You're surprised at her success?
Inferring competence from emotional responses to performance outcomes

Mika Asaba [1], Yang Wu[1], Brandon Carrillo, & Hyowon Gweon
{masaba, yangwu1, bacarr, hyo}@stanford.edu
Department of Psychology, Stanford University, 450 Jane Stanford Way,
Stanford, CA, USA, 94305

Abstract

How do we learn about who is good at what? Others' competence is unobservable and often must be inferred from observable evidence, such as failures and successes. However, even the same performance can indicate different levels of competence depending on the context, and objective evaluation metrics are not always available. Building on recent advances on children's use of emotion as information, here we ask whether expressions of surprise inform inferences about competence. Participants saw scenarios (sports, academics) where two students achieved identical outcomes but a teacher showed surprise to one student and no surprise to the other. In Exp.1, adults inferred that the successful student who elicited the teacher's surprise was less competent than the other student, but this pattern reversed when both students failed. Exp.2 (4-9-year-olds) finds initial evidence for such inferences in school-aged children. These findings have implications for promoting healthy social comparisons and preventing acquisition of negative stereotypes from non-verbal cues.

Keywords: affective cognition; cognitive development; theory of mind; social reasoning; achievement

Introduction

We are curious about who is good at what. Learning about our own and others' competence not only fulfills our curiosity, but also allows us to form accurate beliefs about what we (and others) can do and how best to improve. These beliefs influence social comparisons and academic motivation (Heyman, Dweck, & Cain, 1992; Dweck, 2008) and even important life decisions with long-term consequences (e.g., career choices). Thus, the ability to reason about one's own and others' competence is a critical social skill, and it is especially important for young children who are constantly learning about the world, trying new tasks, and constructing a sense of self.

Evaluating competence, however, is a challenging task. Rather than being directly observable, it must be inferred from observable evidence such as performance outcomes (e.g., successes and failures). Yet, these outcomes are often embedded in complex social contexts, leading to different inferences depending on features of the task (e.g., difficulty; Leonard, Bennett-Pierre, & Gweon, 2019), who is compared to whom (Nicholls & Miller, 1983, 1984), and what standards are used for evaluation (Asaba et al., 2018). Thus, aside from domains with widely-accepted standards for assessing performance (e.g., GRE score percentile), performance alone is often insufficient for accurate inferences about competence.

Another source of information about competence is social feedback (i.e., others' responses to one's performance). Prior work suggests that explicit verbal feedback (praise, criticism) can have powerful implications for learning, motivation, and even how children think about their own abilities (e.g., Mueller & Dweck, 1998; Henderlong & Lepper, 2002; Dweck, 2008). Yet, such direct feedback can often be uninformative or simply unavailable in many real-world contexts. First, people may not always provide informative, honest feedback due to various social considerations (e.g., acting polite or preventing others from feeling bad; Yoon, Tessler, Goodman, & Frank, 2016); indeed, recent work suggests that preschool-aged children distinguish informative praise from uninformative praise based on prior observations of others' past praise and the quality of work praised (Asaba et al., 2018). Second, such information may not always be available. In many educational and professional settings (let alone in more casual contexts), explicit judgments of competence are often withheld; in Westernized societies, for instance, such explicit judgments are believed to cause competitiveness or stress, or hinder creative, innovative thinking.

Of course, others' explicit feedback is not the only form of social feedback we receive from others; we are also sensitive to (and look out for) others' non-verbal responses, such as their emotional reactions to performance outcomes. While some emotional expressions are clearly valenced and correlate highly with outcome (e.g., joy when your team scores a goal, frustration following failures; see Skinner, Olson, & Meltzoff, 2019), others, such as surprise, are valence-neutral but nonetheless informative. For instance, if your colleague looked surprised that you successfully wrote a simple line of Python code, you might infer that he probably thought you were a beginner programmer and perhaps not capable of more challenging tasks. Such indirect, implicit responses are subtle but powerful sources of information.

These inferences can manifest even more clearly in relative judgments of competence. Imagine two students, Adam and Kyle, who are receiving their math exam results. The teacher looks unsurprised as she hands back Adam's "A"; she simply nods and smiles per usual. However, the teacher looks surprised as she hands back Kyle's results, even though he also got an A. Despite the fact that Adam and Kyle both got A's, you might infer that (the teacher thinks) Adam is better than Kyle at math. This is because the teacher's surprised expression indicates a violation of her expectations about Kyle's performance; the teacher *had not* expected Kyle to receive an A (and expected a lower grade instead). On the other hand,

[1]These authors contributed equally to this work.

the teacher's neutral-positive response to Adam suggests that Adam's performance was consistent with her expectations. While seemingly intuitive, such inferences reflect an abstract, causal understanding of how emotional responses are elicited not only by observable outcomes (e.g., getting A's) but also by unobservable mental states (e.g., the teacher's beliefs).

Decades of research in cognitive development have investigated our intuitive, theory-like understanding of emotions (Harris, Johnson, Hutton, Andrews, & Cooke, 1989; Lagattuta, 2014; Wellman & Liu, 2004). More recently, computational approaches to studying human social cognition have also formalized these intuitions as a generative causal process by which an agent's mental states (i.e., beliefs, desires) and external events jointly give rise to various emotions that manifest as the agent's emotional expressions (Ong, Zaki, & Goodman, 2019; Wu, Baker, Tenenbaum, & Schulz, 2018; Saxe & Houlihan, 2017). Such internal causal models can support the observer's inferences in two directions.

First, we can *predict* others' emotions based on their mental states and observed outcomes (forward inference). Much developmental work has focused on forward inferences (e.g., Harris et al., 1989; Lagattuta, 2014; Doan, Friedman, & Denison, 2018; Wellman & Liu, 2004); for instance, recent work suggests that 5 year-old children can integrate others' expectations with event outcomes to infer their emotions (e.g., given identical outcomes, someone with low expectations might feel better than an agent with high expectations; Asaba, Ong, & Gweon, 2019; Lara, Lagattuta, & Kramer, 2019).

Second, such intuitive understanding of others' emotions also supports inferences in the reverse direction (inverse inference); given others' emotional expressions, we can infer the internal mental states (as in the math example above) or external events that gave rise to the emotional expressions. For example, recent work suggests that even infants can infer the probable external events that have elicited others' emotional expressions (Wu, Muentener, & Schulz, 2017). By age 5, children can jointly infer others' beliefs and desires from their emotional responses to anticipated and observed outcomes (Wu & Schulz, 2018): For instance, if an agent looks happy before opening a box and sad after opening it, children can infer that the agent thought there was something desirable in the box (but in fact there's not). Older children (7-year-olds) can even make second-order mental state inferences from emotional expressions displayed in social contexts (Wu & Schulz, 2019). Collectively, this work suggests that at least by the early school years, children can make sophisticated mental state inferences by integrating emotional expressions, event outcomes, and contextual information.

Based on this literature, it may seem plausible that 5- to 7-year-olds can reason about competence in much the same way as adults. For instance, in the math example above, they might understand the teacher expected Adam to receive an A, but not Kyle. However, prior studies on emotion-based inferences have focused on inferring beliefs about concrete, physical states (e.g., contents of a box, location of a desirable treat; Wu & Schulz, 2018) rather than beliefs about unobservable qualities of people such as competence. Although recent evidence suggests that 3- and 4-year-olds use others' observations of their own successes and failures to infer what others think of their competence (e.g., "she thinks I can't activate this toy", Asaba & Gweon, 2019), whether children can draw such inferences based on observers' emotional expressions remains an open question. Furthermore, prior work suggests that children struggle with relative judgments of competence even in late childhood especially given conflicting cues (e.g., given the same outcome, a student who goofed off is better at math than the one who worked hard; Nicholls, 1978) or in the absence of cues that clearly mark underlying competence (e.g., speed, Leonard et al., 2019).

The current study used scenarios similar to the math example above to investigate children's abilities to infer others' competence based on an observer's emotional responses. In particular, we focus on the emotional expression of surprise, given children's understanding of surprise as the mismatch of prior expectations and actual outcomes (Wellman & Banerjee, 1991). Participants were presented with vignettes in which two students showed the same performance at a task (i.e., both succeeded, or both failed) as their teacher observed; the teacher looked surprised at one student's outcome and not surprised at the other student's outcome. At test, participants were asked which agent is better at that task. To examine the generalizability of such inferences, we created two physical (kicking, throwing) and two academic (math, spelling) scenarios. In Exp.1, we first establish adults' intuitions; in Exp.2, we recruit from a broad age range of children (4 year-olds to 9 year-olds) to identify the emergence and developmental change of these judgments[2].

Experiment 1: Adults

Methods

Participants Sixty-seven adults (M_{Age}(SD) = 36.9(10.9), range: 19-64, 39 female) were recruited from Amazon's Mechanical Turk. An additional 4 subjects were excluded for failing 50% (8 of 16) or more of post-test check questions[3].

Stimuli For the teacher, we used two photos of an adult female: one with a surprised expression and one with a smiling expression. We used the smiling face as the baseline (i.e., non-surprised) expression because a neutral expression is often interpreted as negative, and a smiling face was more appropriate in the context of our task (i.e., a teacher observing her students). For the students, we used same-gendered pairs of generic cartoon characters without facial features. Each activity (kicking, throwing, spelling, math) was depicted with simple cartoon images; successes and failures were marked with a green check mark and a red "X", respectively. See Fig.1 for examples.

[2]Analyses, scripts, and stimuli for Exp.1-2: `https://osf.io/4ezvt/?view_only=4ec60eae51174932addaf680c908c3e5`

[3]Including these participants does not change the qualitative results of the study.

Procedure All participants responded to 8 trials (4 different activities, 2 trials each: success, fail). Each trial consisted of three phases shown on the same page: Introduction, Test, and Post-test checks. Order of activities was randomized; within each activity, the fail and success trials were paired and presented in randomized order, and gender was counterbalanced across trial and activity (e.g., girls succeeded in kicking but failed in throwing, succeeded in math, but failed in spelling).

In the Introduction, participants were shown images of the students and the teacher's smiling face (baseline expression) with accompanying text that described the goal of the activity (e.g., "In this game, you throw a ball into a bucket"). To make it more plausible that students' performance may violate the teacher's expectations, participants were told about the variability of students' performance (e.g., "Some of the kids are better than others. But sometimes, kids get lucky and get the ball in, and sometimes, they accidentally make mistakes").

At test, participants were shown two same-gender students who both succeeded (success trial) or failed (fail trial) at the same activity. The only difference between the two students was whether the teacher expressed surprise (henceforth "surprise student") or did not express surprise ("no-surprise student") at the outcome. The teacher's emotional expression was clearly labeled for participants: "The teacher was *not surprised* that Adam made the ball in, and was *surprised* that Kyle made the ball in." The Test Question was: "Who is better at throwing the ball into the bucket?" The other prompts were: "...kicking the ball into the goal?", "...math?", "...spelling?" See Fig.1 for trials for the throwing activity.

In the Post-test check phase, participants were asked to report the teacher's emotional response for each agent's outcome (e.g., "Just to check, was the teacher surprised or not surprised that Emma made the ball into the goal?"); participants answered 16 questions total (2 per trial, 4 activities).

Results and Discussion

Our primary question was whether adults would use the teacher's emotional responses to students' performance outcomes to infer their relative competence. We ran a generalized linear mixed-effects model with trial (fail, success) and activity type (math, spelling, throwing, kicking) as fixed effects and participant as a random effect predicting which student participants chose as better (the surprise student or no-surprise student). First, we found that trial type significantly predicted responses (β = -5.127; $z = -14.958$, $p < .001$), but activity type did not ($p's > .347$). Second, consistent with our predictions, participants selectively chose the no-surprise student in success trials (92.2%, $Z = 7.45$, $p < .001$, Exact Wilcoxon-Pratt Signed-Rank Test) and the surprise student in fail trials (93.0%, $Z = 7.51$, $p < .001$). Finally, participants were equally accurate in success and fail trials ($Z = -.45$, $p = .746$, Exact Wilcoxon-Pratt Signed-Rank Test).

These results provide robust support for our hypothesis that emotional responses inform people's inferences about relative competence. When two agents both succeeded at the same task, the agent to whom the teacher expressed surprise was judged as less competent. In contrast, if both had failed, the agent to whom the teacher expressed surprise was judged as more competent. Thus, participants were not simply using surprise as indicators of competence (or lack thereof); rather, they integrated the teacher's emotional response with the students' outcome to infer the students' underlying competence.

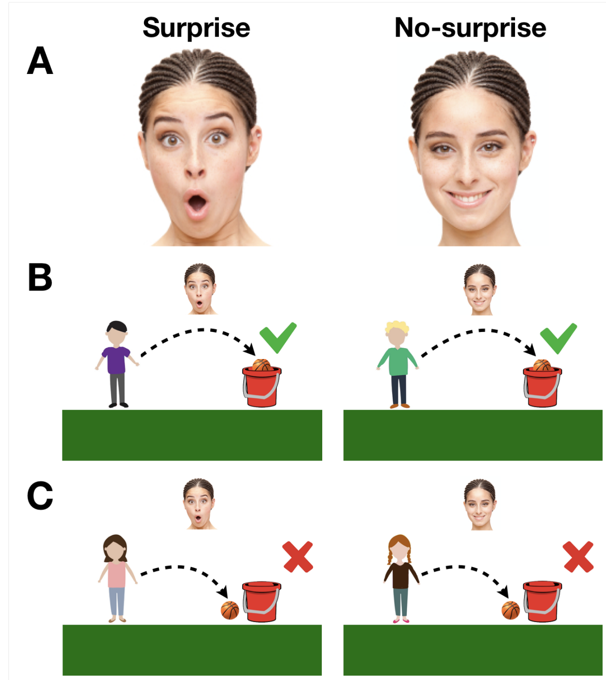

Figure 1: Stimuli for Exp. 1 & 2. A: The teacher's surprised and non-surprised expression. B: Example of a success trial for the throwing activity. The teacher shows a surprise response for one student and a non-surprised expression for the other. Images were shown side-by-side during the test question. C: Example of a fail trial for the throwing activity.

Experiment 2: 4- to 9-year-olds

This task not only requires inferring expectations from emotional responses, but also using and holding such expectations in mind when making competence judgments. So, in Experiment 2, we recruited a relatively broad range of children (4- to 9-year-olds) in an initial sample with the goal of capturing the age at which children successfully make these judgments.

Methods

Participants Twenty-eight children (13 female, M_{Age}(SD) = 6.6(1.7), range: 4.1-9.9) were recruited from a local museum (n = 20) and campus preschool (n = 8). We excluded participants who failed 50% or more of the check questions (n = 1 four-year-old)[4] or who did not respond to the test questions (i.e., responded "both" to all questions; n = 1 six-year-old).

[4]Including this participant does not qualitatively change the results.

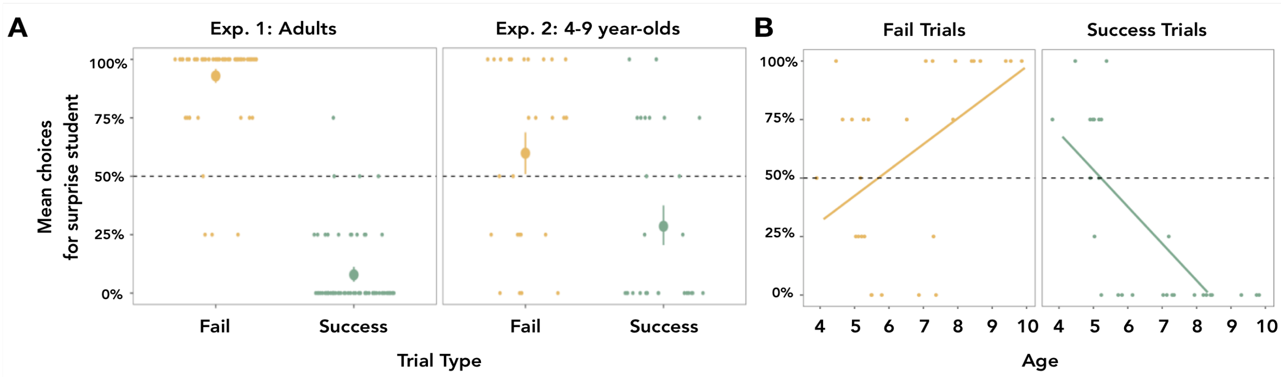

Figure 2: Results for Experiments 1 & 2. A: Exp.1 (adults) and Exp.2 (4-9 year-olds) group means (large dots) and individual means (small dots) for choosing the student who received surprise as more competent for fail and success trials. Error bars represent 95% confidence intervals. B: Scatterplots for Exp.2 individual means by age for fail and success trials.

Stimuli Participants were shown images of the teacher's surprised and non-surprised expressions on laminated paper for warm-up questions at the beginning of the experiment. For the test trials, children were shown the same images from Exp.1 on laminated paper.

Procedure Children were tested in a quiet room in the museum or preschool. In the warm-up phase, subjects were introduced to the teacher and shown her two facial expressions (see Fig 1A) described as "surprised" and "non-surprised," respectively. For each expression, they were asked check questions that if the teacher was surprised or not surprised.

Then, all participants underwent the same eight trials as in Exp.1 with Introduction, Check, and Test phases for each trial. Minor modifications were added to make the task more engaging for children. In each trial, the experimenter remarked on one student's outcome ("look, Emma kicked the ball into the goal"), revealed the teacher's expression ("the teacher was watching and let's look at her face now"), and asked participants the check question about the teachers' emotion ("Is she surprised or not surprised?"); if participants provided an incorrect response to the check question, the experimenter provided the correct response. Then, this sequence was repeated for the other student in the trial who received the other emotional response. Thus, by the test question, participants clearly understood whether the teacher was surprised or not surprised at each student's performance.

Finally, with images of the students' outcomes and the teacher's expressions visible, the experimenter asked, "One of the kids is better at this game. Who is better at [kicking, throwing, math, spelling]?" As in Exp. 1, the fail and success trials were paired and randomized within activity; activity order was randomized. Each participant underwent all 8 trials.

Results

All participants correctly identified the teacher's emotional expressions ("surprised" or "not surprised") in the warm-up questions. For the check questions in each trial about the teacher's emotional response to the student's performance ("surprised" or "not surprised"), the majority of participants (25 of 28) correctly answered all 16 questions; the remaining correctly answered 10 (n = 1 four-year-old), 14 (n = 1 five-year-old), and 15 (n = 1 nine-year-old) of the questions.

We ran a generalized linear mixed-effects model with trial (fail, success), age (continuous), and activity (math, spelling, throwing, kicking) as fixed effects, with an interaction term between trial and age, and participant as a random-effect, predicting participants' choice of student. Consistent with the results from Exp.1, we found a main effect of trial type (β = .142, $z = 5.138$, $p < .001$) but no main effect of activity ($p's > .161$). Additionally, we also found a main effect of age (β = .876, $z = 2.871$, $p = .004$), and an interaction between trial type and age (β = -.267, $z = -5.314$, $p < .001$).

As a group, children chose the no-surprise student in success trials (71.4%, $Z = 2.91$, $p = .004$, Exact Wilcoxon-Pratt Signed-Rank Test), but did not choose the surprise student in fail trials significantly above chance (59.8%, $Z = 1.32$, $p = .211$). To ask whether children were differentially *accurate* in fail and success trials, we dummy-coded responses as correct if the no-surprise agent was chosen in success trials and if the surprise agent was chosen in fail trials; we did not find a significant difference in children's accuracy for success and fail trials ($Z = .98$, $p = .310$, Exact Wilcoxon-Pratt Signed-Rank Test).

Given the wide age range and significant age-by-trial interaction, we median-split children into younger (age: 4.1 - 5.9; N=14) and older age groups (age: 6.2 - 9.9; N=14) and looked at children's choices within each trial. We found that the older group was accurate for success and fail trials (Success: 98.2%, $Z = 3.64$, $p < .001$; Fail: 76.8%, $Z = 2.16$, $p = .039$) and marginally more accurate in success trials than in fail trials ($Z = 2.22$, $p = .063$). The younger group was at chance for both trial types (Success: 44.6%, $Z = -.58$, $p = .707$; Fail: 42.9%, $Z = -.81$, $p = .536$) with no difference between success and fail trials ($Z = .06$, $p = .985$)

Collectively, these results provide initial evidence for children's developing abilities to judge others' relative competence from emotional responses to performance outcomes. However, these abilities were observed only in the older half of the participants; given two agents who achieved the same outcome, children used the teacher's different emotional expressions to figure out who was likely to be better at the game.

General Discussion

Across two experiments, we examined adults' and children's abilities to infer others' relative competence based on an observer's emotional responses to their performance outcomes. The tasks were highly similar between adults and children; two students produced identical outcomes on an activity, but the teacher expressed surprise to only one of the outcomes. In Exp.1, we found a clear difference in adults' responses across trial types (success, failure); adults inferred that the student who elicited surprise was *less competent* when both students were successful, but the student who elicited surprise was *more competent* when both students had failed. These results were observed consistently in all activity types (academic: math, spelling; physical: kicking, throwing). In Exp.2, we found preliminary support for an emerging competence for such inferences: While children between 6 and 9 years of age correctly responded to both trials, younger children (4- to 5-year-olds) did not show such pattern. As an initial study, these results provide suggestive evidence that by the early school years, older children can integrate emotional responses and outcomes to infer an observer's prior expectations and make sense of others' relative abilities.

Unlike older children, younger children (4- to 5-year-olds) did not show evidence of such inferences. This is unlikely to be due to younger children being unable to understand the scenarios or attend to the task; they correctly answered the warm-up and check questions ("Is she surprised or not surprised?"), and appropriately chose only one student at test, suggesting that they were able to recognize the expressions and understand the test question. Although children at this age range show some difficulties labeling surprised expressions in *free-labeling* tasks (Widen, 2013), those difficulties may be, at least partially, due to the high demands of a free-labeling task (Wu & Gweon, 2019).

So, why did younger children struggle with this inference? Below we consider a few possible sources of children's difficulties. One possibility is that younger children are genuinely incapable of drawing the key inference required in this task, i.e., integrating the teacher's emotional response with the student's performance to infer the teacher's prior beliefs about the student's competence. Although past work suggests that even infants and preschoolers understand that expressions of surprise indicate violations of prior expectations (Wellman & Banerjee, 1991; Wu & Gweon, 2019), such studies have used clear outcomes of simple physical events (e.g., sampling a rare ball; Wu & Gweon, 2019). Thus, children in this study may have struggled with generating the alternative possibility (i.e., the idea that the teacher could have expected the student to perform differently) or have difficulty converting the teacher's prior expectations to representations of the students' competence.

Critically, however, it is also possible that younger children are capable of drawing the key inference but are unable to express their understanding due to extraneous task demands; children had to represent two sets of mental states (i.e., the teacher's prior beliefs about the surprise and no-surprise student) and compare their relative competence. Although we chose to use relative judgments to simplify the dependent measure, it is possible that keeping both in mind is beyond their representational capacity (Leahy & Carey, 2020) or executive demands (Carlson & Moses, 2001).

Finally, other aspects of the task might have masked their competence. For instance, with less experience with formal schooling, it is possible that younger children do not readily consider emotional expressions as relevant or informative cues to competence. If so, it is possible that they can benefit from a richer, more motivating cover story that contextualizes its relevance. Future work might investigate the nature of younger children's difficulty by probing the teacher's prior expectations about performance and/or providing the expectation and asking them to judge the students' relative competence.

While older children, as a group, showed above-chance accuracy on the task, some aspects of their responses raise questions about whether their responses are truly adult-like. First, while only a trend, children showed a relatively lower accuracy on the fail trials than the success trials, whereas adults did not show this difference. One possibility for this is that our test question may have been confusing for the fail trials, and it may be more cognitively demanding to reason about "who is better" when both agents failed the task. Simply asking "who is worse" may not resolve the issue, as children may find the question unfamiliar and even pragmatically weird. However, the trends in this initial study should be interpreted with caution because we recruited from a broad age range with few children in each age group. We are currently collecting a larger sample of children to replicate these findings and better determine the effect sizes as well as the age trends.

In the current study, the teacher's emotional response was the only distinguishing factor between the two students. This was an important design decision to isolate the role of emotional expressions in judgments of others' relative competence. However, the fact that we used a smiling expression as the baseline, no-surprise face raises an alternative explanation: Children might associate happiness/smiles with success, and use the association to choose an answer rather than using the presence or absence of surprise to infer the teacher's underlying beliefs. Thus, in success trials, children might have chosen the no-surprise student because the teacher smiled; in fail trials, children might have chosen the surprise (i.e., correct) student simply to avoid choosing the inconsistent pairing between failure and smile. We believe that this is a rather unlikely possibility for several reasons. First, we used the smiling face as the baseline expression even during introduction scenarios where the smiling face was not associated with any performance outcomes. Second, throughout the task, this expression was labeled as the one where the teacher was "not surprised" (rather than "happy"), and phrased our check questions accordingly (i.e., "Is she surprised or not surprised"). More generally, a smiling expression is an appropriate baseline expression not only in the context of our task (i.e., a teacher observing her students) but in many other social contexts. For instance, beyond positive events such as successes, negative events such as failures can also elicit positive emotions especially through encouragements and empathy. Thus, children might associate smiling expressions with a broad range of events. However, future studies may address this concern with more sensitive measures, such as showing participants one agent at a time and asking them to rate their competence on a scale. While using a rating scale poses a different kind of task demand for preschool-aged children, prior work suggests children by 6 years of age are able to use these scales.

In this initial study, we focused on children's and adults' third-party judgments of others' competence. However, children often observe adults' emotional expressions in response to their own performance. In fact, children may be especially motivated to make sense of their own abilities, and may be particularly sensitive to their parents', teachers', and peers' verbal and non-verbal responses to their outcomes and abilities. The current results raise the possibility that children can incorporate others' emotional responses into evaluations of their *own* abilities. Thus investigating first-person evaluations from others' emotional responses is an exciting future direction. Such inferences might manifest particularly in contexts that naturally invite social comparison, such as when other peers or siblings are attempting the same or similar tasks.

Note that the meaning of emotional expressions depends heavily on what we know about the emoter. In this task, we specifically chose the students' *teacher* to be the one providing emotional responses; we assumed that children would readily attribute knowledge of the student's abilities to the teacher, and trust her emotional responses to be reliable and accurate. Imagine instead if the teacher *falsely* thought that the student had previously done poorly, then her surprise at the student's good performance should be discounted when evaluating the student's competence. Thus, the informativeness of others' emotional expressions may depend on our evaluations of the accuracy of others' prior beliefs.

Indeed, in real-world situations, emotion is rarely the only available factor to consider in inferring others' competence. In fact, it is not always clear how heavily emotional expressions are (or should be) weighed relative to other relevant factors, such as performance outcomes or verbal feedback. Interestingly, the congruency between emotional responses and explicit feedback might serve as a cue for sincerity. Future work will investigate how children integrate across explicit (e.g., feedback, praise, criticism) and implicit cues (e.g., emotion) to make sense of others' informativeness and evaluate abstract qualities of the self and others. Furthermore, the relevance of emotion may also depend on the specific activity or skill; especially in cases where there are no clear standards for success (e.g., giving an academic talk at a conference, pitching a research idea), others' emotional responses during and after the event may serve as particularly helpful cues for one's performance and abilities.

Finally, this work has important implications for our understanding of children's acquisition of stereotypes. By 6, girls are less likely to believe that girls are "really, really smart" and avoid activities that are described as being for children who are really smart (Bian, Leslie, & Cimpian, 2017). Compared to the role of linguistic cues in promoting and perpetuating stereotypes (e.g., Chestnut & Markman, 2018) relatively less is known about the role of implicit, non-verbal cues like emotional expressions in the formation of stereotypes. Some recent work suggested that children tend to prefer and imitate the target of positive nonverbal signals (e.g., smiling and leaning in vs. scowling and leaning away), and do so even for the target's novel social group (Skinner, Meltzoff, & Olson, 2017; Skinner et al., 2019). Our findings go beyond these findings and raise the possibility that children are not simply sensitive to the valence of others' expressions; rather, they can draw inferences from others' surprised expressions (i.e., a non-valenced signal) about an agent's competence, around the age that they acquire gender-based stereotypes about intelligence (Bian et al., 2017). Future work will explore whether children can learn about a group's competence from a teacher's emotional expressions, such as surprise. More broadly, studying others' emotional reactions to under-represented minorities' performance in STEM fields is an important area for future work, especially because adults may learn how to explicitly communicate in "unbiased" ways, yet their emotional responses may nonetheless indicate underlying implicit biases.

In sum, reasoning about others' emotional responses is critical for not only interacting with others and intervening to make others feel better, but also for learning about the world, others, and the self. Bridging prior work on children's developing intuitive theories of emotion and their understanding of competence, here we provided initial evidence that emotional responses are a rich source of information about abstract qualities of other agents, like competence or ability; critically, emotional expressions can provide meaning beyond direct observation or verbal feedback.

Acknowledgments

We thank Stephanie Chang for help with data collection, Kate Littlejohn for help with coding, and the Palo Alto Junior Museum & Zoo, and families who participated. This work was supported by an NSFGRFP to MA; the Paul & Lilah Newton Brain Science Award to YW; a McDonnell Scholars Award and a Jacobs Foundation Research Fellowship to HG.

References

Asaba, M., & Gweon, H. (2019). Young children rationally revise and maintain what others think of them. *PsyArXiv*. doi: 10.31234/osf.io/yxhv5

Asaba, M., Hembacher, E., Qiu, H., Anderson, B., Frank, M. C., & Gweon, H. (2018). Young children use statistical evidence to infer the informativeness of praise. In *Proceedings of the 40th annual conference of the cognitive science society* (p. 112-117).

Asaba, M., Ong, D. C., & Gweon, H. (2019). Integrating expectations and outcomes: Preschoolers' developing ability to reason about others' emotions. *Developmental psychology*, *55*(8), 1680–1693.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child development*, *72*(4), 1032–1053.

Chestnut, E. K., & Markman, E. M. (2018). "Girls are as good as boys at math" implies that boys are probably better: A study of expressions of gender equality. *Cognitive science*, *42*(7), 2229–2249.

Doan, T., Friedman, O., & Denison, S. (2018). Beyond belief: The probability-based notion of surprise in children. *Emotion*, *18*(8), 1163–1173.

Dweck, C. S. (2008). *Mindset: The new psychology of success*. Random House Digital, Inc.

Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition & Emotion*, *3*(4), 379–400.

Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological bulletin*, *128*(5), 774–795.

Heyman, G. D., Dweck, C. S., & Cain, K. M. (1992). Young children's vulnerability to self-blame and helplessness: Relationship to beliefs about goodness. *Child Development*, *63*(2), 401–415.

Lagattuta, K. H. (2014). Linking past, present, and future: Children's ability to connect mental states and emotions across time. *Child Development Perspectives*, *8*(2), 90–95.

Lara, K. H., Lagattuta, K. H., & Kramer, H. J. (2019). Is there a downside to anticipating the upside? children's and adults' reasoning about how prior expectations shape future emotions. *Child Development*, *90*(4), 1170–1184.

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78.

Leonard, J. A., Bennett-Pierre, G., & Gweon, H. (2019). Who is better? Preschoolers infer relative competence based on efficiency of process and quality of outcome. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.*

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of personality and social psychology*, *75*(1), 33-52.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child development*, *49*(3), 800–814.

Nicholls, J. G., & Miller, A. T. (1983). The differentiation of the concepts of difficulty and ability. *Child Development*, *54*(4), 951–959.

Nicholls, J. G., & Miller, A. T. (1984). Reasoning about the ability of self and others: A developmental study. *Child Development*, *55*(6), 1990–1999.

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, *11*(2), 338–357.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a bayesian model of theory of mind. *Current opinion in Psychology*, *17*, 15–21.

Skinner, A. L., Meltzoff, A. N., & Olson, K. R. (2017). "catching" social bias: Exposure to biased nonverbal signals creates social biases in preschool children. *Psychological Science*, *28*(2), 216–224.

Skinner, A. L., Olson, K. R., & Meltzoff, A. N. (2019). Acquiring group bias: Observing other people's nonverbal signals can create social group biases. *Journal of personality and social psychology*.

Wellman, H. M., & Banerjee, M. (1991). Mind and emotion: Children's understanding of the emotional consequences of beliefs and desires. *British Journal of Developmental Psychology*, *9*(2), 191–214.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), 523–541.

Widen, S. C. (2013). Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emotion Review*, *5*(1), 72–77.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive science*, *42*(3), 850–884.

Wu, Y., & Gweon, H. (2019). Surprisingly unsurprising! infants' looking time to improbable events is modulated by others' expressions of surprise. *PsyArXiv*. doi: 10.31234/osf.io/8whuv

Wu, Y., Muentener, P., & Schulz, L. E. (2017). One- to four-year-olds connect diverse positive emotional vocalizations to their probable causes. *Proceedings of the National Academy of Sciences*, *114*(45), 11896–11901.

Wu, Y., & Schulz, L. E. (2018). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child development*, *89*(2), 649–662.

Wu, Y., & Schulz, L. E. (2019). Understanding social display rules: Using one person's emotional expressions to infer the desires of another. *Child Development*.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2771–2776).