditional information processing approaches emphasize search and heuristics, often dismissing imitation learning as "rote memorizing" (Katona 1940).

By contrast, current research on imitation learning considers learning by demonstration to be more complex than mere memorizing. Although there is still debate about what counts as imitation and what can be explained away using other cognitive mechanisms, many researchers agree that learning by imitation can involve understanding of mentors' intentions (Carpenter, Call, and Tomasello 2002) and acquisition of complex hierarchical representations (Byrne and Russon 1998).

In previous work on learning by demonstrations in a problem solving task, we found that demonstrations improved participant's accuracy over simply being told if answers are correct or not (Dandurand, Bowen, and Shultz 2004). However, it remained unclear what mechanisms were involved in this accuracy improvement. Although participants may have abstracted or understood certain aspects of the task by watching demonstrations, rote memorizing could also account for the observed accuracy improvement because demonstrations and target problems were identical.

In this project, we seek evidence that improved accuracy was not simply due to rote memorizing, and that some degree of understanding mediates those performance improvements. We operationally define *understanding* as the ability to generalize observed solutions to new, unseen problems. We evaluate participants' abilities to generalize information seen in problem solving demonstrations to a novel, more complex variant of the problem.

0.1 Flexible, context-sensitive imitation

The literature on imitation learning provides evidence that knowledge acquired through observation is used in flexible, context-sensitive ways. For example, in one study, rhesus macaques were trained to respond to photographs (Subiaul et al. 2004). The positions of these photographs randomly changed on every trial so that sequences could not simply be memorized. Monkeys learned new sequences more rapidly after observing an expert execute those sequences than when they learned new sequences by trial and error. Therefore, macaques may have abstracted a general rule from imitation learning with experts.

Another study suggests that even fourteen-month-old infants do not mindlessly memorize and copy actions. Instead, they appear to evaluate whether these actions are deemed the most rational alternative available to achieve a goal (Gergely, Bekkering, and Kiraly 2002). If an adult touches an object with her head while her hands are available, the infant infers that there is an advantage to performing this action with one's head and imitates this action with his head. However, if the demonstrator's arms are covered, the infant imitates the action with his hands.

0.2 Accuracy in the ball weighing experiment

In a previous study about learning in problem solving tasks (Dandurand, Bowen, and Shultz 2004), a ball weighing experiment required that participants identify which ball weighed either more or less than the other eleven balls using a scale at most three times. We manipulated the information available for learning the task. Participants in an imitation learning condition watched five demonstrations of problems that were successfully solved. In contrast, participants in a reinforcement learning condition were only told if their answers were correct or not.

We found that the imitation learning group was more accurate (higher rate of correct answers) than the reinforcement learning group. From a strict information content basis, this result is not surprising; participants performed better with richer, more detailed information. Indeed, participants in the imitation groups were shown exactly how to solve problems, whereas those in the reinforcement learning group only got binary information (i.e., correct or not).

There is abundant evidence that better information quality results in more accuracy. For instance, immediate feedback does not necessarily allow learners to detect and self-correct errors. Furthermore, feedback does not diminish participants' propensity to seek external help from other sources (Reder and Klatzky 1994). Also, students give more correct explanations for their choices, and correct their answers significantly more when an agent explains why an answer is correct rather than simply telling whether answers are correct or not (Moreno and Mayer 2005).

Clearly, richer information resulted in better accuracy, but the mediating mechanism remains unclear. Some mechanisms may involve understanding, for instance insight and problem-independent encoding of features. However, because demonstrations and target tasks were identical, accuracy improvements could simply be attributed to rote memorizing. Also, it remained unclear how feedback improved accuracy compared to the baseline, when participants are not given any information for solving the task.

Two important questions were left unanswered in previous work. First, is binary explicit feedback (correct or not) useful? Second, why did demonstrations improve accuracy? Is understanding or memorizing involved?

Furthermore, we wondered if familiarization with the problem prior to watching demonstrations would increase accuracy. Perhaps participants who have the chance to work on problems before watching demonstrations can learn more because they better understand what is difficult about the task, and thus better focus attention on critical or difficult aspects. In contrast, participants who are not familiar with the problem may feel more overwhelmed with the information presented in the demonstrations.

0.3 New study of feedback and imitation learning in problem solving

We conducted a new study of feedback and imitation learning in problem solving tasks to investigate those questions. The methodology used was essentially the same as in the aforementioned laboratory experiment (Dandurand, Bowen, and Shultz 2004). To increase the attractiveness of the task and make it more fun, participants weighed gizmos that changed on each trial instead of balls. A screen shot of the program is presented in Figure 1.

We operationally defined *understanding* as the ability to generalize what was learned by observation to novel, more difficult problems. Some participants watched demonstrations of a simpler variant of task involving nine gizmos instead of twelve. The ability to generalize from nine-gizmo demonstrations to a twelve-gizmo task is viewed as evidence of understanding.

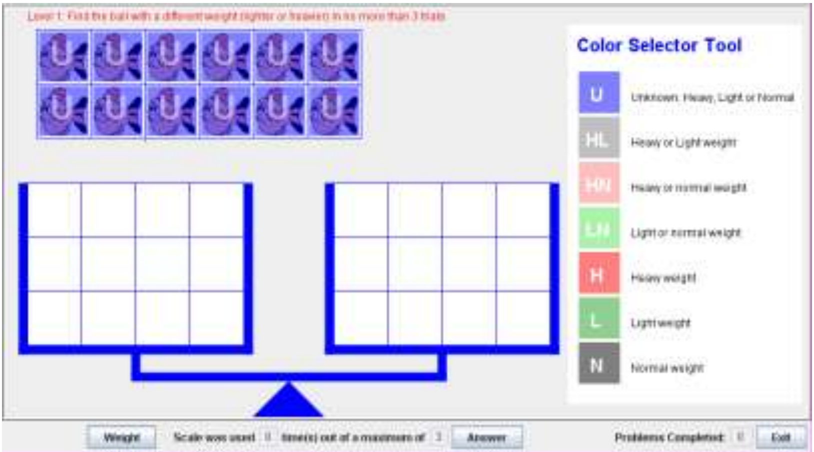

Figure 1 - The Gizmo Problem Solving task. On each trial, the gizmos display a different picture, such as a car or animal. On this screenshot, we see the twelve gizmos in the bank, the scale, and the color selector tool used to tag gizmos with informative labels (H for heavy weight, L for light weight, N for normal weight, LN for light or normal weight, HN for heavy or normal weight, U for unknown weight, HL for heavy or light weight).

0.4 Comparison of nine and twelve-gizmo solutions

The twelve-gizmo task (Figure 2) is more difficult and complex than the nine-gizmo task (Figure 3) because there are 24 cases to discriminate (12 gizmos x 2 weights) compared to 18 for the nine-gizmo variant (9 gizmos x 2 weights). Demonstrations presented to participants were based on five randomly selected branches from the solution trees presented in Figure 2 and Figure 3 respectively for twelve- and nine-gizmo tasks.

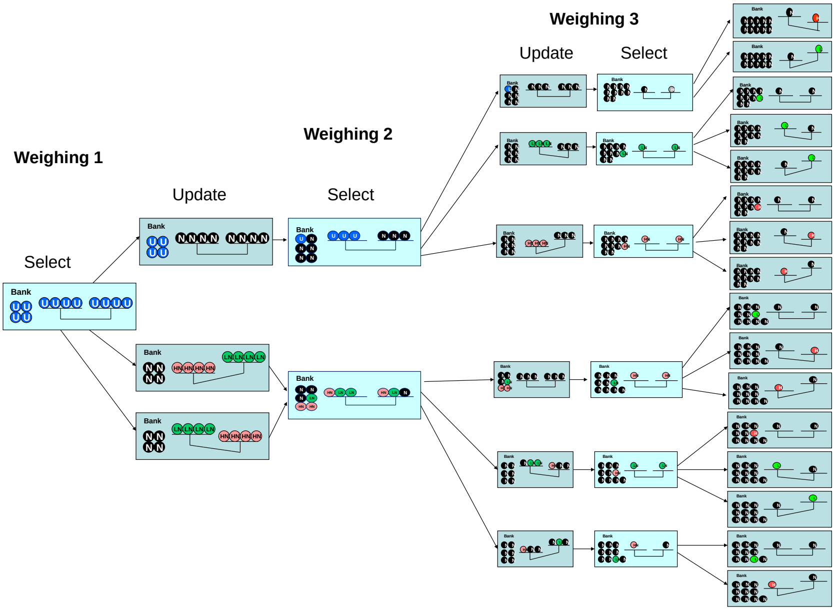

Figure 2 – Complete solution to the twelve-gizmo task. Five random branches were presented as demonstrations to the imitation learning group. For conciseness, two branches have been collapsed so 16 leaves are shown.

We analyzed similarities and differences between the solutions presented. On the first weighing, a 4 vs. 4 selection is presented in twelve-gizmo demonstrations, but a 3 vs. 3 selection is shown in the nine-gizmo variant. On the second weighing, when the scale balances on the first weighing, selections demonstrated are a 3xU vs. 3xN for the twelve-gizmo demonstrations, as opposed to a 2xU vs. 2xN for the nine-gizmo demonstrations. However, when the scale tips on the first weighing, both versions show a HN+2xLN vs. HN+LN+N selection. On the third weighing, solutions tend to overlap despite cascading differences in solutions originating in dissimilarities in the first and second weighings.

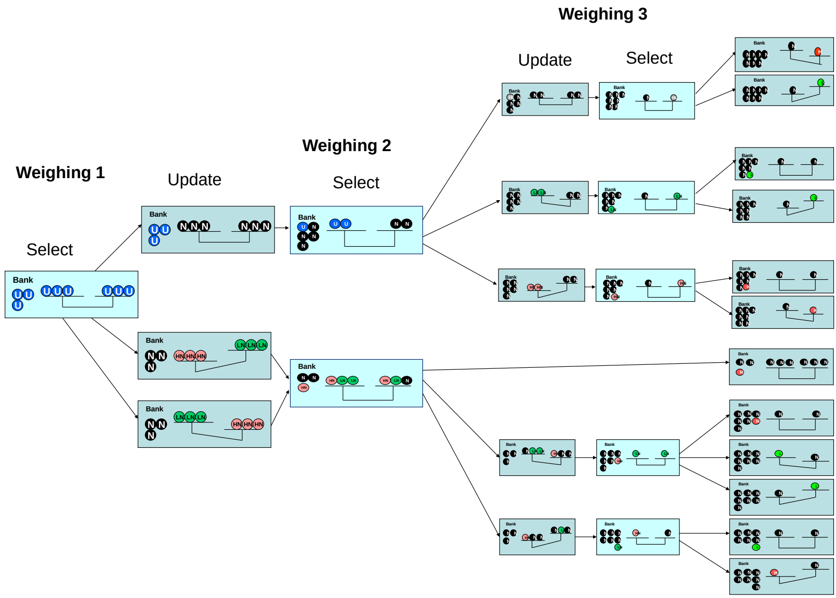

Figure 3 – Complete solution to the nine-gizmo task. For conciseness, two branches have been collapsed so 12 leaves are shown.

What performance could be expected if participants simply memorized nine-gizmo solutions? Participants' attention is naturally drawn to the scale, whereas the bank is often neglected (Dandurand, Shultz and Onishi 2007). Thus weighings that involve identical arrangements of gizmos on the scale should be perceived as equivalent, even though a different amount of gizmos are left in the bank. Therefore, reproducing a nine-gizmo solution from memory means reproducing the presented arrangements of gizmos on the scale. The predicted solution is shown in Figure 4. When the target is one of the three gizmos labeled as Unknown in the highlighted yellow box, participants cannot uniquely identify the target with 3 weighings only, and need to guess among six possibilities (3 gizmos x 2 weights). Therefore memorizing would yield a maximal accuracy of 19/24 (= 79%) with perfect memory[i].

In short, nine-gizmo problems provide relevant information for solving twelve-gizmo problems. However, if participants merely memorize nine-gizmo demonstrations, participants would incorrectly select 3 vs. 3 gizmos on the first weighing, and accuracy would be about 79% of those who watched twelve-gizmo problems, assuming all participants have equivalent abilities to memorize.

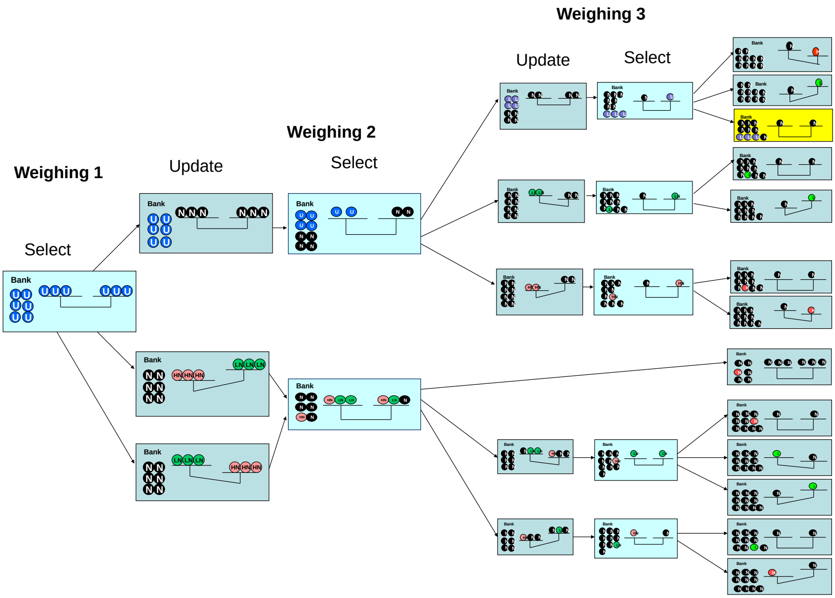

Figure 4 - Predicted solution to the twelve-gizmo problems if the nine-gizmo solutions are memorized. Using this strategy, participants are forced to guess among the 6 cases (3 gizmos x 2 weights) in the yellow box.

1. Method

1.1 Participants

One hundred thirty one people (95 females and 36 males) participated in the Gizmo Problem Solving experiment. Our sample was mainly composed of university students recruited either through the McGill Psychology Department subject pool in exchange for course credit or through personal contacts. The other 25 participants were unselected web users who read about our experiment on web sites that list psychology experiments.

Eleven participants were excluded due to internet and server problems, and power failures. This analysis did not include the verbal instructions group (n=20). The final sample of participants included 100 people. Mean participants' age was 22, ranging from 18 to 68 years of age.

1.2 Design

We used a two-way mixed design, with learning condition as a between-subjects factor, and quartile as a within-subjects factor. Our experiment investigated: (1) familiarization – give participants 5 minutes practice before watching demonstrations, (2) demonstrations – show participants five problem solving demonstrations, and (3) explicit feedback – tell participants if their answers are correct or not.

Although familiarization could be combined with both twelve and nine-gizmo demonstrations, we only tested its effect on participants who watched nine-gizmo demonstrations to limit the number of groups and participants.

1.2.1 Independent variables

Learning condition

All participants worked on twelve-gizmo problems, but were randomly assigned to one of five learning conditions, summarized in Table 1:

- Reinforcement learning: participants were given explicit feedback only.
- Imitation learning: participants watched five successful demonstrations of the twelve-gizmo task, and then worked on problems with no explicit feedback.
- Generalized imitation learning: participants watched demonstrations of a simpler variant of the problem involving nine gizmos, and then worked on problems with no explicit feedback.
- Delayed generalization learning: after familiarization with the task, participants watched demonstrations of the nine-gizmo task. To standardize the total time spent solving problems, they solved problems for 25 minutes after having watched demonstrations.
- Control: participants received no feedback or demonstrations.

Learning condition	Control	Reinforcement*	Imitation*	Generalized imitation	Delayed generalization
Familiarization	No	No	No	No	Yes
Demonstrations	None	None	12 gizmos	9 gizmos	9 gizmos
Explicit feedback	No	Yes	No	No	No

Table 1 – Experimental groups in the gizmo problem solving task
***: indicates groups replicated from (Dandurand, Bowen and Shultz 2004).**

This experiment was performed online using a Java applet. To avoid the potentially confounding effect of the method (lab vs. online), we collected new data even for conditions already run in the laboratory study (Dandurand, Bowen, and Shultz 2004).

Quartile

The second independent variable, called Quartile, was a repeated measure with 4 levels (1, 2, 3 and 4). Each quartile represents one quarter of all trials (e.g., quartile 2 includes trials 25% to 50%). We included this factor to study practice effects.

1.2.2 Dependent variables

There were two dependent variables: accuracy and elapsed time. Accuracy is the proportion correct responses. Elapsed time is the time, measured in seconds, taken to complete problem trials.

1.2.3 Hypotheses

Elapsed time

Consistent with previous work (Dandurand, Bowen, and Shultz 2004), we expected participants to get faster with practice as they reduce exploration and settle on routine, automated solutions.

Accuracy

We made four predictions about accuracy. First, if explicit feedback is useful, the reinforcement learning group should outperform the control group.

Second, if participants simply memorized the solutions shown, we predicted that the two generalization groups (who watched nine-gizmo demonstrations) would reach an accuracy of about 79% of the imitation group (see “Comparison of nine and twelve-gizmo solutions” section). By contrast, a higher accuracy would suggest that generalization participants understood something about demonstrations and did not merely memorize them.

Third, we predicted that if generalization participants memorized, they would reproduce a 3 vs. 3 strategy on the first weighing, but if they understood and correctly abstracted the selection rule, they should use a 4 vs. 4 strategy instead.

Finally, we hypothesized that the delayed generalization group would outperform the generalized imitation group because (1) during familiarization, participants may identify critical and difficult features of the problem, and (2) participants may pay more attention to those features while watching demonstrations and therefore learn more.

1.3 Procedure

To reduce variability typically associated with online experiments, participants were instructed to choose a room where they could avoid distractions.

After solving practice trials involving 3 gizmos and 2 uses of the scale, participants had to find the one gizmo that weighed more or less than the others in a set of 12 gizmos, using a scale at most 3 times. The computer program randomly selected a different target gizmo and weight on each trial. Participants worked on trials for 30 minutes. To begin with, all 12 gizmos were tagged as unknown weight. Problem solvers alternated between selecting gizmos to weigh and updating their labeling using the color selector tool (see Figure 1), until they found a solution, or used up the 3 allowed weighings. Participants used labeling to keep track of their hypotheses about gizmo weights. For example, gizmos installed on the side of the scale that moved down should be marked heavy or normal.

2. Results

2.1 Elapsed time

A log transformation was performed to improve the normality of elapsed times distribution. A two-way mixed ANOVA revealed a main effect of quartile, F (3, 285) = 41, $p < 0.001$, suggesting that participants got faster with practice (see Figure 5). More specifically, we found a linear trend, F (1, 95) = 77, $p < 0.001$, and a quadratic trend, F (1, 95) = 11, $p < 0.01$, suggesting that larger accelerations happened early in training, and leveled off with practice. Furthermore, we found no main effect of learning condition, F (4, 95) = 0.11, $p > 0.05$, or interaction of quartile with learning condition, F (12, 95) = 0.68, $p > 0.05$.

2.2 Accuracy

We applied an arcsine transformation to accuracies (proportions of correct answers) to stabilize variance (Hogg and Craig 1995). A two-way mixed ANOVA revealed a main effect of quartile, F (3, 285) = 4.5, $p < 0.01$, suggesting that participants got more accurate with practice (see Figure 6). More specifically, we found a linear trend, F (1,95) = 6.7, $p < .05$, and a quadratic trend, F (1,95) = 4.7, $p < .05$, suggesting that improvements in accuracy occur early in training, and level off with practice.

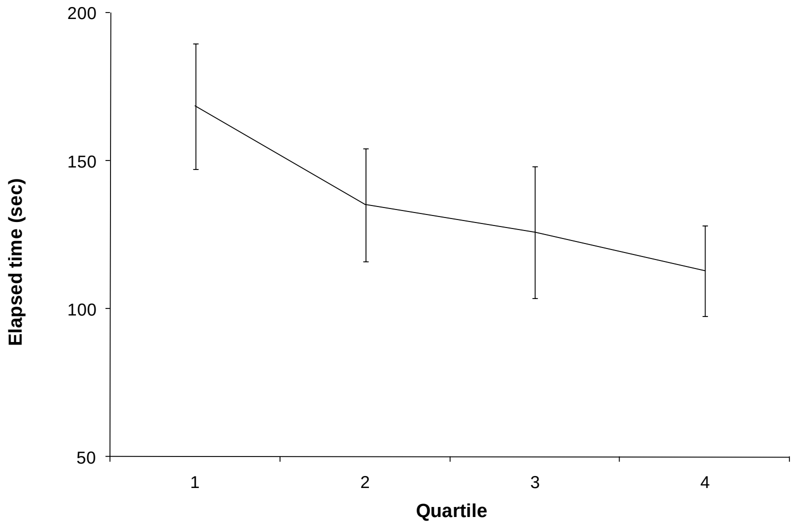

Figure 5 - Elapsed time as a function of quartile. Error bars represent standard errors (SE).

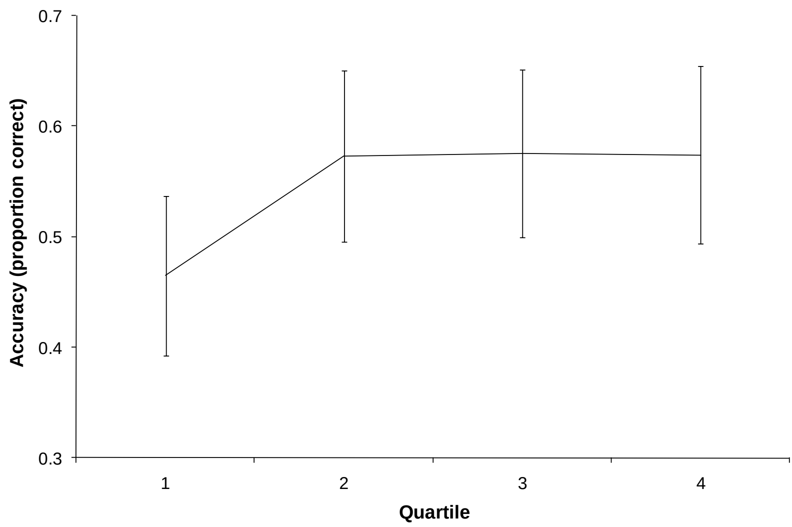

Figure 6 - Accuracy (proportion correct) as a function of quartile. Error bars represent standard errors (SE).

The ANOVA also revealed a main effect of learning condition, F (4, 95) = 3.41, p < 0.05, but no quartile by learning condition interaction, F (12, 285) = 1.70, p > 0.05. Average accuracy per learning group is presented in Figure 7.

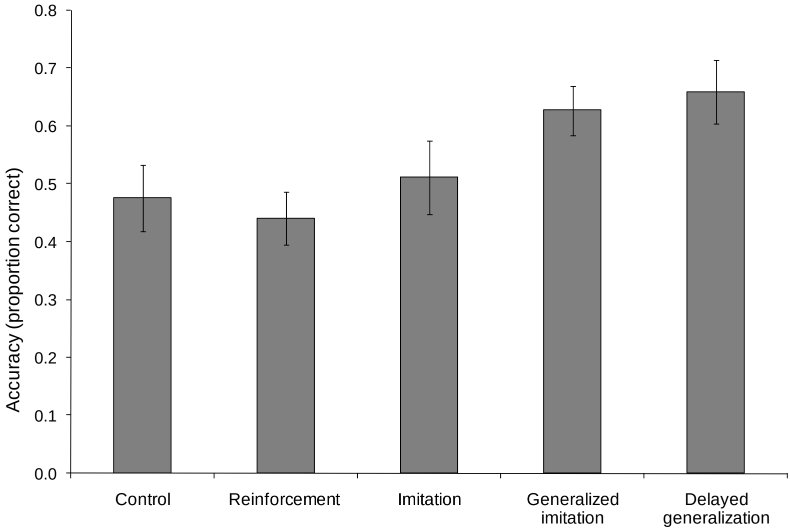

Figure 7 - Accuracy (proportion correct) per learning condition. Error bars represent standard errors (SE).

To further investigate how accuracy varied across learning conditions, we performed post-hoc tests. LSD post-hoc tests are shown in Table 2. Two large clusters emerge: (1) control, reinforcement and imitation groups, and (2) generalization and delayed generalization groups, consistent with a visual inspection of Figure 7.

	Control	**Reinforcement**	**Imitation**	**Generalization**	**Delayed Generalization**
Control					
Reinforcement					
Imitation					
Generalization		*			
Delayed Generalization	*	*	*		

Table 2 - Summary of LSD post-hoc tests. * indicates a significant difference between groups.

2.3 First weighing strategies used

We investigated what participants do on their first weighing. Because participants selected equal numbers of gizmos in 98.6% of their actions (the remaining fraction mainly caused by GUI manipulation errors), there are only 6 possible selections of gizmos, all of which are of unknown weight: 1 gizmo vs. 1 gizmo, 2 vs. 2, 3 vs. 3, 4 vs. 4, 5 vs. 5 and 6 vs. 6. Results are presented in Table 3.

Strategy	**Experimental group**				
Left / right gizmos installed	Control	Reinforcement learning	Imitation learning	Generalization	Delayed generalization
1/1	13.6%	5.1%	7.5%	0.3%	0.0%
2/2	6.2%	12.2%	7.1%	1.0%	0.9%
3/3†	19.9%	30.2%	17.8%*	28.3%*	26.9%*
4/4‡	49.6%	30.5%	63.9%*	67.2%*	64.5%*
5/5	2.7%	3.7%	2.5%	1.0%	6.0%
6/6	7.7%	18.3%	1.2%	2.0%	1.7%

Table 3 – Use of strategies on first weighing as a function of experimental group. †strategy shown to the generalization groups ‡ shown to the imitation learning group (correct strategy for the twelve-gizmo task). * indicate which values were used for the contingency table.

We tested for differences in frequencies of the 3 vs. 3 and 4 vs. 4 strategies in the imitation, generalization and delayed generalization groups using a 2x3 contingency table (see items labeled with * in Table 3). We found no frequency difference, suggesting endorsement rates of the correct strategy were equally high in the generalization groups and in the imitation leaning group ($\chi^2 = 4.23$, df = 2, $p > 0.05$).

3. Discussion

This study had two important goals. First, we wanted to see if explicit feedback is useful. Second, we asked if imitation learners could understand and generalize what they learned by watching demonstrations.

3.1 Is explicit feedback useful?

Participants in the reinforcement learning group were no more accurate (44.1%) than those of the control group (47.6%). Also, the absence of a learning group by quartile interaction suggests they did not get more accurate with practice than participants who received no explicit rewards. Finally, the significant main effect of quartile suggests that all participants, including those who got no explicit feedback, became more accurate from the first (46.5%) to the fourth quartile (57.4%). This is probably due to strategy improvements learned with practice. Together, these three results suggest that explicit feedback is not useful, perhaps because participants correctly reason about their solution accuracy.

Previous work with reinforcement learning simulations of this problem (Dandurand, Shultz and Rivest 2007) suggests other possibilities for why participants did not benefit from explicit feedback: reinforcement signals may be too infrequent (only once per 3 weighings) and not informative enough (binary information only: correct or not), and the problem space may be too large for rewards to be effective in the short time available.

3.2 Does imitation involve understanding?

We predicted that if memory is primarily involved, accuracy of the generalization groups should be about 79% of accuracy in the imitation group (79% x 51.1% = 40.4%). We also predicted that, based on the information content of the demonstrations, the imitation learning group would perform better than the generalization groups.

We found that accuracy was higher in the generalized imitation (62.8%) and the delayed generalization group (65.9%) than the imitation group (51.1%). The difference was not significant in spite of a large difference in percentage. This is probably due to the large sample variance. This unexpected result suggests that not only did generalization participants understand demonstrations, but also that other factors aside from information content may be involved. Perhaps participants in generalization groups perceived twelve-gizmo problems as novel, challenging, and interesting after watching nine-gizmo demonstrations. Indeed, humans are intrinsically motivated to perform novel and challenging activities (Ryan and Deci 2000). By contrast, the task may have appeared less interesting to the imitation learning participants because it was familiar. It is also possible that participants got more out of the nine-gizmo demonstrations because they were simpler and therefore easier to understand.

The memorizing hypothesis predicted that participants who watched nine-gizmo problems would reproduce the 3 vs. 3 strategy observed for the first weighing. In contrast, the understanding hypothesis predicted they would abstract the rule (install 1/3 of gizmos on each side of the scale), and thus correctly generalize to a 4 vs. 4 selection when solving twelve-gizmo problems. Results shown in Table 3 support the understanding hypothesis. The correct strategy was endorsed as highly in the two generalization groups (67.2% and

64.5%) as in the imitation group (63.9%). We also see that about a third of participants (28.3% and 26.9%) in generalization groups used a 3/3 strategy on the target problem, suggesting individual differences. Some participants may have relied more on memorization than others, but globally, participants appeared to have successfully generalized.

In short, abstraction of the correct rule for the first weighing, and high accuracy of the generalization groups are incompatible with predictions based solely on the memorizing hypothesis, and instead, support the understanding hypothesis.

3.3 Familiarization and elapsed time

Based on the results of our experiment, familiarization did not improve accuracy. However, improved experimental designs (e.g., using familiarization as a fully crossed factor) may be more sensitive and uncover an effect of familiarization. This crossed design would thus include a group who get familiarized with the task before watching twelve-gizmo problem demonstrations.

Finally, consistent with our hypothesis, participants became faster with practice, probably due to automatization of the strategies used.

3.4 Does generalization equal understanding?

We found evidence that participants generalized and abstracted correct rules by observing an expert solve a simpler problem variant, which we operationally defined as evidence for understanding. However, because understanding is an evasive concept, difficult to define and measure, our definition may be problematic.

Besides memorizing (which predicted a maximum of 79% accuracy), other simple cognitive mechanisms may explain why certain problem features were generalized, but arguably do not involve understanding. For example, watching demonstrations may have primed participants to use complex gizmo selections without understanding why correct solutions require those complex selections. However, priming cannot account for the correct abstraction of the 1/3 rule on the first weighing. Also, because the quality of the priming effects present in exact information should be at least as good as those of less relevant material, priming cannot explain why performance was better in the generalization groups than the imitation group.

We presented a simple view where memorizing and understanding are contrasted. We succeeded in showing that rote memorizing (and priming) could not account for the observed performance. However, it is possible that memorizing and priming are involved to some degree. In fact, anecdotal evidence suggests that most participants had to memorize certain particularly difficult solution steps. We also noticed important differences among participants' cognitive styles in the imitation learning group, some relying more on memorizing than others.

Although they have some limitations, think aloud protocols could shed light on the mechanisms involved in generalization. We may also be able to address more specific questions such as whether learning by demonstration involves insight, memorizing, reasoning, or priming.

3.5 Implications for social imitation

In this experiment, demonstrations were presented by a computer program and not a human being. Because imitation is inherently social, we can wonder how research done in artificial laboratory setting with computer-mediated demonstrations informs us about apprenticeship learning in naturalistic environments. The influential Turing Test rests on the assumption that humans can engage in social interactions through computer interfaces (Turing 1950). Not only has this been confirmed, but humans even tend to mistake computers for humans rather easily (Shieber 1994). Therefore, the cognitive processes used for learning by observation of computer-mediated demonstrations are likely similar to those involved in imitation in naturalistic settings.

In sum, participants improved with practice even without explicit feedback, suggesting they could evaluate their solution accuracy using reasoning alone, or that the feedback available was insufficient given the task difficulty. Participants also successfully generalized knowledge acquired by demonstrations, thus supporting the understanding hypothesis and discrediting the memorizing hypothesis. Surprisingly, participants performed even better when presented with demonstrations of a simpler task. Further research is needed to assess the role of novelty and motivation in learning by observation of problem solving tasks. These results show that imitation learning of human problem solving, as with previous studies with macaque monkeys (Subiaul et al. 2004) and human infants (Gergely, Bekkering, and Kiraly 2002), involves more than rote memorizing.

4. Acknowledgements

Kayo Nakamura and Kristine H. Onishi for their comments and suggestions.

The research was supported by a McGill Major scholarship to F. D. and a grant to T. R. S. from the Natural Sciences and Engineering Research Council of Canada.

References

Bandura, A. 1986. *Social foundations of thought & action*. Englewood Cliffs, New Jersey: Prentice-Hall.

Byrne, R. W., and A. E. Russon. 1998. Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences* 21 (05):667-684.

Carpenter, M., J. Call, and M. Tomasello. 2002. Understanding “prior intentions” enables two-year-olds to imitatively learn a complex task. *Child Development* 75 (3):1431-1441.

Shieber, S. M. 1993. Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery* 37 (6):70-78.

Dandurand, F., T. R. Shultz, and K. H. Onishi. 2007. Strategies, heuristics and biases in complex problem solving. In *the Proceedings of the twenty-ninth meeting of the Cognitive Science Society (CogSci 2007)*. New-York: Lawrence Erlbaum Associates.

Dandurand, F., T. R. Shultz, and F. Rivest. 2007. Complex problem solving with reinforcement learning. In *the Proceeding of the 6th IEEE International Conference on Development and Learning (ICDL-2007)*: IEEE.

Dandurand, F., M. Bowen, and T. R. Shultz. 2004. Learning by imitation, reinforcement and verbal rules in problem solving tasks. In *the Proceedings of the third International Conference on Development and Learning (ICDL'04)*. La Jolla, California, USA: The Salk Institute for Biological Studies, UCSD.

Gergely, G., H. Bekkering, and I. Kiraly. 2002. Rational imitation in preverbal infants. *Nature* 415:755.

Hogg, R. V., and A. T. Craig. 1995. *Introduction to Mathematical Statistics*. Upper Saddle River, NJ: Prentice-Hall.

Holyoak, K. J. 1995. Problem solving. In *Thinking: An Invitation to Cognitive Science, 2nd edition*, edited by E. E. Smith and D. N. Osherson. Cambridge, MA: MIT Press.

Katona, G. 1940. *Organizing and memorizing*. New York, NY: Columbia University Press.

Moreno, R., and R. E. Mayer. 2005. Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology* 97 117-128.

Newell, A., and H. A. Simon. 1972. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Reder, L. , and R. L. Klatzky. 1994. The effect of context on training: Is learning situated? Pittsburgh: Carnegie Mellon University, School of Computer Science.

Ryan, R. M., and E. L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55:68-78.
Subiaul, F., J. F. Cantlon, R. L. Holloway, and H. S. Terrace. 2004. Cognitive imitation in rhesus macaques. *Science* 305:407-410.
Turing, A. 1950. Computing machinery and intelligence. *Mind* 59:433-460.

---

[i] Out of 24 cases, 18 can be reliably identified, and the chance level for the remaining six is 1/6. Total expected accuracy is thus (18 + 6*(1/6))/24