For Teaching Perceptual Fluency, Machines Beat Human Experts

Ayon Sen (asen6@wisc.edu)[1], Purav Patel (ppatel47@wisc.edu)[2], Martina A. Rau (marau@wisc.edu)[2], Blake Mason[3], Robert Nowak[3], Timothy T. Rogers[4], Xiaojin Zhu[1]

[1] Department of Computer Sciences, [2] Department of Educational Psychology,
[3] Department of Electrical and Computer Engineering, [4] Department of Psychology
University of Wisconsin-Madison

Abstract

In STEM domains, students are expected to acquire domain knowledge from visual representations that they may not yet be able to interpret. Such learning requires perceptual fluency, or the ability to intuitively and rapidly see the underlying concepts in visuals and to translate between them. Perceptual fluency is acquired via nonverbal, implicit learning processes. Thus far, we have lacked a principled approach for identifying a sequence of perceptual fluency problems that promote robust learning. Here, we describe how a novel machine learning technique can generate an optimal sequence of perceptual fluency problems. In a human experiment, we show that a machine-generated sequence outperforms both a random sequence and a sequence generated by a human domain expert. Interestingly, the machine-generated sequence resulted in significantly lower accuracy during training, but higher posttest accuracy. This suggests that the machine-generated sequence induced desirable difficulties. To our knowledge, our study is the first to show that machine learning can yield desirable difficulties for perceptual learning.

Keywords: visuals; perceptual fluency; implicit learning; desirable difficulties; machine learning; machine teaching; chemistry; optimal training; sequence effects

Introduction

Visual representations are ubiquitous instructional tools in science, technology, engineering, and math (STEM) domains (Ainsworth, 2008; NRC, 2006). For example, chemistry instruction on bonding typically includes the visuals shown in Figure 1. While we typically assume that such visuals help students learn because they make abstract concepts more accessible, they can also impede students' learning if students do not know how the visuals show information (Rau, 2017). To successfully learn new domain knowledge from visuals, students need representational competencies — knowledge about how visual representations show information (Ainsworth, 2006). For example, a chemistry student needs to learn that the dots in the Lewis structure (Figure 1a) show electrons and that the spheres in the space-filling model in (Figure 1b) show regions where electrons likely reside.

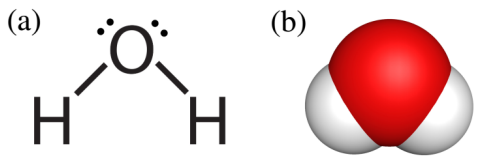

Figure 1: Two common visual representations of water (a: Lewis structure; b: space-filling model).

Instruction that helps students acquire representational competencies mostly focuses on conceptual representational competencies. These include the ability to map visual features to concepts, support conceptual reasoning with visuals, and choose appropriate visuals to illustrate a given concept (Bodemer, Ploetzner, Feuerlein, & Spada, 2004). Less research has focused on a second type of representational competency — perceptual fluency. It involves the ability to rapidly and effortlessly see meaningful information in visual representations (E. J. Gibson, 2000; Goldstone & Barsalou, 1998). For instance, chemists can effortlessly see that both visuals in Figure 1 show water. Perceptual fluency plays an important role in students' learning as it frees cognitive resources for higher-order complex reasoning, thereby allowing students to use visuals to learn new domain knowledge (Rau, 2017).

Students acquire perceptual fluency via implicit inductive processes (E. J. Gibson, 2000; Goldstone & Barsalou, 1998). Consequently, instructional interventions for perceptual fluency engage students in simple problems to quickly judge what a visual shows (Kellman & Massey, 2013). One kind of perceptual-fluency problem may ask students to quickly and intuitively judge whether two visuals like the ones in Figure 1 show the same molecule by using implicit intuitions. The problem sequence is typically chosen so that (1) students are exposed to a variety of visuals and (2) consecutive visuals vary incidental features while drawing attention to conceptually relevant features (Kellman & Massey, 2013; Rau, 2017). However, these general principles leave many possible sequences open. To date, we lack a principled approach capable of identifying sequences of visual representations that yield optimal learning outcomes. Hence, we used an inverse machine-learning technique that selects a sequence of visual representations that was most effective for a learning algorithm. In a human experiment, we then tested whether the machine-generated sequence of visual representations yielded higher learning outcomes compared to (1) a random sequence and (2) a sequence generated by a human expert based on perceptual learning principles.

In the following, we review literature concerning visual representations, perceptual fluency, and our inverse machine-learning paradigm. Then, we describe the methods we used to identify the machine-generated sequence, followed by the methods for the human experiment. We also discuss how our results may guide educational interventions for representational competencies and machine learning more broadly.

Prior Research

Perceptual Fluency

Representations used in instructional materials are defined as external representations because they are external to the viewer. By contrast, internal representations are mental objects that students can imagine and mentally manipulate. External representations can be symbolic (text) or visual (Lewis structures). Unlike symbolic representations, *visual representations* have similarity-based mappings to the referent (Schnotz, 2014).

Perceptual fluency research is based on findings that experts can automatically see meaningful connections among representations, that it takes them little cognitive effort to translate among representations, and that they can quickly and effortlessly integrate information distributed across representations (E. J. Gibson, 2000). Chemistry experts, for example, can see at a glance that the Lewis structure in Figure 1a shows the same molecule as the space-filling model in Figure 1b. Such perceptual expertise frees cognitive resources for explanation-based reasoning (Goldstone & Barsalou, 1998) and is considered an important goal in STEM education.

According to the cognitive theory of multimedia learning (CTML) and the integrated model of text and picture comprehension (ITCP), perceptual fluency involves efficient formation of accurate internal representations of visual representations (Mayer, 2009; Schnotz, 2014). Doing so requires mapping analog internal representations of multiple visual representations to one another (Mayer, 2009; Schnotz, 2014).

Cognitive science literature (E. J. Gibson, 2000; Goldstone, 1997; Koedinger, Corbett, & Perfetti, 2012) suggests that students acquire perceptual fluency via perceptual-induction processes. These processes are inductive because students can infer how visual features map to concepts through experience with many examples (E. J. Gibson, 2000; Goldstone, 1997; Kellman & Massey, 2013). Students gain *efficiency* in seeing meaning in visuals via perceptual chunking. Rather than mapping specific analog features to concepts, students learn to treat each analog visual as one perceptual chunk that relates to multiple concepts. Perceptual-induction processes are thought to be nonverbal because they do not require explicit reasoning (Koedinger et al., 2012). They are implicit because they happen unintentionally and sometimes unconsciously (Shanks, 2005).

Interventions that target perceptual fluency are relatively novel. Kellman and colleagues (2013) developed interventions that engage students in perceptual-induction processes by exposing them to many short problems wherein they have to rapidly translate between representations. Students might receive numerous problems that ask them to judge whether two visuals like the ones shown in Figure 1 show the same molecule. These interventions have enhanced students' learning in STEM domains like chemistry (Rau, Michaelis, & Fay, 2015). Critically, these interventions seek to determine whether perceptual fluency practice on a set of training problems generalizes to unfamiliar posttest problems.

Perceptual learning is strongly affected by the sequence in which problems appear (Rau, 2017). To design effective problem sequences, consecutive problems expose students to systematic variation (often via contrasting cases) so that irrelevant features vary while relevant features appear across several problems (Kellman & Massey, 2013). However, visual representations can differ on a large number of features. Thus, many possible problem sequences can systematically vary these visual features. We addressed this issue using Zhu's machine teaching paradigm (Zhu, 2015; Zhu, Singla, Zilles, & Rafferty, 2018).

Machine Teaching

Machine teaching, the inverse problem of machine learning, has been applied in fields including cognitive psychology and education (Patil, Zhu, Kopeć, & Love, 2014). It requires a cognitive model that takes the form of a learning algorithm. This algorithm mimics how human students learn a mapping between two visual representations (e.g., the ones shown in Figure 1). Given the cognitive model, machine teaching seeks a sequence of learning problems (optimal training sequence O), such that when given O, the learning algorithm learns the mapping. To evaluate whether a training sequence is effective, we test the cognitive model's performance at mapping visual representations using a different test set of perceptual fluency problems than were used in training. Typically, a set of training problems (known as training instances in machine learning) is drawn from a perceptual fluency training distribution (P_t). The set of test problems (known as test instances in machine learning) comes from a separate distribution (P_e). The goal is to minimize the test error rate on P_e. The goal of machine teaching then becomes:

$$O = \underset{S \in C_t}{\operatorname{argmin}} P_{(x,y)\sim P_e}\left(\mathcal{A}(S)(x) \neq y\right) \tag{1}$$

Here, C_t is the set of all possible training sequences generated from P_t and $\mathcal{A}(S)$ is the learned hypothesis after training on S. To properly construct the optimal training sequence O in this setting, we must understand (1) the nature of the to-be-learned domain knowledge and (2) the learning algorithm $\mathcal{A}$ used by the cognitive model. In this paper, the to-be-learned domain knowledge is a binary judgment of whether or not two molecules in different visual representations are the same. Further, we identified a cognitive model that mimics how humans learn these mappings. Our goal is to investigate whether, when the mappings and the cognitive model are well understood, machine teaching can identify a training sequence that is more effective than (a) an expert-chosen sequence based on perceptual learning principles and (b) a random sequence.

Cognitive Model

We now describe how we constructed the cognitive model that was used to optimize the training sequence. To this end, we describe the perceptual-fluency problems, how we

formally represented these problems, the learning algorithm used by the cognitive model, and finally how we used the cognitive model to identify the optimal training sequence.

Perceptual-Fluency Problems

Perceptual-fluency problems are single-step problems that ask students to make simple perceptual judgments. In our case, students were asked to judge whether two visual representations show the same molecule, as shown in Figure 2. Students were given two images. One image was of a molecule represented by a Lewis structure and the other image was a molecule represented by a space-filling model. Their task was to judge whether or not the two images show the same molecule.

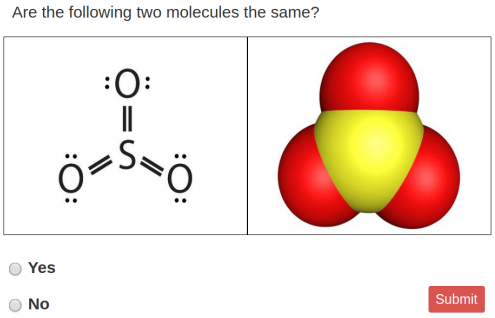

Figure 2: In this sample perceptual-fluency problem, students judged whether or not the Lewis structure and the space-filling model showed the same molecule. The answer is yes.

Visual Representation of Molecules

In our experiment, we used visual representations of chemical molecules common in undergraduate instruction. To identify these molecules, we reviewed textbooks and web-based instructional materials. We counted the frequency of different molecules using their chemical names (e.g., H_2O) and common names (e.g., water). We chose the 142 most common molecules. In order to formally describe the visual representations, we quantified visual features such as the number of lines or dots in the Lewis structure or the color of spheres in the space-filling models. To this end, we created feature vectors for each of the molecules (Figure 3) that describe which visual features the representation contains, as described in (Rau, Mason, & Nowak, 2016). Specifically, feature vectors of Lewis structures contained 27 features and feature vectors for space-filling models contained 24 features. These feature vectors were used by the learning algorithm.

Learning Algorithm

Learning was modeled using a feedforward artificial neural network (ANN) (Demuth, Beale, De Jess, & Hagan, 2014) that takes two feature vectors as input (corresponding to the two visual representations in the task) and is trained to output 1 when they represent the same molecule and 0 otherwise. To produce accurate predictions, the model must learn to generate internal representations that are proximal when the two

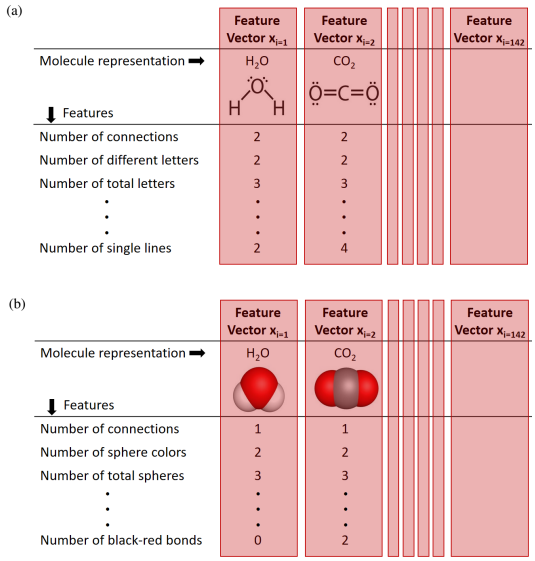

Figure 3: Example features for H_2O and CO_2 molecule representations with feature vectors in red (a: Lewis structure; b: space-filling model).

feature vectors depict the same molecule and distal when they depict different molecules. In this sense, the model captures the intuition from perceptual fluency theory that internal representations are used to discern the underlying similarity between different visual representations of the same structure. To this end, we included two separate subnetworks in the ANN learning algorithm (one for each input feature vectors), which is atypical for a general ANN structure. The subnetworks generated the internal representations for the two input feature vectors as discussed above. The model architecture is shown in Figure 4.

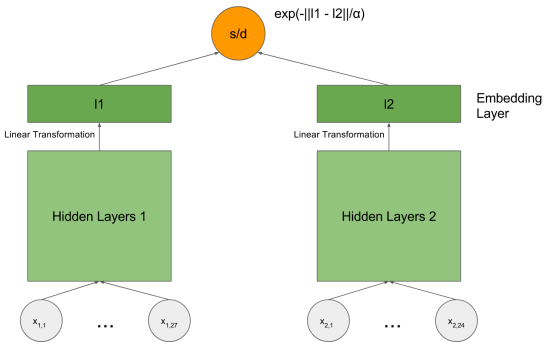

Figure 4: Neural network learning algorithm architecture.

Pilot Experiment to Train the Learning Algorithm

First, we needed to train the learning algorithm to mimic human perceptual learning. To this end, we conducted a pilot experiment to find a good set of hyperparameters. For cognitive models, good hyperparameters make predictions that match human behavior on the posttest. We matched the algorithm's predictions to summary statistics of human perfor-

mance on the posttest. For our pilot experiment, we recruited 47 undergraduate chemistry students. The series of problems they were provided was similar to the ones we describe later in the Human Experiment section. Specifically, they were provided with random training sequences generated from the training distribution P_t. We then used standard coordinate descent with random restart to find a good hyperparameter set. The hyperparameters that we tuned include learning rate, number of hidden layers and number of units in each layer.

Finding an Optimal Training Sequence

Next, we used the ANN learning algorithm to generate an optimal training sequence for the perceptual-fluency problems by solving Equation 1. We did so by searching over the space of all possible training sequences. We set the size of the candidate training sequences to 60, thereby aligning with prior perceptual learning research (Rau et al., 2015). We used a modified hill climbing algorithm to find an optimal training sequence. Hill climb search takes a greedy approach. Procedurally, we started with one particular training sequence. Then, we evaluated neighbors of that particular training sequence to determine whether a better one existed. If so, we moved to that one. This process stopped when no such neighbors are found. This search algorithm is defined with its states and neighborhood definition. The states of the search algorithm were any training sequence $S \in \mathcal{C}_t$ of size 60. Two training sequences were identified as neighbors if they differ by one problem. For computational efficiency, we restrict ourselves to only inspecting 500 neighbors for a given training sequence.

Human Experiment

To evaluate whether the optimal training sequence yields higher learning outcomes, we conducted a randomized experiment with humans.

Participants

We recruited 368 participants using Amazon's Mechanical Turk (MTurk) (Buhrmester, Kwang, & Gosling, 2011). Among them, 216 were male and 131 were female. The rest did not disclose their gender. Most participants were below the age of 45 (86%) and the largest number (192) fell in the age group $24-35$.

Test Set

To reiterate, our goal was to assess transfer of learning from the training sequence to a novel test set. Thus, we chose the problems from separate distributions. We randomly divided the 142 molecules we selected into two sets of 71 (training molecules, $\mathcal{X}_t$ and test molecules, $\mathcal{X}_e$). One set was used to create the test distribution and the other one was used to create the train distribution. The test distribution P_e is particularly important because our goal was to reduce humans' error rates on this distribution. The test distribution was created by the following procedure. $x_1 \sim p_1$, where p_1 is a marginal distribution on $\mathcal{X}_e$. p_1 is "importance of molecule x_1 to chemistry education" and was constructed by manually searching a corpus of chemistry education articles for molecule text frequency. With probability $1/2$, set $x_2 = x_1$ so that the true answer $y = 1$. Otherwise, draw $x_2 \sim p_2(\cdot \mid x_1)$. The conditional distribution p_2 is based on domain experts' opinion that favors confusable x_1, x_2 pairs in an education setting. Also note that $p_2(x_1|x_1) = 0, \forall x_1$. Taken together, $P_e(x_1,x_2) = \frac{1}{2}p_1(x_1)\mathbb{I}_{\{x_1=x_2\}} + \frac{1}{2}p_1(x_1)p_2(x_2 \mid x_1)$. Both the pretest and posttest judgment problems were sampled from this distribution across all conditions.

Experimental Design

We compared three training conditions. In the *machine-training sequence* condition, we used the training sequence O found by the search algorithm. For all $(x_1,x_2) \in O$ the corresponding true answer y was the indicator variable on whether x_1 and x_2 were the same molecule: $y = \mathbb{I}_{\{x_1=x_2\}}$. We presented x_1 and x_2 in Lewis and space-filling representation to human participants, respectively. Participants gave their binary judgment $\hat{y} \in \{0,1\}$. We then provided the true answer y as feedback to the participant. In the *human training sequence* condition, the training sequence was constructed by domain expert using perceptual learning principles. Specifically, an expert on perceptual learning sequences visuals constructed the sequence based on the contrasting cases principle (Kellman & Massey, 2013; Rau et al., 2015), so that consecutive examples emphasized conceptually meaningful visual features, such as the color of spheres that show atom identity or the number of dots that show electrons. The rest of this condition was the same as the machine training sequence condition. In the *random training sequence* condition, each training problem (x_1,x_2) was selected from the training distribution P_t with $y = \mathbb{I}_{\{x_1=x_2\}}$. The training distribution P_t for this condition was created in a similar way as the test distribution but on the training molecules. The rest of this condition was the same as the previous ones.

Procedure

We hosted the experiment on the Qualtrics survey platform (Qualtrics, 2005) and NEXT (Jamieson, Jain, Fernandez, Glattard, & Nowak, 2015). Participants first received a brief description of the study and then completed a sequence of 126 judgment problems (yes or no). Problems were divided into three phases. Phase one was the pretest and it included 20 test problems without feedback. Phase two was for training and it included 60 training problems with correctness feedback. We assumed that participants learned during this phase because they received feedback. Phase three was the posttest with 40 test problems displayed without feedback.

In addition, one *guard problem* was inserted after every 19 problems throughout all three phases. A guard problem either showed two identical molecules depicted by the same representation or two highly dissimilar molecules depicted by Lewis structures. We used these simple guard problems to filter out participants who responded haphazardly. During modeling, we disregarded all guard problems. When the two

molecules were shown to participants, the position (left/right) was randomized so that no representation was privileged.

Results

Of the 368 participants, we excluded 43 participants who failed any of the guard questions. The final sample size was $N = 325$. The final number of participants in the conditions random, human, and machine training sequence were 108, 117 and 100 respectively. Figure 5 reports accuracy on the pretest, training sequence and posttest by condition.

Effects of condition on training accuracy First, we tested whether training condition affected participants' training accuracy using ANCOVA. We found a main effect of condition on training accuracy, $F(2,321) = 18.8, p < .001, \eta^2 = .082$. Tukey post-hoc comparisons revealed that (a) the machine training sequence condition had significantly lower training accuracy than the human training sequence condition ($p < .001, d = -0.32$), (b) the machine training sequence condition had significantly lower training accuracy than the random training sequence condition ($p < .001, d = -0.26$), and (c) no significant differences existed between the human and random training sequence conditions ($p = .592, d = 0.05$).

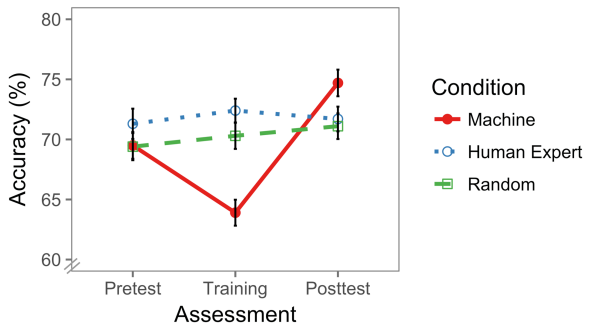

Figure 5: Means and standard errors by condition during each assessment phase.

Effects of condition on posttest accuracy Next, we tested whether training condition affected participants' posttest accuracy using ANCOVA. We found a main effect of condition on posttest accuracy, $F(2,321) = 5.02, p < .01, \eta^2 = .023$. Tukey post-hoc comparisons revealed that (a) the machine-training sequence condition had significantly higher posttest accuracy than the human training sequence condition ($p < .05, d = 0.16$), (b) the machine-training sequence condition had significantly higher posttest accuracy than the random sequence condition ($p < .05, d = 0.14$), and (c) no significant differences existed between the human and random training sequence conditions ($p = .960, d = -0.02$).

Discussion

Our goal was to investigate whether machine learning can help identify a sequence of visual representations that enhances students' learning from perceptual-fluency problems. To this end, we used machine teaching to reverse-engineer an optimal training sequence for a machine learning algorithm. Next, we conducted an experiment with humans that compared the machine teaching sequence to a random sequence and to a sequence generated by a human expert on perceptual learning. The machine teaching sequence resulted in lower training accuracy, but higher posttest accuracy.

These results significantly advance the perceptual learning literature. First, our results can inform the instructional design of perceptual-learning problems. Even though prior research yields principles for effective sequences of visual representations, numerous potential sequences can satisfy these principles. This study revealed how machine teaching can help solve this problem. Given a learning algorithm that constitutes a cognitive model of students learning a task, instructors can identify problem sequences that likely yield higher learning outcomes. Second, our results expand theory on perceptual learning. The fact that the machine learning sequence yielded lower performance during training, but higher posttest scores suggests that this sequence induced desirable difficulties (Soderstrom & Bjork, 2015).

Desirable difficulties refers to interventions yielding lower performance during training, but higher long-term learning. To explain this phenomenon, Soderstorm and Bjork (2015) proposed that more difficult learning interventions induce more active processing during training. This impedes immediate performance due to the increased difficulty but results in more durable memories and long-term learning. Our results suggest that the machine teaching approach was successful because it identified a training sequence that induced desirable difficulties. To our knowledge, our study is the first in which a machine-generated instructional intervention used desirable difficulties to support perceptual fluency.

This study also contributes to the machine learning literature. We provide empirical evidence that an ANN learning algorithm constitutes an adequate cognitive model of learning with visual representations. To our knowledge, the machine teaching paradigm has thus far only been applied to learning with artificial visual stimuli that vary on one or two dimensions (e.g. Gabor patches (B. R. Gibson, Rogers, Kalish, & Zhu, 2015)). Our study is the first to demonstrate that machine teaching can model and improve learning with high-dimensional visual representations like Lewis structures and space-filling models of chemical molecules.

Our findings were limited in several ways. First, the population of MTurk workers limits generalization to the target population of undergraduate chemistry students. MTurk workers have highly variable prior knowledge about chemistry. Second, the search algorithm we used to find an optimal training sequence did not test all possible training sequences of size 60. Exhaustively finding a solution is not practically feasible. Thus, we settled for a suboptimal training sequence that still yielded a small risk on the test distribution. Third, our study was constrained in the use of chemistry representations as stimuli. While we used more high-dimensional rep-

resentations than prior perceptual learning studies (B. R. Gibson et al., 2015), the complexity of our representations does not represent all realistic stimuli. Sparser and richer visuals exist and it is possible that machine teaching will yield greater benefits for these visuals.

Conclusion

Visual representations are used in many domains, but it can be cognitively demanding to learn from them. Perceptual fluency can help by freeing up cognitive resources for higher-order reasoning. Here, we tested a machine teaching technique for developing perceptual fluency in chemistry. The machine-generated optimal training sequence improved learning compared to a training sequence generated by a human expert who used perceptual-learning principles and compared to a random sequence. These results are promising, as they suggest that machine teaching can help create more effective sequences of perceptual-fluency problems. Given that visual representations are ubiquitous in STEM domains, we anticipate that our findings will be broadly useful.

Acknowledgement

This is supported in part by NSF grant IIS 1623605. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also thank Michael Mozer and Brett Roads (University of Colorado, Boulder) for their support and comments regarding the cognitive model.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and instruction*, *16*(3), 183–198.

Ainsworth, S. (2008). The educational value of multiple-representations when learning complex scientific concepts. *Visualization: Theory and practice in science education*, 191–208.

Bodemer, D., Ploetzner, R., Feuerlein, I., & Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learning and Instruction*, *14*(3), 325–341.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, *6*(1), 3–5.

Demuth, H. B., Beale, M. H., De Jess, O., & Hagan, M. T. (2014). *Neural network design*. Martin Hagan.

Gibson, B. R., Rogers, T. T., Kalish, C., & Zhu, X. (2015). What causes category-shifting in human semi-supervised learning? In *Cogsci.*

Gibson, E. J. (2000). Perceptual learning in development: Some basic concepts. *Ecological Psychology*, *12*(4), 295–302.

Goldstone, R. L. (1997). Perceptual learning. *San Diego, CA: Academic Press*.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*(2), 231–262.

Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. (2015). Next: A system for real-world development, evaluation, and application of active learning. In *Advances in neural information processing systems* (pp. 2656–2664).

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *The psychology of learning and motivation*, *58*, 117–165.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, *36*(5), 757–798.

Mayer, R. E. (2009). *Cognitive theory of multimedia learning*. The cambridge handbook of multimedia learning (2nd ed., pp. 31-48). New York, NY: Cambridge University Press.

NRC. (2006). *Learning to think spatially*. Washington, D.C.: National Academies Press.

Patil, K. R., Zhu, X., Kopeć, Ł., & Love, B. C. (2014). Optimal teaching for limited-capacity human learners. In *Advances in neural information processing systems* (pp. 2465–2473).

Qualtrics. (2005). *Qualtrics©2018.* `https://it.wisc.edu/services/surveys-qualtrics`, last visited 01-18-2018. Provo, Utah, USA.

Rau, M. A. (2017). Conditions for the effectiveness of multiple visual representations in enhancing stem learning. *Educational Psychology Review*, *29*(4), 717–761.

Rau, M. A., Mason, B., & Nowak, R. D. (2016). How to model implicit knowledge? similarity learning methods to assess perceptions of visual representations. In *Edm* (pp. 199–206).

Rau, M. A., Michaelis, J. E., & Fay, N. (2015). Connection making between multiple graphical representations: A multi-methods approach for domain-specific grounding of an intelligent tutoring system for chemistry. *Computers & Education*, *82*, 460–485.

Schnotz, W. (2014). An integrated model of text and picture comprehension. *In R.E. Mayer (Ed.), The Cambridge handbook of multimedia learning (2 ed., pp. 72-103); New York, NY: Cambridge University Press*.

Shanks, D. R. (2005). Implicit learning. *Handbook of cognition, London: Sage*, 202–220.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199.

Zhu, X. (2015). Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Aaai* (pp. 4083–4087).

Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). An overview of machine teaching. *arXiv preprint arXiv:1801.05927*.