COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13470

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13470

Teaching Without Thinking: Negative Evaluations of Rote Pedagogy

Ilona Bass,[a,b] Cristian Espinoza,[b] Elizabeth Bonawitz,[b] Tomer D. Ullman[a]

[a]*Department of Psychology, Harvard University*
[b]*Graduate School of Education, Harvard University*

Received 9 September 2023; received in revised form 11 April 2024; accepted 16 May 2024

Abstract

When people make decisions, they act in a way that is either automatic ("rote"), or more thoughtful ("reflective"). But do people notice when *others* are behaving in a rote way, and do they care? We examine the detection of rote behavior and its consequences in U.S. adults, focusing specifically on pedagogy and learning. We establish *repetitiveness* as a cue for rote behavior (Experiment 1), and find that rote people are seen as worse teachers (Experiment 2). We also find that the more a person's feedback seems similar across groups (indicating greater rote-ness), the more negatively their teaching is evaluated (Experiment 3). A word-embedding analysis of an open-response task shows people naturally cluster rote and reflective teachers into different semantic categories (Experiment 4). We also show that repetitiveness can be decoupled from perceptions of rote-ness given contextual explanation (Experiment 5). Finally, we establish two additional cues to rote behavior that can be tied to quality of teaching (Experiment 6). These results empirically show that people detect and care about scripted behaviors in pedagogy, and suggest an important extension to formal frameworks of social reasoning.

Keywords: Social reasoning; Pedagogy; Commonsense reasoning; Habitual behavior

1. Introduction

You may have had this experience: You are at an academic conference and attending your first talk of the day. At the end of the talk, an audience member asks an astute and piercing question. Impressed, you continue on to attend another presentation, and notice that same audience member in the crowd. At the end of this second talk, they ask a question—a very

The preregistrations, analysis scripts, study materials, and deidentified datasets generated and analyzed for these studies are publicly available in a repository on OSF: https://osf.io/uwycg.

Correspondence should be sent to Ilona Bass, Department of Psychology, Harvard University, 52 Oxford St., Cambridge, MA 02138, USA. E-mail: ibass@fas.harvard.edu

similar question, in fact, to the one they asked earlier. As the day progresses, you notice that this person is asking the same question again and again. What at first seemed like a person making a helpful point, engaged with the presentation's content, now gives the impression of an automaton, going through its preprogrammed motions. You find yourself making a mental note that perhaps you should not take this person's feedback too seriously if you ever come across it in the future.

While the example above is taken from academia, it highlights a fundamental, everyday distinction in human behavior, likely familiar to anyone. People can act in a rote and automatic way, or they can behave in a more reflective and thoughtful way. This dichotomy has been explored at length theoretically, empirically, and neurally, in a variety of decision-making frameworks and across species (Botvinick, 2012; Dickinson, 1985; Dolan & Dayan, 2013; Etkin, Büchel, & Gross, 2015; Kahneman, 2011; Liljeholm, Tricomi, O'Doherty, & Balleine, 2011). Both of these different decision strategies are rational, in that they make better sense in different contexts. Automatic, habitual responses are quick and efficient in familiar circumstances, but relatively inflexible to changes in the environment. Reflective, thoughtful responses are flexible and environment-contingent, but they are slower and more cognitively taxing (Keramati, Dezfouli, & Piray, 2011; Schneider & Shiffrin, 1977; Wood & Neal, 2007).

It is hard to overstate the reach and influence of the past work examining the dichotomy between thoughtful and rote action. However, it has mostly examined how people themselves act and think. Our focus here is instead on the opposite direction: Not on how people themselves act, but on how people think about, and interpret the actions of *others* as thoughtful or rote. Going back to the opening example, our interest is not in how the question-asker chose to act in a rote way, but how other people realized they were acting in this way, and how it changed their learning. This question has received very little attention, certainly compared to the examination of people's own decisions, but we propose it is both a ubiquitous and important mental computation.

Thinking about the mental states of others (goals, beliefs, emotions) from their actions is the domain of *theory-of-mind*, which has been the focus of many years of research in cognitive and developmental psychology (Baillargeon, Scott, & He, 2010; Premack & Woodruff, 1978; Tomasello, 2018; Wellman, Cross, & Watson, 2001; Wimmer & Perner, 1983). Recent formal frameworks of theory-of-mind reasoning explain how we may use people's behavior to infer their intentions and mental states (Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jern & Kemp, 2015). While these frameworks have been successful, they start from the assumption that other people's behavior is driven by goals, beliefs, and intentions. Yet, as shown by the decades of research mentioned above, many behaviors are not goal-driven, but automatic, habitual, or scripted. Few approaches to reasoning about the mental states of others have accounted for this (though see Gershman, Gerstenberg, Baker, & Cushman, 2016; Schank & Abelson, 1977, for some notable exceptions). Our work expands on the theory-of-mind approach to reasoning about others, and asks: Do people make inferences about whether others' actions are driven by contemplative thought or automatic scripts, and what are the consequences of such inferences for social reasoning, decision-making, and learning?

In this paper, we examine people's reasoning about rote and reflective behavior in others, in the context of informal pedagogy—the kind of setting described in our opening example. Pedagogy is a natural domain for examining reasoning about automatic behavior for at least three reasons. First, it is a common setting in which a speaker (presenter, teacher, pedagogue) is under the competing pressures of practicing and engaging. That is, teaching others often entails a tension between *being* prepared and rehearsed, but *appearing* engaged and off-the-cuff. Moreover, speakers often have to simultaneously communicate information to many listeners at once, while also attempting to engage with individuals' learning goals and needs.

Second, as highlighted by the opening example, there are reasons to think that the detection of rote or automatic behavior has important consequences for how a listener will learn and attend. Past research in education supports this intuition. Students learn more readily from more engaged teachers (Ahn, Chiu, & Patrick, 2021; Roth, Assor, Kanat-Maymon, & Kaplan, 2007; Skinner & Belmont, 1993; Wentzel, 2009; Wigfield, Cambria, & Eccles, 2012), and students who perceive their teachers as caring have better academic outcomes (Gasser, Grütter, Buholzer, & Wettstein, 2018; Smart, 2014). It stands to reason then that if people detect that another person is using automatic processes, they will disengage and learn less.

Third, beyond empirical work, there are existing formal models of pedagogical reasoning that can both illuminate the inference of rote behavior and benefit from its empirical demonstration. Models of pedagogical reasoning provide a framework for understanding the cognitive processes that underlie learning. These models propose a recursive process, in which knowledgeable, helpful teachers should present the information that would be most efficient in conveying a concept; students in turn should update their beliefs based on the observed evidence, with the assumption that the teacher is knowledgeable (Bonawitz, Shafto, Yu, Gonzalez, & Bridgers, 2020) and has actually selected the information in a helpful, intentional way (Bonawitz & Shafto, 2016; Bonawitz et al., 2011; Shafto, Goodman, & Griffiths, 2014). The recursive mental-state reasoning (e.g., "I know that you know that I know...") that underlies these models provides a theoretical framework for understanding why engagement with a teacher might be important for learning: A person who is not engaged with the beliefs, needs, and goals of another person will be less likely to select the best possible evidence for them. However, the extent to which students might be more or less likely to learn from rote, scripted teaching approaches also creates something of a puzzle for these models: These models ultimately predict that the quality of teachers should be determined by the quality of the evidence they present. Rote teachers may very well present information that is indistinguishable from what an engaged teacher would have presented. So, if it is empirically shown that learning is negatively affected by the perception of rote behavior independent of the evidence, it means significant and important amendments are needed in our current computational models of pedagogy.

We note that there has been a host of recent work examining different aspects of reasoning about other people's thinking in a way that goes beyond simple theory-of-mind inference, and which relates to our interests here, but is distinguished from it. For example, recent work by Hawkins, Gweon, and Goodman (2021) has used extensions of the Rational Speech Acts framework (Goodman & Frank, 2016a) to model less-than-ideally-informative speakers that do not perfectly weight the perspective of listeners in a visual perspective-taking task, and the

consequent behavior of listeners. They showed that "scripted" speaker utterances were seen as less informative than those produced naturally, and that listeners adjusted their behavior as a result. Importantly, this framework and task is suited for a perspective-taking situation in which both speaker and listener share the burden of perspective-taking, distinct from the pedagogy models concerning us here. Second, the (reasonable) response of a listener to an uninformative speaker, as shown by the models and experiments, is to take more of the burden on themselves. We are interested in scripted situations, particularly in pedagogy, that lead the learner to disengage from the situation. Beyond this work, a recent preprint (Berke, Tenenbaum, Sterling, & Jara-Ettinger, 2023) has considered inference in a theory-of-mind framework in which the observer infers the amount of mental effort another person puts into pursuing their goals, in particular accounting for situations in which another person may be distracted (daydreaming) or relying on a known solution when solving a puzzle in a game of rush-hour. We see that work as a step in the right direction, but one that still models other agents as starting from the point of pursuing goals under some beliefs, as opposed to enacting scripts and habits using other decision- making modules, and without demonstrating the negative consequences of this inference in pedagogy.

We also note that a host of other work beyond Berke et al. (2023) has examined people's reaction times as an indication of underlying preferences and goals. For example, Gates, Callaway, Ho, and Griffiths (2021) used an inverse drift-diffusion model to capture the inferred strength of preference from revealed reaction times (quickly choosing an apple over an orange indicates a much higher preference for the apple over the orange). Going beyond this, Konovalov and Krajbich (2023) noted that people may be sensitive to the fact that their own decision times can reveal the strength of their preferences, and might use this strategically to obscure their preferences in certain situations. This work, however, still uses the notion that other people are using rational planning and decision-making (going from beliefs and goals, costs and rewards to actions), rather than the idea that rote behavior may not rely on goals and beliefs in this way, as we assume here. It also does not consider the potential negative downstream consequences for inferring that another agent is being rote.

Here, we ask two empirical questions: (1) Are people sensitive to whether teachers are acting in a rote way? And, assuming that they are, (2) How does the perception of a person as rote influence evaluations of their teaching? We primarily operationalized rote behavior using *repetitiveness*, where more repetitive feedback to different people should indicate more rote reasoning. We note that repetitiveness is neither a necessary nor sufficient cue for rote reasoning. Listeners use both the content and features of others' speech, such as disfluencies and prosody, to make inferences about the subject matter (Arnold, Kam, & Tanenhaus, 2007; Heller, Arnold, Klein, & Tanenhaus, 2015; Xie, Buxó-Lugo, & Kurumada, 2021) or the speaker's mental states (Fox Tree, 2002; Kidd, White, & Aslin, 2011; Loy, Rohde, & Corley, 2017) and teaching goals (Bascandziev, Shafto, & Bonawitz, 2021; see also Goodman & Frank, 2016b). Reaction time may also serve as a cue to how much cognitive effort a speaker is exerting (Richardson & Keil, 2022). While we leverage some of these additional cues in Experiment 6, repetitiveness serves as a reasonable starting point for exploring this phenomenon, as has been suggested by past work (Gershman et al., 2016), and is illustrated by the intuitions set up by our opening example: An audience member that asks an astute

question seems insightful; an audience member that asks the exact same question over and over across many different talks seems like a marionette.

In six preregistered experiments, we showed participants videos of a person providing feedback to students in a classroom setting. In Experiments 1–5, this feedback varied with respect to its repetitiveness across different groups of students; in Experiment 6, we varied the teacher's attentiveness (Experiment 6A) and use of verbal disfluencies (Experiment 6B). Experiment 1 verifies whether people associate repetitiveness with rote reasoning. In Experiments 2 and 3, we ask participants to make a variety of judgments about people exhibiting rote or reflective behavior. Experiment 4 investigates whether people naturally cluster rote and reflective teachers into distinct semantic categories. Experiment 5 asks whether repetition can be decoupled from perceived automaticity (e.g., if an explanation is provided for the teacher's repetitiveness). In Experiment 6, we tie automaticity to two additional behavioral cues: attentiveness and speech disfluencies. We expected that participants would indeed recognize when others are acting in a rote way, and that perceptions of a person as rote would broadly be associated with more negative evaluations of their teaching. While it may seem intuitive to associate rote behavior with a negative evaluation in general, we highlight that there are many contexts in which it might be expected or acceptable for a social partner to be acting automatically (e.g., conversational scripts, as when ordering in a restaurant). Further, pedagogical contexts are a particularly consequential context for the inference of rote behavior. Our goal, then, is to provide preliminary evidence for the detection of rote behavior—an intuitive yet understudied aspect of commonsense social reasoning—leveraging informal pedagogy as an especially relevant setting, and examining the expected consequences for learning.

2. Experiments

All experiments were approved by the Harvard University Institutional Review Board. Participants provided written informed consent before beginning the studies. Methods were carried out in accordance with relevant guidelines and regulations. For each experiment, we collected a small pilot sample to estimate effect sizes. Final sample sizes were determined based on these estimated effect sizes, with the goal of reaching 90% power. All aspects of our experiments—including sample sizes, analysis plans, study materials, and inclusion criteria—were preregistered with the Open Science Framework before beginning data collection.

2.1. Participants and materials

For Experiments 1 through 4, participants were convenience samples of adults recruited from Amazon Mechanical Turk via CloudResearch, which has built-in screening tools for assuring higher data quality. For Experiments 5 and 6, we moved away from Mechanical Turk (due to reasons outlined in Peer, Brandimarte, Samat, & Acquisti, 2017), and participants were instead recruited through the Prolific platform (https://www.prolific.com). Responses were submitted online using Qualtrics surveys.

We created three videos of a person (Teacher) providing feedback to three different groups of students on in-class projects they were ostensibly working on. The Teacher approached

Condition	Feedback Content (paraphrased)
Identical	F1: "Expand the last section a bit more" F2: "Expand the last section a bit more" F3: "Expand the last section a bit more"
Similar	F1: "Expand the last section a bit more" F2: "The last section needs more content" F3: "Flesh out the points in the last section"
Unique	F1: "Expand the last section a bit more" F2: "Balance the use of text and images" F3: "Make sure the introduction is clear"

Fig. 1. Screenshots of video stimuli and description of conditions for Experiments 1 through 5. Participants watched a video of a teacher providing feedback to three different groups of students. This feedback was either identical, similar, or unique across the three groups. The above images are of the actors in our video stimuli and not of study participants. Informed consent and express permission to publish this image was obtained from all actors.

each group one at a time and looked at their project for about 5 s. She then provided a single piece of feedback before moving on to the next group. The feedback the Teacher provided across groups was either *Identical*, *Similar*, or *Unique*; see Fig. 1. Because the actors in these videos wore masks, all participants saw exactly the same visuals with different audio dubbed on top, depending on the condition to which they had been randomly assigned. Each video was about 40 s long. Although all of the experiments described below (except Experiment 6) used these same videos, we wanted to ensure that our results could not be explained by effects of priming or other potential confounds. So, we ran separate experiments with unique samples of participants. Crucially, across all the videos used in our experiments, participants could not actually see the projects that the students were working on. This was intentional, as we wanted to ensure that the actual appropriateness or relevance of the feedback for the projects was equally ambiguous across conditions. (In the General Discussion, we address how relevance of feedback might trade off with perceptions of a teacher's automaticity to influence teaching evaluations and learning.)

The preregistrations, analysis scripts, study materials, and deidentified datasets generated and analyzed for these studies are publicly available in a repository on OSF: https://osf.io/uwycg.

2.2. *Experiment 1*

In Experiment 1, we ask whether people use repetitiveness as a cue that another person's behavior is automatic.

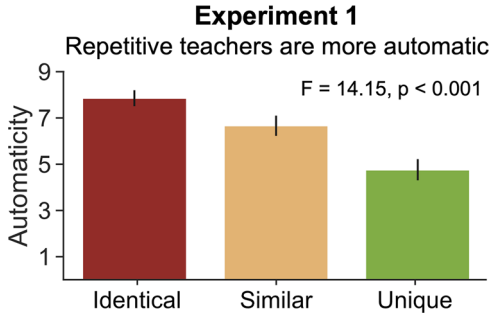

Fig. 2. Results from Experiment 1. This figure shows average automaticity ratings across conditions; error bars represent +/− the standard error. Teachers who provide repetitive feedback across learners are perceived as more automatic than teachers who provide unique feedback.

2.2.1. Method

Sixty participants were randomly assigned to view one of the three videos described above ($N = 20$ in the Identical condition, $N = 20$ in the Similar condition, $N = 20$ in the Unique condition, between-subjects). Five additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 4$) or technical difficulties experienced during the task ($N = 1$). After watching the video, participants were given a working definition of automaticity: *"We are interested in people's thoughts about automatic behavior. By automatic, we mean behavior that appears robotic or rehearsed, as though the person is following a script and not thinking deeply."* Participants rated how automatic they thought the teacher was on a 1–9 Likert scale (1 = not at all automatic, 9 = completely automatic).

One might be concerned that this definition could lead participants to a negative interpretation of the notion of automaticity. As noted in the introduction, there are many contexts in which it might be expected or acceptable for a social partner to act automatically. Similarly, terms like "rehearsed," "robotic," or "not thinking deeply" may have differently valenced connotations depending on the social context. A barista that is acting robotically and not thinking deeply in response to a coffee order is probably to be expected, and it would be strange for them to do otherwise. So, for the description of automaticity presented to participants in this task, we opted for a fuller explanation of what we meant by scripted behavior.

2.2.2. Results

A one-way ANOVA revealed significant differences between conditions ($F(2, 57) = 14.15$, $p < .001$, $\eta^2 = 0.33$): Participants thought the teacher in the Identical video was more automatic than the teacher in the Unique video, with the Similar teacher falling in between (see Fig. 2). These results establish that people use repetitiveness as a cue for inferring whether others are relying on rote, automatic reasoning processes, setting the stage for the next experiments.

2.3. Experiment 2A

In Experiment 2, we asked how the perception of rote behavior influences evaluations of teaching. In our main experiment (Experiment 2B), we presented participants with one of the same three videos from Experiment 1 and asked them to evaluate the teacher on several pedagogy metrics. However, we first wanted to ensure that the individual pieces of feedback that the teacher provides are not in themselves biased or leading, so that any significant effects we might see in the main study can be attributed to our experimental manipulation of repeating the feedback. So, in Experiment 2A, we ran a pre-test control measure to validate that the different pieces of feedback provided across conditions are not seen as differentially helpful.

2.3.1. Method

For this experiment, we split the three videos from Experiment 1 (Identical, Similar, Unique) into individual clips of the teacher providing feedback to only one group of students in isolation, as opposed to three groups consecutively. The first piece of feedback was identical across all three conditions, while the second and third pieces of feedback were not, resulting in seven unique clips. (See Fig. 1: F1 was identical across videos, while F2 and F3 were different across videos. So, the seven clips were F1, F2 + F3 from the Identical video, F2 + F3 from the Similar video, and F2 + F3 from the Unique video.) A new group of 140 participants were randomly assigned to view one of these seven clips ($N = 20$ per clip). Ten additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 6$) or technical difficulties experienced during the task ($N = 4$). After watching their assigned clip, participants were asked five questions about the quality of the teacher: (1) *How good was this teacher in this interaction?* (2) *How good is this teacher in general?* (3) *How much will the students learn?* (4) *How much will the projects be improved?* (5) *How much was the teacher really thinking about each group?* Each of these questions was on a 1–9 Likert scale. These were the same five evaluative questions that we would ask participants in the main experiment (see Experiment 2B).

2.3.2. Results

In line with our preregistered analysis plan, we ran a one-way ANOVA on each of the evaluative measures by clip. As predicted, we found no significant differences in any of the ratings across clips ($F(6, 133) \leq 1.83$, $p \geq .10$). In general, the feedback was seen as moderately helpful (grand mean across all clips and questions = 6.38).

2.4. Experiment 2B

The results from Experiment 2A reveal that the different pieces of feedback in our videos are not differentially useful or informative per se. In Experiment 2B, we ask whether this feedback is perceived as less helpful when it is repeated across learners.

2.4.1. Method

A new group of 60 participants were randomly assigned to view one of the same three videos from Experiment 1 ($N = 20$ in the Identical condition, $N = 20$ in the Similar

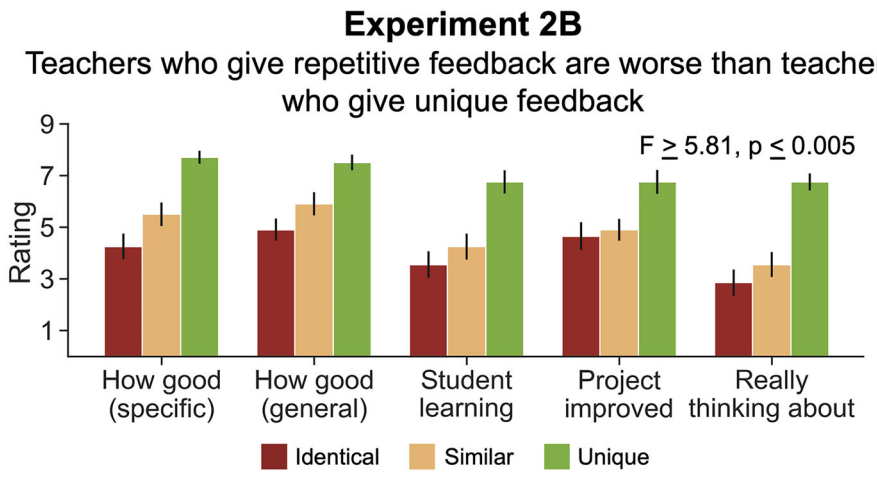

Fig. 3. Results from Experiment 2B. This figure shows average ratings for all five evaluations across conditions; error bars represent +/− the standard error. Teachers who provide repetitive feedback are rated worse along all five pedagogy metrics.

condition, $N = 20$ in the Unique condition, between-subjects). No participants were dropped and replaced in this experiment. After watching the video, participants were asked the same five evaluative questions described in Experiment 2A.

2.4.2. Results

A series of one-way ANOVAs revealed that the teacher who provided identical feedback was perceived as worse across all five dimensions (see Fig. 3). Participants thought the Identical teacher was less helpful, both in the specific interactions observed in the videos ($F(2, 57) = 17.48$, $p < .001$, $\eta^2 = 0.38$) and in general ($F(2, 57) = 10.86$, $p < .001$, $\eta^2 = 0.28$), and that they were thinking about each group less ($F(2, 57) = 21.69$, $p < .001$, $\eta^2 = 0.43$). Participants also thought the students paired with the Identical teacher would learn less ($F(2, 57) = 11.84$, $p < .001$, $\eta^2 = 0.29$), and that the quality of the projects they were working on would suffer as a result ($F(2, 57) = 5.81$, $p = .005$, $\eta^2 = 0.17$). The Similar teacher fell between the other two conditions on all metrics.

In addition to the main results above, we ran an exploratory analysis that directly compared the findings of Experiments 2A and 2B. To do this, we computed the average of all five evaluations to create an overall quality score for each participant in both experiments. We ran three independent samples t-tests comparing these scores from each condition in Experiment 2B ($N = 20$ per condition) with scores for the clips that comprised each of those videos in Experiment 2A ($N = 60$ per set of 3 clips). Because these analyses were not preregistered, we set $\alpha = 0.015$ to correct for family-wise error. We found that ratings of the Identical video in Experiment 2B ($M = 4.04$) were significantly lower than ratings of its component clips in Experiment 2A ($M = 6.26$; $t(78) = 4.96$, $p < .001$, $d = 1.21$). Similarly, ratings of the Similar video in Experiment 2B ($M = 4.82$) were significantly lower than ratings of its component clips in Experiment 2A ($M = 6.53$; $t(78) = 4.07$, $p < .001$, $d = 0.99$). However, ratings of

the Unique video in Experiment 2B ($M = 7.09$) did not differ from ratings of its component clips in Experiment 2A ($M = 6.39$; $t(78) = 1.77$, $p = .08$, $d = 0.47$). These results highlight a key point regarding the evaluation of the teacher's feedback itself: The exact same feedback is perceived as helpful when given once (Experiment 2A), but less helpful when it is repeated across learners (Experiment 2B). This lends preliminary evidence to the idea that evaluation of a teacher's quality could be negatively affected by the perception of rote behavior independent of the evidence.

2.5. Experiment 3

Experiment 2 established that rote people are seen as worse teachers, but this was a discrete distinction. In Experiment 3, we examine the connection between perceptions of rote behavior and evaluations of teaching quality more quantitatively. The Similar condition had the teacher give feedback that was similar across student groups, but not identical (see Fig. 1). The degree to which such feedback is perceived as unique was then more open to individual interpretation than in the other two conditions, allowing us to examine individual differences that directly link perceptions of repetitiveness to rote-ness.

2.5.1. Method

Forty new participants were assigned to the Similar condition only. Two additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 1$) or technical difficulties experienced during the task ($N = 1$). After watching the video, participants first provided the same five evaluations as in Experiment 2. The reliability among these various metrics of quality was quite high ($\alpha = 0.928$); in line with our pre-registration, we averaged these evaluations to create an "overall quality" score for each participant. Then, participants separately rated how similar they thought the teacher's feedback was across groups (also on a 1–9 scale).

2.5.2. Results

We found that similarity judgments were significantly and negatively related to the overall quality scores ($r(38) = -.327$, $p = .039$, $R^2 = .11$): The more similar participants thought the feedback was, the worse the teaching was rated. See Fig. 4. Importantly, all participants in this study saw the teacher give exactly the same set of feedback. Thus, subtle differences in perceptions of exactly how similar the feedback was (and so, how automatically the teacher was behaving) may be quantitatively related to differences in perceptions of the teacher's quality.

We note that these findings may be driven by the data points in the upper-left quadrant of Fig. 4; that is, relatively few participants gave similarity ratings lower than 8 (9/40), but those who did also rated the teacher more favorably overall. We did not preregister any outlier exclusion procedures, because we wanted to capture the full range of variability in participants' responses. One interpretation of our findings is that the connection between perceived feedback similarity and teacher evaluations may be directional, such that *less* similar feedback is linked with *better* teaching (as opposed to *more* similar feedback being linked with

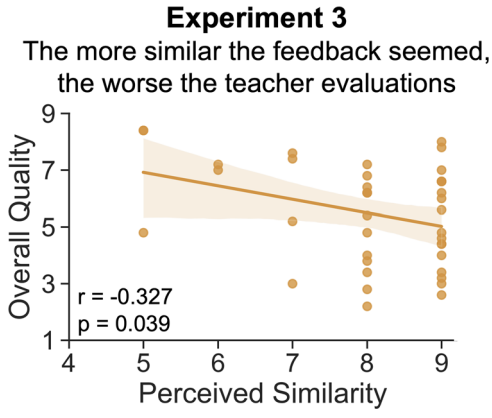

Fig. 4. Results from Experiment 3. This figure shows each participant's similarity rating of the feedback (x-axis) and overall quality rating of the teacher (y-axis), with a best-fit linear trend line and bootstrapped 95% confidence interval of the estimate. The more similar participants thought a person's feedback was, the worse they evaluated that person's teaching.

worse teaching). While we do not pursue this data pattern further in this paper, it would be an interesting question for future work.

2.6. Experiment 4

Experiments 1–3 show converging evidence that people associate repetitiveness with automatic behavior, and that perceived automaticity is related to worse evaluations of teaching. However, the previous experiments restricted participants' response categories. It is possible that when freely describing the behavior of teachers, our provided categories and evaluations are not the ones that come naturally to people. To investigate the spontaneous associations people make with rote and reflective teaching behaviors, participants in Experiment 4 described each of the teachers in an open-response task. We examined the semantic space of the words generated in the different conditions.

2.6.1. Method

One-hundred fifty participants viewed one of the same three videos as in Experiments 1 and 2 ($N = 50$ in the Identical condition, $N = 50$ in the Similar condition, $N = 50$ in the Unique condition, between-subjects). Seventeen additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 10$) or technical difficulties experienced during the task ($N = 7$). After watching the video, participants were asked to list five words they would use to describe the teaching they observed. We corrected any misspellings before proceeding to analysis.

2.6.2. Results

We first examined the valence of the words participants provided. We used the HuggingFace pipeline abstraction to the distilbert-base-uncased-finetuned-sst-2-english model, classifying each term as positive or negative. We found that the Unique teacher was described

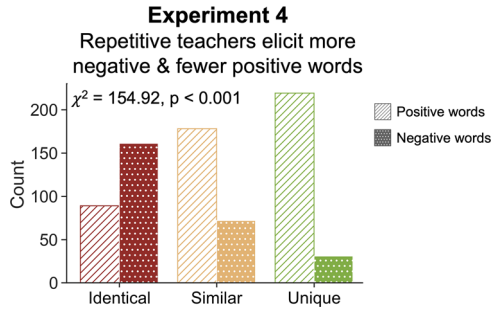

Fig. 5. Results from Experiment 4. This figure shows the frequency of participants' words that were categorized as positive and negative across the three conditions. Participants generated more negative words and fewer positive words for rote teachers.

more positively than the Identical teacher, with the Similar teacher falling in between ($\chi^2(2, N = 750) = 154.92$, $p < .001$, $V = 0.45$; see Fig. 5).

Next, we examined the similarity of participants' words to two preregistered groups of words that were representative of the semantic concepts in which we were interested: a "Rote" word cluster (*automatic, robotic, scripted, rehearsed, rote, reflexive, thoughtless, mechanical*), and a "Reflective" word cluster (*attentive, considerate, careful, reflective, engaged, thoughtful, contemplative, spontaneous*). We used Semantic Projection (Grand, Blank, Pereira, & Fedorenko, 2022) to locate each participant-generated term in a multidimensional vector space using the BERT language model through the SentenceTransformers package with the paraphrase-MiniLM-L6-v2 pretrained model. We applied the same technique to the eight terms in each word cluster, and then calculated a centroid for each cluster by taking the mean of every dimension over all eight component terms. For visualization, we applied tSNE to flatten every word vector (and the two centroid vectors) into a point in 2D space. Then, we drew a vector between the two centroids to create a feature subspace for "automaticity," visualized in Fig. 6A. We calculated the cosine similarity between each participant-generated term and both centroids (Rote and Reflective). The average cosine similarity between the words people gave and the Rote cluster was highest in the Identical condition and lowest in the Unique condition ($F(2, 147) = 4.41$, $p = .014$, $\eta^2 = 0.06$). In contrast, similarity to the Reflective cluster was highest in the Unique condition and lowest in the Identical condition ($F(2, 147) = 10.13$, $p < .001$, $\eta^2 = 0.12$; see Fig. 6B). These results suggest a natural and spontaneous rote-versus-reflective distinction in people's semantic space when observing relevant behavior.

2.7. Experiment 5

The findings reported so far show that teachers who provide repetitive feedback across learners are seen as acting more automatically, and that people associate such repetitiveness with worse teaching and distinct rote/reflective semantic spaces. However, we cannot rule out the possibility that participants' negative evaluations of the Identical teacher simply reflected a sensitivity to *repetitiveness* in pedagogy, and not automaticity per se. That is, it is possible

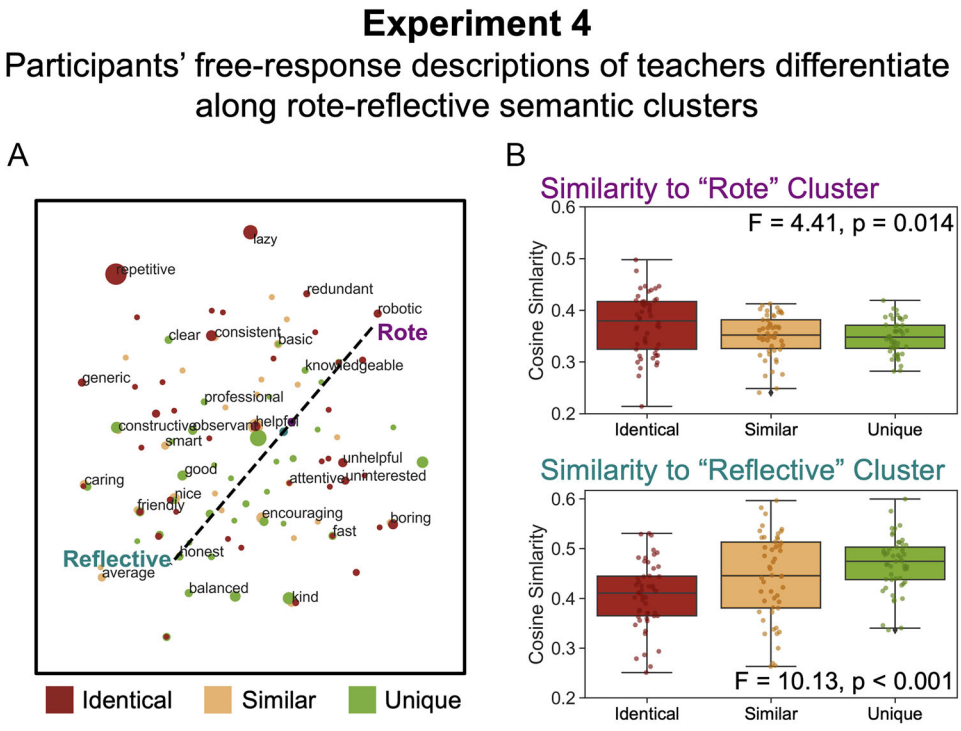

Fig. 6. Results from Experiment 4. (A) A feature subspace for "automaticity." The size of a dot corresponds to the frequency of the word. (B) Boxplots showing the cosine similarity between the words participants generated in each condition and the Rote cluster (top) and the Reflective cluster (bottom). Participants who saw the rote teacher generated words that were more semantically similar to the rote keyword cluster, while those who saw the reflective teacher generated words that were closer to the reflective keyword cluster.

that there is a simple and more direct link between repetition and pedagogy, such that repeated behavior is seen as worse in and of itself. To address this concern, we wanted to show that (1) being repetitive does not always signal automatic behavior, such that if repetition is explained in a nonautomatic way, the downstream consequences will change; and (2) one can be seen as automatic without repetition, which will in turn lead to the same downstream consequences as repetition, because automaticity is the important factor. We address point (1) in Experiment 5, by investigating whether explaining away a teacher's repetition can mitigate perceptions of automaticity and negative teaching evaluations. Point (2) is addressed in Experiment 6.

2.7.1. Method

Two-hundred and three participants viewed the Identical video from the previous experiments ($N = 101$ in the Pretext condition, $N = 102$ in the Control condition, between-subjects; conditions differences are described below). Sixty-seven additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 28$ in the Pretext condition, $N = 37$ in the Control condition) or technical difficulties experienced during the task ($N = 2$).

After watching the video, participants rated the teacher's automaticity (using the same question as in Experiment 1) and quality (using the same five questions as in Experiments 2 and 3). The order in which these evaluations were made was counterbalanced across participants. Then came the primary experimental manipulation. After answering all six questions, participants in the Pretext condition were given the following explanation for the teacher's repetition:

> "It is important that you have some additional context about the video you just saw. On the day the students got their group project assignment, the Teacher was sick and was not able to come to class. Instead, the project was given to the students by a Substitute Aid (not shown in the video). When assigning the project to the students, the Substitute Aid was not clear about which parts of the project the students should focus on. The Substitute Aid was also not clear about how the projects should be structured. The video you saw took place during the following class session. The Teacher, now recovered, returned to class and provided the students feedback on their current progress on the group projects (assigned by the Substitute Aid)."

In the Control condition, participants were given an unrelated text prompt. Content was provided in this control text so that we could conduct memory and attention checks in the control condition paralleling the Pretext condition. This allowed us to ensure that an approximately equal number of participants in each condition would be included via passing the attention checks. (Indeed, the number of participants dropped due to attention check failures did not differ across conditions: $\chi^2(1, N = 268) = 0.88$, $p = .35$.) The text prompt read:

> "Thank you for your attention to the videos and questions so far. It is important that the people who participate in our online research studies read all of the instructions carefully. We will now give you some information that you will need to remember at the end of the study. In this study, we are interested in understanding people's thoughts about teaching. To study this, we show people like you videos of teaching scenarios. The results of this study will inform the field of social cognition. Towards the end of this study, we will ask you three questions: 1) What are we studying in this task? 2) How do we study this? 3) What field will this study inform? Please select the answers thoughts about teaching, showing videos, and social cognition in that order."

Participants then watched the video a second time and provided the same ratings again (automaticity and quality).

As in Experiment 3, we computed the reliability among the five quality questions at Time 1 and Time 2. Reliability was high at both times ($\alpha \geq 0.929$), so we once again averaged these evaluations to create an "overall quality" score for each participant at Time 1 and at Time 2.

2.7.2. Results

If it is possible to explain away the teacher's repetition, we should see more favorable evaluations of the teacher at Time 2 in the Pretext condition than in the Control condition.

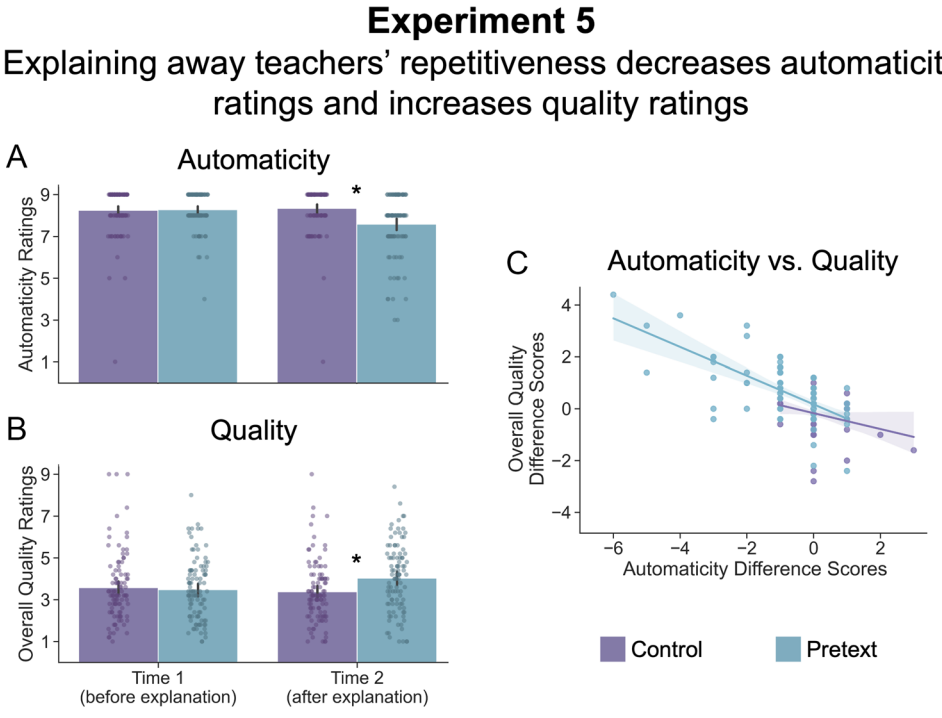

Fig. 7. Results from Experiment 5. (A) Automaticity ratings at Time 1 and Time 2 across conditions. Ratings were lower at Time 2 in the Pretext condition than in the Control condition. (B) Overall quality ratings at Time 1 and Time 2 across conditions. Ratings were higher at Time 2 in the Pretext condition than in the Control condition. (C) This figure shows each participant's difference score (Time 2 rating – Time 1 rating) for automaticity (x-axis) and overall quality (y-axis). The more participants decreased their automaticity ratings of the teacher from Time 1 to Time 2, the more they increased their quality ratings.

Indeed, at Time 1, automaticity and quality ratings did not differ across conditions ($p \geq .66$). At Time 2, however, participants in the Pretext condition rated the teacher as less automatic ($t(201) = 4.12$, $p < .001$, $d = 0.578$; see Fig. 7A) and of higher quality ($t(201) = 3.01$, $p = .001$, $d = 0.423$; see Fig. 7B). Providing an explanation for the teacher's repetition thus disentangled the cue of repetition from perceptions automaticity per se.

We also wanted to know whether changes in participants' automaticity ratings from Time 1 to Time 2 were connected to changes in their quality ratings, at the individual level. That is: after a second viewing of the video, if a participant thought the teacher was less automatic than they did at Time 1, did they also tend to think she was a better teacher? We correlated difference scores (i.e., Time 2 ratings – Time 1 ratings) for automaticity and overall quality, both within and across conditions. All three of these correlations were significant and negative (Pretext: $r(99) = -.678$, $p < .001$, $R^2 = .460$; Control: $r(100) = -.248$, $p = .012$, $R^2 = .061$; combined: $r(201) = -.653$, $p < .001$, $R^2 = .427$; see Fig. 7C). These results once again highlight the tight link between perceptions of automaticity and evaluations of teaching at the individual level.

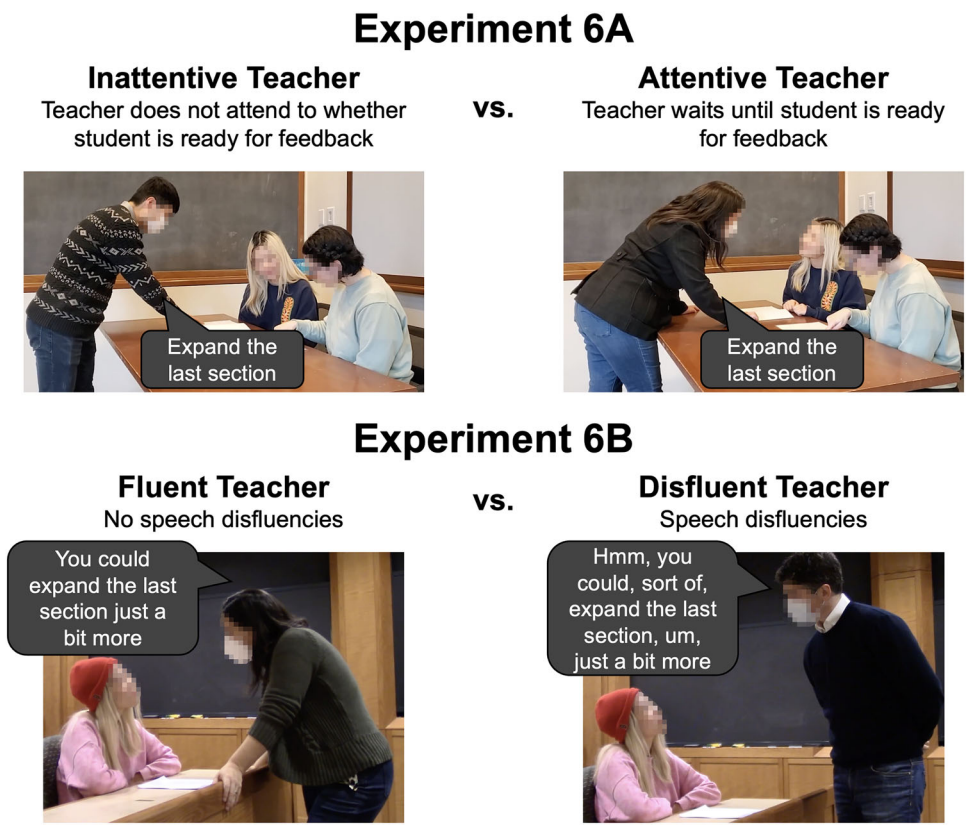

Fig. 8. Screenshots of video stimuli and description of conditions for Experiment 6. Participants watched two videos of two different teachers providing feedback to a student. In Experiment 6A, the teacher was either attentive or inattentive to whether the student was ready to receive the feedback. In Experiment 6B, the feedback was delivered either with or without speech disfluencies.

2.8. Experiment 6A

Experiment 5 demonstrates that it is not always automatic to be repetitive. In Experiment 6, we aim to show that one can be perceived automatic in absence of repetition. To that end, we ask whether other behavioral cues can be linked to inferences about automaticity in pedagogical contexts. We leverage two such cues: attentiveness (Experiment 6A) and speech difluencies (Experiment 6B), both of which extend naturally from past work (e.g., Heller et al., 2015; Hawkins et al., 2021).

2.8.1. Method

In order to investigate additional cues that could be tied to automaticity, we created new video stimuli that manipulated qualities of the teacher's feedback other than its repetitiveness across students. For Experiment 6A, we focused on the teacher's *attentiveness* to whether or not a student was prepared to receive feedback (see Fig. 8, top). We created videos of two

different teachers providing feedback to a student. In both videos, the teacher approached the student and looked at their project for about 5 s. They then provided a single piece of feedback (which was the same across teachers). The "Inattentive" teacher provided the feedback without looking up at the student and attending to whether they were ready to receive the feedback. The "Attentive" teacher looked up at the student and waited until they were ready before providing the feedback. Participants were presented with both of these videos, within-subjects. We counterbalanced which actor played which teacher and the order in which the videos were presented across participants.

One-hundred and fifty-three participants viewed the Attentive and Inattentive teacher videos. Four additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 3$) or technical difficulties experienced during the task ($N = 1$). After watching each video, participants rated the quality of the teacher they had just seen using the same five evaluative questions from the previous experiments. Finally, participants judged the relative automaticity of the two teachers by answering the question, *Which teacher was more automatic?* This question was on a 1–6 Likert scale, from "1st Teacher was much more automatic" to "2nd Teacher was much more automatic." For analysis, we reverse-coded ratings from participants who saw the Inattentive teacher second. This way, across counterbalancing orders, ratings of 1–3 represent the Inattentive teacher being rated as more automatic, and ratings of 4–6 represent the Attentive teacher being rated as more automatic.

As in the previous experiments, reliability was high among the five evaluative questions for both teachers ($\alpha \geq 0.928$). So, we averaged these evaluations to create an "overall quality" score for each teacher by participant.

2.8.2. Results

Our primary question was whether participants discerned a difference in the automaticity of the Attentive and Inattentive teachers. We investigated this in two ways. First, we compared the average automaticity rating across participants ($M = 3.14$, $SD = 1.44$) to the midpoint of the scale (3.5) by one-sample t-test, which was significant ($t(152) = 3.12$, $p = .002$, $d = 0.252$; see Fig. 9A). So, people found the Inattentive teacher to be more automatic than the Attentive teacher. We also performed a binomial test comparing the number of people who said the Inattentive teacher was more automatic (i.e., provided ratings of 3 or lower) to chance (50%). We found that 92 out of 153 participants thought the Inattentive teacher was more automatic, which is significantly more than would be expected due to chance ($p = .015$, two-tailed). People thus used the teacher's attentiveness as a cue to their automaticity.

We also wanted to know whether there was a quantitative connection between the relative automaticity and quality of the teachers. To do this, we first calculated difference scores between overall quality ratings for the two teachers for each participant (i.e., Attentive – Inattentive), such that positive scores represent the Attentive teacher being rated more favorably, and negative scores represented the Inattentive teacher being rated more favorably. We correlated these quality difference scores with automaticity ratings (where lower values represent the Inattentive teacher being rated as more automatic, and higher values represent the Attentive teacher being rated as more automatic). We found that these two measures were significantly and negatively correlated with one another ($r(151) = -.602$, $p < .001$, $R^2 = .363$;

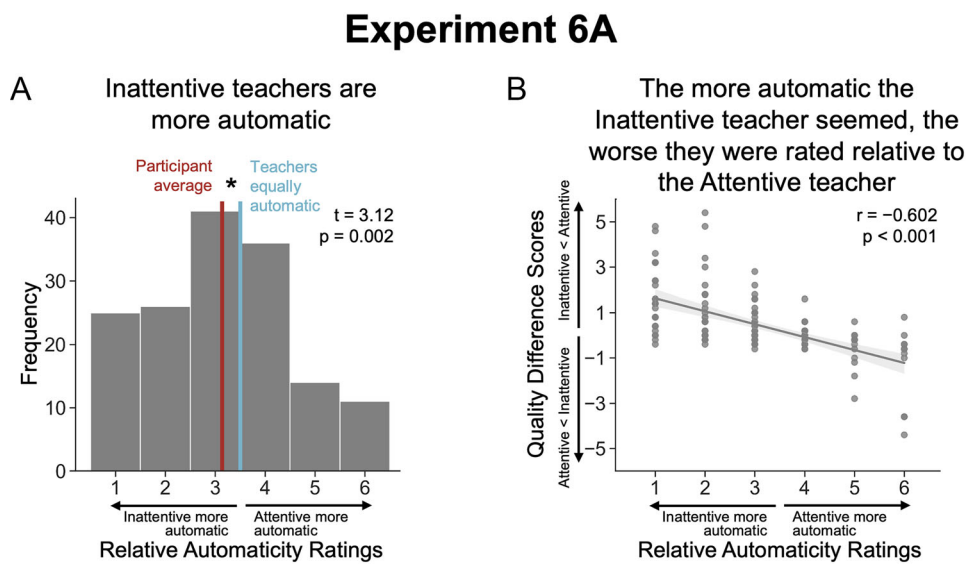

Fig. 9. Results from Experiment 6A. (A) A histogram of participants' relative automaticity ratings of the two teachers. The average rating across participants (red line) was significantly lower than the midpoint of the scale (blue line), meaning participants thought the Inattentive teacher was more automatic than the Attentive teacher. (B) This figure shows each participant's relative automaticty rating (x-axis) and the difference between their overall ratings of the two teachers (Attentive – Inattentive). The more automatic participants thought the Inattentive teacher was, the worse the Inattentive teacher was rated relative to the Attentive teacher.

see Fig. 9B). That is, to the degree that participants thought the Inattentive teacher was acting more automatically than the Attentive teacher, they also rated the Inattentive teacher more negatively. Overall, we also found that participants thought the Inattentive teacher was worse than the Attentive teacher ($t(152) = 3.72$, $p < .001$, $d = 0.300$). This suggests that attentiveness is both a cue to better teaching as well as a cue to less-automatic teaching. That automaticity directly correlated with teacher evaluation scores demonstrates the independent contribution of automaticity inferences on teaching scores above and beyond attentiveness.

2.9. Experiment 6B

For Experiment 6B, we explored *speech disfluencies* as a possible cue to automaticity (or lack thereof).

2.9.1. Method

As in Experiment 6A, we once again created new videos of two different teachers providing feedback to a student. In both videos, the teacher approached the student and looked at their project for about 5 s. They then provided a single piece of feedback (which was the same across teachers). The "Fluent" teacher's feedback did not contain speech dislfuencies, while the "Disfluent" teacher's feedback did (see Fig. 8, bottom). Participants were presented with

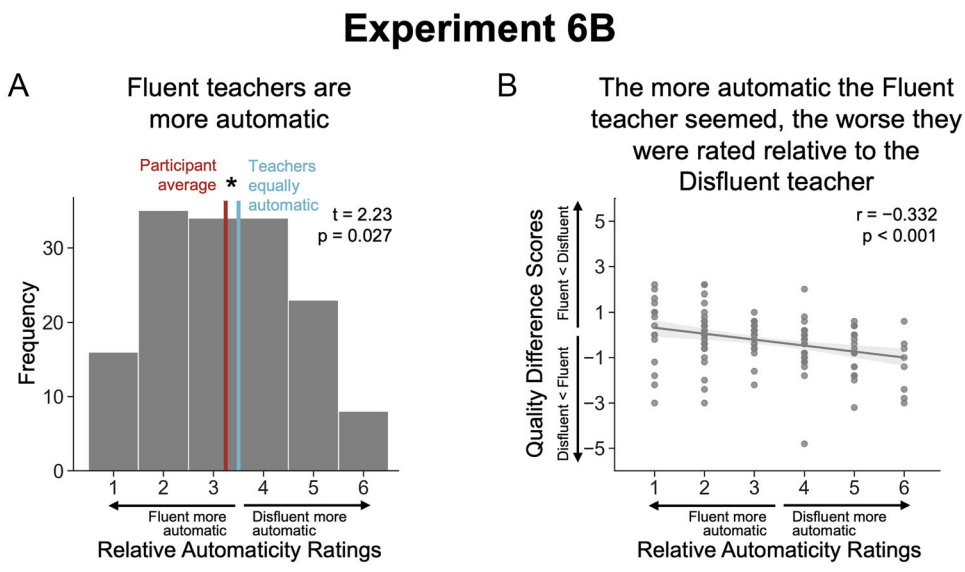

Fig. 10. Results from Experiment 6B. (A) A histogram of participants' relative automaticity ratings of the two teachers. The average rating across participants (red line) was significantly lower than the midpoint of the scale (blue line), meaning participants thought the Fluent teacher was more automatic than the Disfluent teacher. (B) This figure shows each participant's relative automaticty rating (x-axis) and the difference between their overall ratings of the two teachers (Disfluent – Fluent). The more automatic participants thought the Fluent teacher was, the worse the Fluent teacher was rated relative to the Disfluent teacher.

both of these videos, within-subjects. We counterbalanced which actor played which teacher and the order in which the videos were presented across participants.

One-hundred and fifty participants viewed the Fluent and Disfluent teacher videos. Six additional participants were dropped and replaced due to failure to pass built-in attention check questions ($N = 3$) or technical difficulties experienced during the task ($N = 3$). Other than the use of these new videos, the method was identical to Experiment 6A. Participants rated the quality and relative automaticity of both teachers. We reverse-coded automaticity ratings from participants who saw the Fluent teacher second, so that ratings of 1–3 represent the Fluent teacher being rated as more automatic, and ratings of 4–6 represent the Disfluent teacher being rated as more automatic.

Again, reliability was high among the five evaluative questions for both teachers ($\alpha \geq 0.898$), so we averaged these ratings to create an "overall quality" score for each teacher by participant.

2.9.2. *Results*

We first investigated the relative automaticity of the two teachers. The average automaticity rating across participants ($M = 3.25$, $SD = 1.39$) was significantly lower than the midpoint of the scale (3.5) by one-sample t-test ($t(149) = 2.23$, $p = .027$, $d = 0.182$; see Fig. 10A). This means that participants rated the Fluent teacher as more automatic overall (although a

binomial test on the number of participants who rated the Fluent teacher as more automatic was not significant: $85/150$, $p = .121$).

We also related automaticity ratings to overall quality ratings. We again calculated difference scores between overall quality ratings for the two teachers for each participant (i.e., Disfluent – Fluent), such that positive scores represent the Disfluent teacher being rated more favorably, and negative scores represented the Fluent teacher being rated more favorably. We correlated these quality difference scores with automaticity ratings (where lower values represent the Fluent teacher being rated as more automatic, and higher values represent the Disfluent teacher being rated as more automatic). As in Experiment 6A, we again found that these two measures were significantly and negatively correlated with one another ($r(148) = -.332$, $p < .001$, $R^2 = .111$; see Fig. 10B). So, to the extent that individual participants thought the Fluent teacher was behaving more automatically, they were also evaluated as a worse teacher.

Interestingly, the Fluent teacher was rated more favorably overall than the Disfluent teacher ($t(149) = 3.11$, $p = .002$, $d = 0.254$). This aligns with past work showing that verbal disfluencies are associated with uncertainty, both in people's own behavior (Smith & Clark, 1993) and in inferences about others (Brennan & Williams, 1995; Swerts & Krahmer, 2005). Also, certainty and knowledgeability are hallmark qualities of good teachers in the developmental literature (e.g., Jaswal & Malone, 2007; Sabbagh & Baldwin, 2001). Given that we see knowledgeability and automaticity as independent factors in our framework, it is not surprising that fluency can have an effect in two independent ways: it can indicate that a person is reasoning automatically (and to the degree that it does, it should lead to poorer teacher evaluations); but it can also indicate that a person is certain and confident (which should lead to better teacher evaluations). This can exactly lead to the results here, where overall fluency indicates knowledge and so increases ratings on average, but it can independently cue automaticity, and (most importantly for our focus) we find that to the degree that fluency signals automaticity, there is a negative correlation between relative automaticity and quality ratings. This aligns with the independent contribution of automaticity inferences on the evaluation of teaching.

3. General discussion

Do we notice when another person is "really there"? How do we tell? And why does it matter to us? We showed people are sensitive to, and care about whether other people are behaving in a rote way, in the everyday context of pedagogy. We established that people consistently make inferences about automatic behavior in others (both when prompted, and when giving free-form responses), that repetition, inattentiveness, and a lack of speech disfluencies are cues to automaticity, and that automatic-seeming people are perceived as worse teachers, along several dimensions. These results are the first to show that people detect and care about scripted behaviors in teaching, adding to a growing body of work on people's reasoning about the cognitive processes underlying others' actions in pedagogical contexts (Bass et al., 2022).

Our findings provide a proof-of-concept of people's sensitivity to rote behavior in others, setting the stage for future work to build out this research program in a number of exciting

directions. For one, the participants in our study were third-party evaluators of pedagogical interactions, and not actually learning from these teachers themselves. So, whether people actually learn information less effectively when it is transmitted by an automatic source of teaching is yet unknown, but has real-world implications: The rapid advancement of our technologies goes hand in hand with its possible use in the classroom. Educational tools, particularly in the era of remote-learning, are moving toward scripted and automated teaching, including asynchronous learning, prerecorded lectures, virtual classrooms, and app-based design. If students naturally pick up on this automation, and stop caring when they notice it, learning outcomes may well deteriorate. It is crucial then for future work to explore how inferences about automaticity influence learning outcomes.

In our video stimuli, we intentionally left the ground-truth needs of each of the students ambiguous. As a result, a possible interpretation of the findings from Experiments 1 through 3 is that participants assumed the students in the videos were likely not making exactly the same mistakes, and the teacher's repetition indicated a lack of *relevance* to the differing needs of the groups, as opposed to automaticity per se. Furthermore, in Experiment 5, when we explained away the teacher's repetition, we necessarily also provided a reason to believe that the students might require the same feedback (i.e., a sub-par initial teaching of the material from a substitute aid)—and indeed, evaluations of the teacher's repetition became more positive as a result. A lack of pedagogical relevance is one crucial way in which rote reasoning could lead to poor teaching in practice: A teacher who is not engaged with a student's beliefs, needs, and goals will be less likely to select the best possible evidence for them. We also know from a wealth of prior work that a teacher's accuracy plays an important role in how they are evaluated as informants (e.g., see: Harris, Koenig, Corriveau, & Jaswal, 2018; Tong, Wang, & Danovitch, 2020, for recent reviews). However, the results of the word-embedding analysis from Experiment 4, which finds that repetitiveness is specifically associated with a "rote" semantic space, cannot be fully explained by a feedback relevance account; nor can the results of Experiments 6A and 6B, because we operationalized rote-ness using cues other than repetitiveness. Given this, we believe that both automaticity and a lack of feedback relevance are important, but distinct constructs. An exciting direction for future work would be to more fully disentangle them, in order to understand how they interact to affect learning outcomes.

The extent to which *children* are sensitive to automatic behavior in others is also an open question. Studying the developmental trajectory of this inference will be important for at least three reasons. First, it will provide insight into whether the detection of rote behavior is part of an intuitive class of social reasoning with early developmental origins (Csibra & Gergely, 2009; Gerstenberg & Tenenbaum, 2017; Wellman & Gelman, 1992). Second, it could show how inferences about automatic behavior in others relate to the development of other processes such as theory-of-mind, linguistic development, or executive function. Third, studying children is a prerequisite for connecting the current body of work to formal education. The current findings ideally prepare future work to investigate both development and learning outcomes.

Another important future direction of this research is in integrating the inference of rote behavior into formal frameworks of social reasoning, including theory-of-mind and especially models of pedagogy (e.g., Jara-Ettinger et al., 2016; Shafto et al., 2014). Current

theory-of-mind frameworks are often predicated on the premise that other people are *utility maximizers*, in that they have goals, constraints, and beliefs, and choose actions to maximize rewards and minimize costs. We suggest that a large part of reasoning about others does not take others to be planning meaningfully in this sense. While the findings are intuitive, for many current models of pedagogy, it is difficult to explain why the rote teachers in our task (who provided ostensibly helpful information) are seen as worse. A particularly fruitful future direction would integrate existing theory-of-mind models with *resource rational* frameworks (Lieder & Griffiths, 2020) to account for cases in which agents may fall back on less cognitively taxing rote reasoning strategies (and see also Berke et al., 2023).

In proposing to expand theory-of-mind models to go beyond current frameworks, we do *not* mean to suggest that cost and reward calculations do not enter into reasoning about other people's rote-behavior, scripts, and habits. Rather, in line with the great deal of work on decision-making, it is likely that people can make a meta-decision of whether to deploy model-based, goal-belief-action planning, or to fall back on scripts. This meta-decision is based implicitly on the presumed cost of one option versus the other, weighed against the degree to which one cares about a social partner. Such a decision can also be made strategically, to signal to a partner that one is willing to incur the higher cost of model-based planning, in order to maintain or communicate the value of a relationship—and again (important for our present purposes), such a signal is only useful if it can be picked up by a person reasoning in the inverse direction. In line with this overall suggestion, recent frameworks of social affiliation suggest that even infants make inferences about other people's social attitudes, relationships, and goals, based on their willingness to incur personal costs for the benefit of social partners (Davis, Carlson, Dunham, & Jara-Ettinger, 2023; Powell, 2022). If reflective engagement with students is seen as incurring a higher cognitive cost on the behalf of others, this could help explain why automatic-seeming teachers are evaluated more negatively, even when the evidence they present is the same.

We focused on the negative implications of the perception of rote behavior, but there are likely many contexts in which rote behavior is expected, such that deviating from it may even be seen as odd and cause social friction. Past work (Gershman et al., 2016; Kahneman, 2011; Schank & Abelson, 1977) has shown cases in which rote, automatic reasoning is rational for a decision-maker, and so it is likely to be expected by a social partner in such cases. To take a simple example, if you enter a store and the clerk asks "how are you?", the script is to reply "fine, thanks," regardless of how not-fine you are. The clerk is not actually asking how you are, and it would violate social norms to pause, reflect, and give an honest answer. Many scripts are culture-specific ("how are you?" —"fine, thanks" may be particularly familiar to North Americans). But the existence of scripts, and the expectation that they ought to be followed in some social contexts, is likely more general. In particular, prior work has already shown situations in which fast, nondeliberative responses are both expected and positive. For example, Oktar and Lombrozo (2022) have considered decision-making situations in which people expect others to not deliberate in their response. Such behavior seems intuitive—when asking someone to marry you, one expects a nondeliberative "Yes!", though we note that such nondeliberation is not itself indicative of a rote or scripted response, and as discussed in Oktar and Lombrozo (2022), it is rather a sign of authenticity or strength of

preference. Future work could explore the pervasiveness of sensitivity to rote behavior across contexts.

Moreover, we note that the cues that people use to infer automatic behavior in others, and the situations in which rote-ness is perceived as appropriate, may also differ by cultural context. While the U.S.-based adults recruited for these studies generally agreed that cues such as repetition, fluency, and attentiveness can be used to infer rote behavior, and that rote behavior leads to worse teaching, whether these findings generalize to broader populations is an open question.

In the current work, we focused on pedagogy as an important domain, but the detection of automaticity is pervasive beyond pedagogy: A relationship that sours when a partner nods without listening, a politician that loses popularity for being too-polished when compared to the off-the-cuff brute that seems genuine, an emotional story that loses its punch in the retelling. The findings and analysis in this paper begin to examine the shared basic reasoning underlying these seemingly disparate situations and others. But more is needed to examine whether, when, and how people notice rote behavior; and those answers will not come automatically.

References

Ahn, I., Chiu, M. M., & Patrick, H. (2021). Connecting teacher and student motivation: Student-perceived teacher need-supportive practices and student need satisfaction. *Contemporary Educational Psychology*, *64*, 101950. https://doi.org/10.1016/j.cedpsych.2021.101950

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914–930. https://doi.org/10.1037/0278-7393.33.5.914

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Bascandziev, I., Shafto, P., & Bonawitz, E. (2021). The sound of pedagogical questions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Bass, I., Bonawitz, E., Hawthorne-Madell, D., Vong, W. K., Goodman, N. D., & Gweon, H. (2022). The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition*, *222*, 104999. https://doi.org/10.1016/j.cognition.2021.104999

Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023). Thinking about thinking as rational computation. *PsyArXiv preprint: 10.31234/osf.io/e65p3*.

Bonawitz, E., & Shafto, P. (2016). Computational models of development, social influences. *Current Opinion in Behavioral Sciences*, *7*, 95–100. https://doi.org/10.1016/j.cobeha.2015.12.008

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*(3), 322–330. https://doi.org/10.1016/j.cognition.2010.10.001

Bonawitz, E., Shafto, P., Yu, Y., Gonzalez, A., & Bridgers, S. (2020). Children change their answers in response to neutral follow-up questions by a knowledgeable asker. *Cognitive Science*, *44*(1), e12811. https://doi.org/10.1111/cogs.12811

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Brennan, S., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383–398. https://doi.org/10.1006/jmla.1995.1017

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Davis, I., Carlson, R., Dunham, Y., & Jara-Ettinger, J. (2023). Identifying social partners through indirect prosociality: A computational account. *Cognition*, *240*, 105580. https://doi.org/10.1016/j.cognition.2023.105580

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78. https://doi.org/10.1098/rstb.1985.0010

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Etkin, A., Büchel, C., & Gross, J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience*, *16*(11), 693–700. https://doi.org/10.1038/nrn4044

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*(1), 37–55. https://doi.org/10.1207/S15326950DP3401_2

Gasser, L., Grütter, J., Buholzer, A., & Wettstein, A. (2018). Emotionally supportive classroom interactions and students' perceptions of their teachers as caring and just. *Learning and Instruction*, *54*, 82–92. https://doi.org/10.1016/j.learninstruc.2017.08.003

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, *217*, 104885. https://doi.org/10.1016/j.cognition.2021.104885

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016). Plans, habits, and theory of mind. *PLOS ONE*, *11*(9), e0162246. https://doi.org/10.1371/journal.pone.0162246

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Goodman, N. D., & Frank, M. C. (2016a). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., & Frank, M. C. (2016b). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behavior*, *6*, 975–987. https://doi.org/10.1038/s41562-022-01316-8

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*, 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, *45*(3), e12926. https://doi.org/10.1111/cogs.12926

Heller, D., Arnold, J. E., Klein, N. M., & Tanenhaus, M. K. (2015). Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and Speech*, *58*(02), 190–203. https://doi.org/10.1177/0023830914528107

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development*, *8*(3), 263–283. https://doi.org/10.1080/15248370701446392

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38. https://doi.org/10.1016/j.cognition.2015.05.006

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Computational Biology*, *7*(5), e1002055. https://doi.org/10.1371/journal.pcbi.1002055

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925–934. https://doi.org/10.1111/j.1467-7687.2011.01049.x

Konovalov, A., & Krajbich, I. (2023). Decision times reveal private information in strategic settings: Evidence from bargaining experiments. *Economic Journal*, *133*(656), 3007–3033. https://doi.org/10.1093/ej/uead055

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. https://doi.org/10.1017/S0140525X1900061X

Liljeholm, M., Tricomi, E., O'Doherty, J. P., & Balleine, B. W. (2011). Neural correlates of instrumental contingency learning: Differential effects of action–reward conjunction and disjunction. *Journal of Neuroscience*, *31*(7), 2474–2480. https://doi.org/10.1523/JNEUROSCI.3354-10.2011

Loy, J. E., Rohde, H., & Corley, M. (2017). Effects of disfluency in online interpretation of deception. *Cognitive Science*, *41*(S6), 1434–1456. https://doi.org/10.1111/cogs.12378

Oktar, K., & Lombrozo, T. (2022). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. *Cognition*, *223*, 105021. https://doi.org/10.1016/j.cognition.2022.105021

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, *17*(5), 1215–1233. https://doi.org/10.1177/17456916211048487

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioural and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Roth, G., Assor, A., Kanat-Maymon, Y., & Kaplan, H. (2007). Autonomous motivation for teaching: How self-determined teaching may lead to self-determined learning. *Journal of Educational Psychology*, *99*(4), 761–774. https://doi.org/10.1037/0022-0663.99.4.761

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, *72*(4), 1054–1070. https://doi.org/10.1111/1467-8624.00334

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Smart, J. B. (2014). A mixed methods study of the relationship between student perceptions of teacher–student interactions and motivation in middle level science. *RMLE Online*, *38*(4), 1–19. https://doi.org/10.1080/19404476.2014.11462117

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, *53*(1), 81–94. https://doi.org/10.1016/j.jml.2005.02.003

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, *115*(34), 8491–8498. https://doi.org/10.1073/pnas.1804761115

Tong, Y., Wang, F., & Danovitch, J. (2020). The role of epistemic and social characteristics in children's selective trust: Three meta-analyses. *Developmental Science*, *23*(2), e12895. https://doi.org/10.1111/desc.12895

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375. https://doi.org/10.1146/annurev.ps.43.020192.002005

Wentzel, K. R. (2009). Students' relationships with teachers as motivational contexts. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 301–322). New York: Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203879498

Wigfield, A., Cambria, J., & Eccles, J. S. (2012). Motivation in education. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 463–478). New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399820.013.0026

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review*, *114*(4), 843–863. https://doi.org/10.1037/0033-295X.114.4.843

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619. https://doi.org/10.1016/j.cognition.2021.104619