Experimenter identity: An invisible, lurking variable in developmental research

Thomas St. Pierre[1], Katherine S. White[2], & Elizabeth K. Johnson[1,3]

(thomas.stpierre@utoronto.ca, white@uwaterloo.ca, elizabeth.johnson@utoronto.ca)

[1]University of Toronto Mississauga

[2]University of Waterloo

[3]University of Toronto

Author Note

Correspondence concerning this article should be addressed to Thomas St. Pierre, Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Rd., Mississauga, Ontario, L5L 1C6, CANADA. Email: thomas.stpierre@utoronto.ca

Abstract

Developmental researchers are well aware that children behave differently around different people. Nevertheless, researchers rarely consider (and report on) who is running their studies. Indeed, in a survey of articles published in the last 3 years in 4 top developmental journals, we find that the vast majority of studies fail to report any information about experimenter identity, despite the fact that child-adult interactions may be strongly influenced by the social inferences that individuals draw from one another. Here, we explain why this must change: Developmental researchers need to acknowledge how experimenter identity could be acting as an invisible, lurking variable, influencing the outcome and generalizability of studies. We provide simple suggestions for how researchers and journals can begin to address this issue, thereby improving the quality and depth of the work in our field.

Keywords: experimenter effects, open science, social cognition, implicit bias, impression formation, experimental control, experimental design, replicability

An invisible, lurking variable in developmental research: The importance of considering experimenter identity

Introduction

In a famous 1972 *Atlantic* article, sociolinguist William Labov took issue with a prevailing hypothesis at the time that differences in White and Black children's relative verbal skills at school were due to differences in verbal input at home. Labov argued that what appeared to be linguistic inaptitude on the part of Black children was instead a by-product of children's testing environment. He noted that, while Black children often appeared to lack communicative skills when tested by White teachers in an academic setting, in more relaxed settings with Black interviewers, these same children were seen to be much more loquacious. Even when the 'testing' situation remains constant, sociolinguists have long observed how the identity of an interviewer alone (and their relationship to an interviewee) can dramatically influence interviewees' behavior. This notion, that the very presence of an interviewer/experimenter influences the (verbal) behavior of a participant, is referred to by sociolinguists as the observer's paradox, and in the field of sociolinguistics, as in some other areas of social science, great care is taken when collecting data to mitigate the influence of interviewer identity on participants' behavior.

Experimental psychologists have similarly recognized for a long time the ways in which experimenters might inadvertently influence participants' behavior, and as a result, the field as a whole has adopted a number of practices to minimize experimenter effects (Masling, 1966; Rosenthal, 1966). For the most part, however, experimental psychologists have focused primarily on controlling an experimenter's *behavior* during an experiment, through such means as rehearsed scripts, neutral facial expressions, and relatively flat vocal affect. One experimenter

effect in particular, experimenter bias, has perhaps received the most attention; this refers to situations where an experimenter's behavior—which is unknowingly influenced by their knowledge about the hypotheses of the study—inadvertently prompts participants to respond in ways that are consistent with those hypotheses (Gilder & Heerey, 2018; Rosenthal, 1966). As a result of such awareness-raising, it has become customary in experimental work for researchers to limit experimenters' knowledge/behavior through practices like double blind studies, and to report on such methodological decisions in publications.

For the most part, however, most experimental researchers (including ourselves) tend to have a blindspot when it comes to reporting on the characteristics of their experimenters. This is particularly surprising, given some work that has specifically investigated the effects of experimenter identity on participants' behavior (see Innis & Fraser, 1971; McGuigan, 1963; Silverman, 1974 for discussions about this issue). For example, work with adults has demonstrated ways in which the race (e.g., Cilliers et al., 2015; Lowery et al., 2001; Marx & Goff, 2005; Nagendra et al., 2018; Thames et al., 2013), gender (e.g., Kállai et al, 2004; see Chapman et al., 2018 for a review), accent (Hay et al., 2009), age (Sindi et al., 2013), and sexual orientation (e.g., Berry, 2015; Cuenot & Fugita, 1982) of experimenters (among other traits) can differentially influence participants' responses across a wide variety of domains, and there is some evidence for similar effects in children (e.g., Corenblum & Annis, 1987; Greenberg & Gordon, 1983; Katz et al., 1975; Kwong See et al., 2012; Shatz & Gelman, 1973).

This issue should be especially concerning for developmental researchers. In many adult studies, participants can often simply read experimental instructions on a computer screen while experimenters wait in another room. In contrast, in developmental work, experimenters are much more heavily involved in guiding children through experiments, making the influence of

experimenter identity on children's behavior potentially more likely. In the current paper, we argue that it is time for developmental researchers to take into account the potential influence of experimental identity on children's behavior in their research. We begin by discussing in more detail why we believe researchers should consider—and importantly, report on—the attributes of their experimenters. In particular, we highlight how children may react differently to individual experimenters based on their identity, and how an experimenter's behavior, too, may differ as a result of whom they are testing. We then provide a survey of all of the experimental articles published in the last 3 years in 4 top developmental journals (*Child Development*, *Developmental Science*, *Developmental Psychology*, and *Journal of Experimental Child Psychology*), showing that the vast majority of studies fail to report on the attributes of their experimenters. Finally, we provide recommendations for how researchers can approach the issue of experimenter identity effects in their research.

Why should we care about experimenter attributes, anyway?

Developmental psychologists are well aware of children's sensitivity to the social properties of others (Rhodes & Baron, 2019), and how children's behavior can be influenced by their perceptions of others, making it all the more striking that researchers do not routinely report on the characteristics of experimenters. Starting in infancy, children begin to categorize individuals based on categories such as language (Kinzler, 2013), gender (Leinbach & Fagot, 1993), and race (Anzures et al., 2010), and make use of category membership to generate predictions and make inferences about newly encountered individuals (e.g., Schmidt et al., 2012; Weatherhead & White, 2018). Importantly, children's reactions to/behavior towards individuals can vary depending on their perceptions of those individuals. For example, infants are more

fearful of male strangers compared to female strangers (Benjamin, 1961; Hahn-Holbrook et al., 2010), and show more empathy for their mothers than strangers (Young et al., 1999; Zahn-Waxler et al., 1992). In one word learning study, infants were found to demonstrate more engagement in the task (repeating words and pointing to objects more) when interacting with their mother compared to a stranger, which ultimately led to better word learning (van Rooijen et al., 2019). Older children, too, behave differently based on who they are interacting with; elementary school-aged children, for example, make decisions about whom they should trust based on factors such as appearance (Bascandziev & Harris, 2014), and will adjust their language when speaking to adults vs. younger children (Shatz & Gelman, 1973). In other words, children's behavior in response to different individuals is influenced by how familiar those individuals are and by the inferences that children draw about them (based on the social categories they belong to).

In some cases, children's responses to others will depend on whether those individuals are perceived as fellow members of the same social category or not (i.e., whether they are in-group or out-group members). For example, children will more readily imitate (Buttelmann et al., 2013; de Klerk et al., 2019; Howard et al., 2015), learn from (Corriveau et al., 2013; Xiao et al., 2017), show greater empathy for (Masten et al., 2010), trust (Ma & Woolley, 2013), and are more credulous toward (McDonald & Ma, 2016) in-group members compared to out-group members. While children's sensitivity to specific social properties in such studies is typically examined through the manipulation of study stimuli, there is little reason to believe that children are not sensitive to the characteristics of experimenters themselves, and behave differently depending on whom they are interacting with (see Egalite et al., 2015, showing how low performing students benefit from being assigned to race-congruent teachers).

Moreover, children's inferences about an experimenter may not be driven just by properties of the experimenter themselves, but also how the experimenter interacts with others. Indeed, children make additional inferences about the character traits and likely group status of individuals based on how those individuals interact with others (third-party interactions). Children use nonverbal behaviors that others direct towards an individual (e.g., smiling, nodding, neutral expression, etc.) to evaluate how nice/smart that individual is (e.g., Brey & Shutts, 2018), and will even generalize their positive/negative evaluations to other group members (Skinner et al., 2020). In addition, infants expect people with shared preferences to be more likely to affiliate (Liberman et al., 2014; Liberman et al., 2021), expect members of the same social group to act alike (Powell & Spelke, 2013), and perceive of imitative behavior as prosocial and positive (Powell & Spelke, 2018). In the context of an experimental session, there is ample opportunity for a child to observe the experimenter's interactions with others and for this to influence their own impressions of and reactions to the experimenter. For example, before the experiment begins, the experimenter will need to interact with a child's parent (and potentially with other lab staff). Children may note the degree to which these interactions appear natural and aligned (particularly when the interactions are between parent and experimenter). Moreover, given that parents are typically co-present with children in the testing room, children might even pick up on parents' body language/reactions to the experimenter during the experiment itself. All of this may serve to influence children's impressions of the experimenter and the degree to which they feel affiliated with them. At a more fundamental level, a parent's demonstrated comfort or discomfort with an experimenter may be transferred to the child and this state may affect the child's performance on the task.

But of course, it is not just the child who may make inferences in a testing situation. Experimenters might also behave differently around different children based on their own inferences and biases against certain demographic groups (especially inexperienced or less well-trained experimenters). Research has shown, for example, that teachers' evaluations of (and behavior towards) students varies by such factors as gender and race (Okonofua & Eberhardt, 2015). For example, in the United States, teachers have been shown to direct more positive speech to European American students than Latino/a and African American students (Tenenbaum & Ruck, 2007), and when asked to search for and anticipate problematic behaviors in videos of preschoolers, are more likely to gaze longer at Black children (especially Black boys) than White children (Gilliam et al., 2016). And just as children's responses to others vary based on their group status (in- or out-group), so too may experimenters behave differently based on whether a child is a fellow in-group member or not. Black children, for example, have been shown to receive worse evaluations from White teachers than Black teachers (Bates & Glick, 2013; Downey & Pribesh, 2004). In both their interactions with children, and in tasks requiring subjective judgments of children, experimenters' biases may systematically influence how they respond to children from varying demographic groups.

Importantly, slight differences in how children are treated can in turn influence their behavior. In school contexts, researchers have long observed how teacher expectations about students can impact children's self-perceptions and achievement (i.e., the so-called Pygmalion effect; Rosenthal & Jacobson, 1968; see Johnston et al., 2019 and Wang et al., 2018 for recent reviews). For example, despite performing similarly to other ethnic groups at the beginning of a school year, Maori children in one study showed the least gains by the end of the year compared to children in other ethnic groups; critically, children performed in a way that aligned with

teachers' expectations of student achievement, which tended to be lower for Maori children than other ethnic groups (Rubie-Davies et al., 2010). Thus, slight differences in how teachers (and by extension experimenters) interact with kids can lead to measurable differences in children's performance on various tasks.

Controlling experimenter behavior cannot solve everything

However, even in an ideal world, where an experimenter's behavior was truly identical across all participants, there would still be the possibility that the same behavior could be interpreted differently by different children, depending on the identity of the experimenter. As children interact with experimenters throughout a study, their inferences about the experimenter might serve as a kind of lens through which they evaluate the experimenter's behavior (Harris et al., 1992). In other words, children's interpretations of an experimenter's behavior may depend on how their behavior aligns with children's expectations of their behavior (i.e., their stereotypes). For example, elementary school-aged children's perceptions of their relationships with their teachers have been shown to correlate with their expectations about their teacher in the summer before the new school year starts (Gurland & Evangelista, 2015). In addition, the rapport that children perceive having with an unfamiliar adult is related to their prior expectations of that unfamiliar adult when the behavior of the adult is controlled for (Gurland & Grolick, 2003), showing that children's interpretation of an individual's behavior depends on their a priori expectations about that individual.

Additional work on intergroup cognition has shown that children evaluate the same behaviors differently as a function of group membership. For example, children will punish selfish behavior from out-group members more than in-group members (Jordan et al., 2014), and

will protest more when in-group members violate conventional norms compared to out-group members (Schmidt et al., 2012). In addition, in ambiguous scenarios in which an actor may be committing a negative (e.g., stealing another child's money) or positive act (e.g., picking up and returning the money), children are significantly more likely to interpret a situation negatively when the actor is an out-group member compared to an in-group member (Dunham & Emory, 2014). Thus, even when meticulously controlling an experimenter's behavior, and instituting double-blind procedures, children may nevertheless interpret an experimenter's behavior differently based on their identity, leading to systematically different responses across different experimenters.

In the language sciences, too, a fair amount of work has shown that the same linguistic stimulus is perceived and evaluated differently depending on who listeners think is speaking. For example, the same vowel spoken by Detroit English speakers is perceived as more 'Canadian' when adult listeners believe the speaker is Canadian, but more American when they think the speaker is from Detroit (e.g., as in a word like *about*, whose vowel may be perceived as more stereotypically Canadian 'aboot' or not; Niedzielski, 1999; see also Strand, 1999 for similar work with gender). In addition, when asked to identify native-accented words embedded in noise, Canadian English-speaking participants are less accurate at identifying them when they are accompanied by an image of an Asian-Canadian person compared to a European-Canadian person, presumably due to a mismatch between actual vs. expected pronunciations (Babel & Russell, 2015; see also McGowan, 2015 for a similar study in a US context). Importantly, even children as young as 16 months will form expectations about how a person will speak based on their identity (Weatherhead & White, 2018). In addition, listeners' *evaluations* of speech also depend on who they think is speaking. In one famous study, Rubin (1992) found that participants

were more likely to rate native-accented speech as more 'accented' and less comprehensible when paired with the image of an Asian face compared to a Caucasian face, suggesting that evaluations of accent and comprehensibility are not entirely due to differences in processing the acoustic input, but also to listeners' subjective experience of the speech based on their expectations (see also Kang & Rubin, 2009 and Rubin, 2002 for a review).

Beyond the perception of speech sounds, listeners may also interpret the intentions behind what experimenters say differently, depending on who is speaking. For instance, typically, when children encounter a familiar and an unfamiliar object (e.g., an apple vs. a windmill), and hear an experimenter refer to one of the objects with an unfamiliar word (e.g., *Where is the dax?*), they will assume that the novel word applies to the unfamiliar and not the familiar object (otherwise the experimenter should have used the familiar word *apple*; Markman & Wachtel, 1988). However, if children believe that a speaker is somehow confused or not reliable, they may be less likely to make this assumption (Kwong See & Nicoladis, 2010). In this way, children may respond differently to the same rehearsed script in an experiment based on who is delivering the script and what their intentions/abilities might be.

Experimenter as part of the larger environment

In many ways, experimenter identity can also be considered an aspect of an overall lab environment that contributes to children's impressions of their surroundings. A good example comes from studies investigating stereotype threat, in which individuals underperform in certain tasks because they are anxious about confirming negative stereotypes about the group(s) they belong to (Steele & Aronson, 1995). For example, young girls tend to perform worse on math assessments when their gender identity is primed prior to assessment (e.g., having them color a

girl with a doll), suggesting that negative gender stereotypes (e.g., 'girls can't do math') impede young girls' performance in math (Ambady et al., 2001; see also Désert et al., 2009 and McKown & Weinstein, 2003 for stereotype effects related to SES and race/ethnicity, respectively). Critically, in the same way that an external object/activity (coloring a picture of a girl) activates an individual's identity and the negative stereotypes associated with it, so too can the identity of an experimenter also serve as a cue for such stereotypes. In work with adults, for example, Black participants have been found to perform worse on a verbal test when the test was administered by a White examiner compared to a Black examiner, presumably because the Black examiner alleviated the stereotype threat that test-takers felt (Marx & Goff, 2005). In an experimental setting, one could similarly imagine how the identity of an experimenter could either alleviate or exacerbate children's experience of stereotype threat, depending on the situation.

The makeup of lab staff more generally might also serve to make children's group membership more or less salient. Indeed, children from relatively smaller social groups are often more cognizant of group differences, and may develop stronger group biases than children from majority groups (Brown & Bigler, 2002), since they often stand out more in various social situations. As an example, Brown and Bigler (2002) discuss how a brown-haired girl entering a bus filled with mostly brown-haired individuals might attend less to hair color than a brown-haired girl entering a bus with mostly blond individuals, the idea being that minority status in any given situation makes in-group and out-group distinctions more salient. In a lab setting, seeing multiple lab staff members who all look different from them might make children more aware of the group differences between them and their experimenter than if they had only seen the experimenter.

Ultimately, given that children around the world grow up in highly variable cultural contexts, the specific ways in which experimenter identity may influence children's behavior will depend more broadly on their local socio-cultural environment, for example, their neighborhood demographics (Hwang et al., 2021), the social structure (Shutts et al., 2011), and how children are socialized (Lamm et al., 2018). Many of these socio-cultural factors which differ from locale to locale, and which undoubtedly influence children's perceptions of and interactions with different experimenters, may justifiably lead to divergent findings between labs and failures to replicate. Including demographic information about experimenters in manuscripts (in addition to participants) could help the field to consider the possible role of experimenter identity when confronted with mismatching findings.

Looking for diverse participants. What about experimenters?

In recent years, researchers, funding agencies, and journals alike (including this one) have admirably been pushing to get more diverse samples of participants represented in research, given that participants from highly educated families living in rich, urban, industrialized environments are overrepresented in experimental work (Fernald, 2010). Since this subpopulation sometimes performs quite differently from more diverse groups of participants (e.g., Henrich et al., 2010; Nielson et al., 2017), it is often difficult to make broad generalizations about human psychology based on such a narrow pool of participants. To overcome this limitation, there have been increased calls and efforts to recruit more diverse samples of participants (and report on them) by casting wider nets (e.g., through testing in multiple labs/countries or online; Frank et al., 2017; Rhodes et al., 2020), and actively recruiting

traditionally underrepresented groups of participants (e.g., multiracial/multiethnic, intersex, and transgender children; Dunham & Olson, 2016; Nishina & Witkow, 2020).

As we push to have more diverse samples, we need to become more cognizant than ever before of how the identity of an experimenter might differentially influence children from diverse backgrounds. Children might respond to the same experimenter differently depending on their own identity and life circumstances (Hwang et al., 2021). For example, in a recent study we conducted in a highly diverse, North American city, we tested the degree to which children's wait times in a delayed gratification task were influenced by experimenter identity, manipulating both linguistic (accent and grammaticality) and race cues (White or Asian) of the experimenters recruited for the experiment ([blinded]). Although overall, the identity of the experimenter was not found to influence children's wait times, when considering the race of the children in relation to that of the experimenter, a different pattern emerged. Specifically, the White children in our study, who represented 40% of the children tested, were significantly more likely to wait for the second treat after interacting with an Asian experimenter compared to a fellow White experimenter, suggesting that the children may have felt less at liberty to eat the first treat (should they have wanted to) after interacting with out-group members compared to in-group members. Had we not considered children's identity, and, more importantly, children's identity *in relation* to that of the experimenter, we would have missed crucial differences in the behavior of children from different demographic groups. Thus, researchers who set out to test a diverse sample of children may find that children do not behave differently simply because they are members of different demographic groups; rather, their behavior may vary based on how they perceive the experimenter in relation to themselves.

At the same time that researchers are calling for more diversified samples of participants, there is little to no discussion on similarly diversifying the pool of experimenters that run studies. This is despite the fact that there are ongoing efforts to try to involve more diverse groups of young researchers in various STEM fields (e.g., Tsui, 2007), with many institutions and funding agencies asking researchers to proactively work to increase participation from under-represented groups (e.g., the Natural Sciences and Engineering Research Council of Canada, 2017). While such efforts serve larger social goals of making science more accessible, dynamic, and representative of the population as a whole, they present additional methodological opportunities to incorporate more diverse groups of experimenters in our research. Indeed, just as researchers have argued that we need to recruit more diverse subject pools to better ensure the generalizability of our results (Henrich et al., 2010; Nielson et al., 2017), so too might researchers also want to make sure that the phenomena we observe extend beyond specific groups of experimenters that are frequently used to test children (e.g., in Western labs, White, female undergraduate students).

How often do experimenter attributes actually get reported?

To underscore the magnitude of the issue, we sought to quantify how often experimenter attributes (and which ones) actually get reported in developmental research. We further wanted to see how many studies explicitly report whether one experimenter ran all test subjects, or whether multiple experimenters were involved. To answer these questions, we surveyed the last 3 years of experimental articles (January 2019 – December 2021)—covering a breadth of areas in developmental research—in 4 top developmental journals (*Child Development*, *Developmental Psychology*, *Developmental Science*, and the *Journal of Experimental Child*

Psychology), and documented the types of details about experimenters (and also participants) that were reported in the studies. After excluding studies on animals and adults, meta-analyses, and reviews, we were left with 1521 articles (CD = 379, DP = 287, DS = 279, and ECP = 576;). A subset of the articles (15%, or 228 articles), which were coded by two (rather than one) coder, showed greater than 97% agreement across all dimensions. The data can be found at

https://osf.io/n5j7c/?view_only=21779e356e0946d9bf0ef932959efe60.

The vast majority of studies included an interaction with an experimenter (76% or 1154 articles), which included anything from experimenters explaining instructions to children (or simply asking if they understood computer presented instructions) to engaging in back-and-forth interactions throughout any of the experimental tasks reported in the methods section. An experimenter silently setting up experimental equipment or chatting with parents before the experiment began was not included (though see above for why that might matter). For each of the 1154 articles that included interactions with experimenters, we coded whether or not the authors explicitly mentioned whether the same experimenter(s) ran all the participants; we also coded whether the language(s), accent, gender, race, and exact age of both experimenters and participants were explicitly reported in the methods section (or any of the methods sections in studies with multiple experiments). Using only pronouns to refer to experimenters did not count as mentioning experimenter gender, since both male and female pronouns are sometimes used as generic pronouns. With regards to race, we included—in addition to explicitly outlined race information—both partially outlined (e.g., 80% of participants were X) and regionally defined information for participants (e.g., participants were from town/city Y, which is predominantly Z). Table 1 shows the results of the survey.

[TABLE 1 NEAR HERE]

Only 6% of studies explicitly mentioned whether one experimenter (vs. multiple experimenters) ran all the participants. Notably, there is a dearth of studies including information on the attributes of the experimenter, with most attributes being reported less than one percent of the time (and the most reported attributes still less than 10% of the time). In contrast, studies were substantially more likely to report on the attributes of participants, with the exception of accent, which is rarely reported for both experimenters and participants. Even still, given the push to provide thorough descriptions of participant samples in developmental work, there are nevertheless some attributes which appear to be relatively under-reported (e.g., language and race), in comparison to attributes that are reported in over 90% of studies (age and gender), though to be fair, the language used by experimenters and participants is often easily inferable from experiments (we only coded for explicitly mentioned language background).

We next examined a small subset of studies which explicitly focused on one of the characteristics of interest. To be considered in this subset, the study had to examine, as part of the main research question(s), either (1) the influence of one of the characteristics on some measured variable(s) (e.g., children's abilities to detect emotion in own-race and other-race faces; Segal et al., 2019), or some theoretical aspect of the characteristic itself (e.g., whether story time closer to bed time improves word learning after sleep; Henderson et al., 2021). Interestingly, even when looking at studies focusing on one of the characteristics of interest, we still find that the majority of studies fail to report on those same attributes of the experimenter(s), as seen in Table 2.

[TABLE 2 NEAR HERE]

Altogether, these results confirm that experimenter attributes are extremely underreported in developmental studies, both overall and in comparison to participant attributes.

The almost complete failure to mention accent for both experimenters and participants may suggest that researchers generally consider accent to be a negligible variable, despite numerous studies showing how accent can dramatically influence children's behavior, even beyond work specifically designed to test linguistic phenomena. For example, children are more likely to believe (McDonald & Ma, 2016), imitate (Kinzler et al., 2011; Wagner, Dunfield, & Rohrbeck, 2014), play with (Myers-Burg & Behrend, 2021), and want to be friends with (Kinzler et al., 2007; St. Pierre & Johnson, 2020) native-accented speakers of their language over foreign-accented speakers, presumably because children generally have preferences for in-group members over out-group members (Dunham, 2018). During late childhood, children even begin developing specific attitudes towards particular dialects, with 10-year-olds, for example, tending to rate children with Southern American accents as more friendly, but less intelligent than northern-accented children, consistent with adult stereotypes of those accents (Kinzler & DeJesus, 2013). Given that the accent of an individual's speech also affects how friendly they are perceived to be (Giles & Watson, 2013), how much listeners attend to and remember what they say (Lev-Ari, 2015), and what listeners assume they have knowledge of (e.g., Rett & White, 2019), it is clear that the accent of an experimenter has the potential to influence children's responses along a variety of different measures. Beyond the social inferences children might make based on an experimenter's accent, we must also consider children's abilities to cope with various accents based on their previous language/accent exposure (see Johnson et al., 2022 for a review). As researchers who study children's perceptions of accents, we find the omission of accent information both in children and experimenters particularly striking (and something that we are guilty of ourselves!).

What should we do?

Given that experimenter identity has real potential to influence children's behavior in developmental work, where should we go from here? Below we outline several steps that researchers can immediately take to address this important issue:

1) Report on experimenter attributes

Even if researchers are not interested in potential effects of experimenter identity, or if they do not believe that experimenter identity could conceivably influence the manipulated variables of interest, we nevertheless recommend that researchers report demographic information about their experimenter(s) in their studies. Some journals (including *Infant and Child Development*) already require researchers to report demographic information about participants, even if such variables are not of theoretical interest and not considered in analyses. *Child Development*, for example, requests that authors include demographic information about participants such as "sex, SES, race or ethnicity" (https://www.srcd.org/research/journals/child-development/child-development-submission-guidelines) and *Developmental Psychology* says that "major demographic characteristics should be reported, such as sex, age, socioeconomic status, race/ethnicity, and, when possible and appropriate, disability status and sexual orientation" in order to "provide a more complete understanding of the sample and of the generalizability of the findings" (https://www.apa.org/pubs/journals/dev/index?tab=1). In a similar vein, demographic information about experimenters should additionally be included, at the very least attributes that are typically reported for participants such age, sex, and race or ethnicity. Attributes which seem easily inferable should nevertheless be stated explicitly to avoid potential confusion. In this way, readers can get a sense of how the identity of the experimenter relates to that of the participants. We strongly recommend that researchers also begin providing information about the language

background/accent of experimenters (and also participants!), given that even subtle differences in pronunciation can influence children's evaluations and comprehension (e.g., Fennell & Byers-Heinlein, 2014; Kinzler, 2021; Schmale et al., 2010). When appropriate, researchers should additionally report on other experimenter attributes (e.g., disability status, SES status growing up, sexual orientation, etc.), especially if they think they could conceivably influence children's behavior. Researchers' decisions to report on particular attributes or not should consider which experimenter attributes might be important for their particular culture and subdiscipline.

That being said, we acknowledge that there are complications inherent to reporting experimenter attributes, given that many characteristics (e.g., race, accent, and gender) do not readily divide into discrete categories, and the ways in which experimenters identify themselves (and how others perceive them) may vary from context to context (e.g., Pauker et al., 2018; Remedios & Sanchez, 2018). In addition, many cultures and countries might be more or less hesitant to collect and report on some attributes (e.g., race or health status), and journals should be respectful of these cultural differences. Nevertheless, researchers should strive to be as open and transparent as possible about who their experimenters are (as appropriate for their cultural context) in order to provide a more complete picture of the studies they report on.

An obvious benefit to reporting experimenter identity, and providing a more complete picture of how an experiment was run, relates to issues of replicability. In recent years, the field of psychology has struggled with a replicability crisis, in which the results of many classic studies have been found to be difficult or impossible to replicate (Open Science Collaboration, 2015). While in many cases, an inability to replicate may be due to original spurious results, in other cases, a failure to replicate might instead be due to small changes in the experimental procedure between labs. One such factor that could contribute to replication failure is

experimenter identity, which may influence children's behavior in subtle ways from lab to lab, leading to successful replications in some instances but not others. Being more open about the identity of experimenters will help researchers better understand the potential factors that may influence children's behavior across different experiments, and help move the field forward.

2) Properly account for multiple experimenters

In cases where multiple experimenters *are* involved in a study, researchers should take steps to ensure that the results they obtain are likely not due to experimenter effects. First, it is not only important for experimenters to be blind to the conditions that they are running—as has been discussed extensively in the literature (e.g., Rosenthal, 1966)—it is also critical that each experimenter run relatively equal numbers of participants within each condition, so that differences between conditions can be confidently attributed to the experimental manipulation, rather than to differences caused by different experimenters running each condition. In their manuscripts, researchers should report how many participants each experimenter ran in each condition (including the number of drop-outs). Additionally, when possible, researchers should include an effect of experimenter—either fixed or random—in their statistical models, in order to be more confident that any observed effect of theoretical interest is not caused by the undue influence of a minority of experimenters but generalizes across all experimenters. Such design features should be specified in pre-registrations when possible.

In an ideal world, experiments would be conducted with many different experimenters in order to ensure the generalizability of results, but this is typically not a viable option for most labs. In recent years, however, there has been an uptick in the number of large-scale, multi-lab studies being conducted, including Many Labs (Klein et al., 2014) and Psychological Science

Accelerator (Moshontz et al., 2018) for mostly adult studies and ManyBabies (Frank et al., 2017) for developmental research. While a primary goal behind these endeavors is to conduct highly powered studies whose results generalize across many different lab contexts, these types of studies present additional opportunities to investigate how slight differences between labs (e.g., differences in lighting) influences children's behavior. These large-scale, highly powered studies spanning across multiple labs could become fruitful places to explore how the identity of experimenters—which has not yet been considered by the ManyBabies Consortium—influences children's behavior in different types of paradigms. For example, a large-scale study could explore the potential influence of various experimenter attributes on drop-out rates, a topic that would be of interest to developmental psychologists in general.

3) Consider the theoretical implication of experimenter effects

Finally, rather than treating the identity of an experimenter as a nuisance variable that needs to be controlled for, many areas of psychology could benefit from considering the identity of an experiment as a variable worth investigating in its own right. In many cases, rather than include live interactants at all, researchers will instead present stimuli to children in the form of disembodied audio recordings or images/videos on a computer screen in order to exercise as much control as possible over their experiment. In some cases, the ways in which children react to live interactants may vary considerably from their reactions to the same stimuli presented audiovisually, so the use of live interactants (i.e., experimenters) might lead to new, more ecologically valid insights about a broad range of phenomena that cannot be obtained from using more traditional stimuli alone. As a case in point, most studies investigating language learning in a laboratory setting typically present children with audiovisual input, but in some studies,

researchers have found evidence of learning only when exposing children to live interactants rather than disembodied audio (e.g., Paquette-Smith et al., 2021; Roseberry et al., 2014). For example, Kuhl et al. (2003) found that English-learning 6- to 12-month-olds were able to perceive a phonetic contrast in Mandarin speech only after exposure to live, but not recorded, speakers of Mandarin, highlighting how the presence of live social partners qualitatively differs from more traditional means of stimuli presentation.

In a recent study that we conducted, we showed how the presence of live interactants might provide additional insights on children's perceptions of other individuals beyond what can be shown using audio-visual input alone. During the experiment, 5- to 7-year-old children were told a surprising (and false) claim by either a native- or a foreign-accented experimenter. Upon discovering that the claim of the experimenter was false, children were more likely to directly confront a native experimenter compared to a non-native experimenter, presumably because (the native-accented) children found it less threatening to disagree with a fellow native-accented experimenter compared to the non-native-accented experimenter ([blinded]). Such a result is unlikely to be observed in more traditional methods; only through the use of live experimenters were we able to examine how children responded to in-group and out-group members in a more real-world situation.

Conclusion

Developmental researchers, and experimentalists more generally, strive to control as much of their experiments as possible. This means thoroughly vetting children to make sure they fulfill the experimental criteria, confirming that coding is systematic and reliable, and ensuring that the lab environment is consistent across participants. Researchers even think about the time

of day or season that children are tested (Frank et al., 2017; Johnson & Zamuner, 2010; van der Velde & Junge, 2020), and ensure that children in one condition are not tested at systematically different times than children in another condition. Despite the increasingly stringent control that researchers exercise over their experiments, most studies fail to consider (or at least report on) the identity of the experimenter. In the current paper, we have argued that that experimenter identity should be taken more seriously by developmental researchers, given the very real potential for experimenter identity to systematically influence children's behavior in experiments.

As an immediate step, researchers can report on the demographics of their experimenters in their studies, and where appropriate, include them in statistical analyses. As a longer-term goal, we should make concerted efforts to diversify experimenter pools, not simply to make our results more generalizable, but to fulfill important social goals as well. Not only would this help to immediately expand participation in STEM fields to more diverse groups of individuals, it might additionally encourage children from backgrounds underrepresented in STEM to pursue a career in STEM themselves. Indeed, for many children—particularly potential first-generation university students—participating in university experiments offers them a unique window into what life as a scientist looks like, and having a diverse lab staff which resembles the population of children that it tests could make all the difference in the world for inspiring a more diverse range of future scientists.

References

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, *12*(5), 385-390.

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., & Lee, K. (2010). Categorization, categorical perception, and asymmetry in infants' representation of face race. *Developmental Science*, *13*(4), 553-564.

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823-2833.

Bascandziev, I., & Harris, P. L. (2014). In beauty we trust: Children prefer information from more attractive informants. *British Journal of Developmental Psychology*, *32*(1), 94-99.

Bates, L. A., & Glick, J. E. (2013). Does it matter if teachers and schools match the student? Racial and ethnic disparities in problem behaviors. *Social Science Research*, *42*(5), 1180-1190.

Benjamin, J. D. (1961) Some developmental observations relating to the theory of anxiety. *Journal of the American Psychoanalytic Association*, *9*(4), 652-668.

Berry, B. A. (2015). Experimenter characteristics, social desirability, and the Implicit Association Test. *Psi Chi Journal of Psychological Research*, *20*(4), 247-257.

Brey, E., & Shutts, K. (2018). Children use nonverbal cues from an adult to evaluate peers. *Journal of Cognition and Development*, *19*(2), 121-136.

Brown, C. S., & Bigler, R. S. (2002). Effects of minority status in the classroom on children's intergroup attitudes. *Journal of Experimental Child Psychology*, *83*(2), 77-110.

Buttelmann, D., Zmyj, N., Daum, M., & Carpenter, M. (2013). Selective imitation of in-group over out-group members in 14-month-old infants. *Child Development*, *84*(2), 422-428.

Chapman, C. D., Benedict, C., & Schiöth, H. B. (2018). Experimenter gender and replicability in science. *Science Advances*, *4*(1), e1701427.

Cilliers, J., Dube, O., & Siddiqi, B. (2015). The white-man effect: How foreigner presence affects behavior in experiments. *Journal of Economic Behavior & Organization*, *118*, 397-414.

Corenblum, B., & Annis, R. C. (1987). Racial identity and preference in Native and White Canadian children. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, *19*(3), 254-265.

Corriveau, K. H., Kinzler, K. D., & Harris, P. L. (2013). Accuracy trumps accent in children's endorsement of object labels. *Developmental Psychology*, *49*(3), 470-479.

Cuenot, R. G., Fugita, S. S. (1982). Perceived homosexuality: Measuring heterosexual attitudinal and nonverbal reactions. *Personality and Social Psychology Bulletin*, *8*(1), 100-106.

de Klerk, C. C. J. M., Bulgarelli, C., Hamilton, A., & Southgate, V. (2019). Selective facial mimicry of native over foreign speakers in preverbal infants. *Journal of Experimental Child Psychology*, *183*, 33-47.

Désert, M., Préaux, M., & Jund, R. (2009). So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. *European Journal of Psychology of Education*, *24*, 207.

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, *77*(4), 267-282.

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Science*, *22*(9), 780-793.

Dunham, Y., & Emory, J. (2014). Of affect and ambiguity: The emergence of preference for arbitrary ingroups. *Journal of Social Issues*, *70*(1), 81-98.

Dunham, Y., & Olson, K. R. (2016). Beyond discrete categories: Studying multiracial, intersex, and transgender children will strengthen basic developmental science. *Journal of Cognition and Development*, *17*(4), 642-665.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44-52.

Fennell, C., & Byers-Heinlein, K. (2014). You sound like mommy: Bilingual and monolingual infants learn words best from speakers typical of their language environments. *International Journal of Behavioral Development*, *38*(4), 309-316.

Fernald, A. (2010). Getting beyond the "convenience sample" in research on early cognitive development. *Behavioral and Brain Sciences*, *33*(2-3), 91-92.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421-435.

Gilder, T. S. E., & Heerey, E. A. (2018). The role of experimenter belief in social priming. *Psychological Science*, *29*(3), 403-417.

Giles, H., & Watson, B. M. (Eds.). (2013). *The social meanings of language, dialect, and accent: International perspectives on speech styles*. Peter Lang.

Gilliam, W. S., Maupin, A. N., Reyes, C. R., Accavitti, M., & Shic, F. (2016). *Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions?* Yale University Child Study Center.

Greenberg, R. P., & Gordon, M. (1983). Examiner's sex and children's Rorschach productivity. *Psychological Reports*, *53*(2), 355-357.

Gurland, S. T., & Evangelista, J. E. (2015). Teacher-student relationship quality as a function of children's expectancies. *Journal of Social and Personal Relationships*, *32*(7), 879-904.

Gurland, S. T., & Grolnick, W. S. (2003). Children's expectancies and perceptions of adults: Effects on rapport. *Child Development*, *74*(4), 1212-1224.

Hahn-Holbrook, J., Holbrook, C., & Bering, J. (2010). Snakes, spiders, strangers: How the evolved fear of strangers may misdirect efforts to protect children from harm. In J. M. Lampinen & K. Sexton-Radek (Eds.), *Protecting children from violence: Evidence-based interventions* (pp. 263-289). Psychology Press.

Harris, M. J., Milich, R., Corbitt, E. M., Hoover, D. W., & Brady, M. (1992). Self-fulfilling effects of stigmatizing information on children's social interactions. *Journal of Personality and Social Psychology*, *63*(1), 41-50.

Hay, J., Drager, K., & Warren, P. (2009). Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics*, *29*(2), 269-285.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Henderson, L. M., van Rijn, E., James, E., Walker, S., Knowland, V. C. P., & Gaskell, M. G. (2021). Timing storytime to maximize children's ability to retain new vocabulary. *Journal of Experimental Child Psychology*, *210*, 105207.

Howard, L. H., Henderson, A, M. E., Carrazza, C. & Woodward, A. L. (2015). Infants' and young children's imitation of linguistic in-group and out-group informants. *Child Development*, *86*(1), 259-275.

Hwang, H. G., Debnath, R., Meyer, M., Salo, V. C., Fox, N. A., & Woodward, A. (2021). Neighborhood racial demographics predict infants' neural responses to people of different races. *Developmental Science*, *24*(4), e13070.

Innes, J. M., & Fraser, C. (1971). Experimenter bias and other possible biases in psychological research. *European Journal of Social Psychology*, *1*(3), 297-310.

Johnson, E. K., van Heugten, M., & Buckler, H. (2022). Navigating accent variation: A developmental perspective. *Annual Review of Linguistics*, *8*, 365-387.

Johnson, E. K., & Zamuner, T. S. (2010). Using infant and toddler testing methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 73–93). John Benjamins Publishing Company.

Johnston, O., Wildy, H., & Shand, J. (2019). A decade of teacher expectations research 2008-2018: Historical foundations, new developments, and future pathways. *Australian Journal of Education*, *63*(1), 44-73.

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(35), 12710-12715.

Kállai, I., Barke, A., & Voss, U. (2004). The effects of experimenter characteristics on pain reports in women and men. *Pain*, *112*(1), 142-147.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*(4), 441-456.

Katz, P. A., Sohn, M., & Zalk, S. R. (1975). Perceptual concomitants of racial attitudes in urban grade-school children. *Developmental Psychology*, *11*(2), 135-144.

Kinzler, K. D. (2013). The development of language as a social category. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 314–317). Oxford University Press.

Kinzler, K. D. (2021). Language as a social cue. *Annual Review of Psychology*, *72*, 241-264.

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Children's selective trust in native-accented speakers. *Developmental Science*, *14*(1), 106-111.

Kinzler, K. D., & DeJesus, J. (2013). Northern = smart and Southern = nice: The development of accent attitudes in the United States. *Quarterly Journal of Experimental Psychology*, *66*(6), 1146-1158.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(30), 12577-12580.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., …

Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142-152.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9096-9101.

Kwong See, S. T., & Nicoladis, E. (2010). Impact of contact on the development of children's positive stereotyping about aging language competence. *Educational Gerontology*, *36*(1), 52-66.

Kwong See, S. T., Rasmussen, C., & Pertman, S. Q. (2012). Measuring children's age stereotyping using a modified Piagetian conservation task. *Educational Gerontology*, *38*(3), 149-165.

Labov, W. (1972). Academic ignorance and black intelligence. *The Atlantic Monthly*, *229*(6), 59-67.

Lamm, B., Keller, H., Teiser, J., Gudi, H., Yovsi, R. D., Freitag, C., Poloczek, S., Fassbender, I., Suhrke, J., Teubert, M., Vöhringer, I., Knopf, M., Schwarzer, G., & Lohaus, A. (2018). Waiting for the second treat: Developing culture-specific modes of self-regulation. *Child Development*, *89*(3), e261-e277.

Leinbach, M. D., & Fagot, B. I. (1993). Categorical habituation to male and female faces: Gender schematic processing in infancy. *Infant Behavior and Development*, *16*(3), 317-332.

Lev-Ari, S. (2015). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, *5*, 1546.

Liberman, Z., Kinzler, K. D., & Woodward, A. L. (2014). Friends or foes: Infants use shared evaluations to infer others' social relationships. *Journal of Experimental Psychology: General*, *143*, 966-971.

Liberman, Z., Kinzler, K. D., & Woodward, A. L. (2021). Origins of homophily: Infants expect people with shared preferences to affiliate. *Cognition*, *212*, 104695.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology, 81*(5), 842-855.

Ma., L., & Woolley, J. D. (2013) Young children's sensitivity to speaker gender when learning from others. *Journal of Cognition and Development*, *14*(1), 100-119.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121-157.

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology*, *44*(4), 645-657.

Masling, J. (1966). Role-related behavior of the subject and psychologist and its effects upon psychological data. *Nebraska Symposium on Motivation*, *14*, 67-103.

Masten, C. L., Gillen-O'Neel, C., Brown, C. S. (2010). Children's intergroup empathic processing: The roles of novel ingroup identification, situational distress, and social anxiety. *Journal of Experimental Child Psychology*, *106*(2-3), 115-128.

McDonald, K. P., & Ma, L. (2016). Preschoolers' credulity toward misinformation from ingroup versus outgroup speakers. *Journal of Experimental Child Psychology*, *148*, 87-100.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502-521.

McGuigan, F. J. (1963). The experimenter: A neglected stimulus object. *Psychological Bulletin*, *60*(4), 421-428.

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development*, *74*(2), 498-515.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C., R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501-515.

Myers-Burg, M. R., & Behrend, D. A. (2021). More than just accent? The role of dialect words in children's language-based social judgments. *Journal of Experimental Child Psychology*, *204*, 105055.

Nagendra, A., Twery, B. L., Neblett, E. W., Mustafic, H., Jones, T. S., Gatewood, D., & Penn, D. L. (2018). Social cognition and African American men: The roles of perceived discrimination and experimenter race on task performance. *Psychiatry Research*, *259*, 21-26.

Natural Sciences and Engineering Research Council of Canada. (2017). Guide for applicants: Considering equity, diversity, and inclusion in your application. https://www.nserc-crsng.gc.ca/_doc/EDI/Guide_for_Applicants_EN.pdf

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62-85.

Nielson, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31-38.

Nishina, A., & Witkow, M. R. (2020). Why developmental researchers should care about biracial, multiracial, and multiethnic youth. *Child Development Perspectives*, *14*(1), 21-27.

Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science*, *26*(5), 617-624.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943.

Paquette-Smith, M., Cooper, A., & Johnson, E. K. (2021). Targeted adaptation in infants following live exposure to an accented talker. *Journal of Child Language*, *48*(2), 325-349.

Pauker, K., Meyers, C., Sanchez, D. T., Gaither, S. E., & Young, D. M. (2018). A review of multiracial malleability: Identity, categorization, and shifting racial attitudes. *Social and Personality Psychology Compass*, *12*(6), e12392.

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(41), E3965-E3972.

Powell, L. J., & Spelke, E. S. (2018). Third party preferences for imitators in preverbal infants. *Open Mind, 2*, 61-71

Remedios, J. D., & Sanchez, D. T. (2018). Intersectional and dynamic social categories in social cognition. *Social Cognition*, *36*(5), 453-460.

Rett, A. & White, K. S. (2019). Who knows what to call this toy? Children's learning from native and foreign-accented speakers. Poster presented at Society for Research in Child Development, Baltimore, MD.

Rhodes, M., & Baron, A. (2019). The development of social categorization. *Annual Review of Developmental Psychology*, *1*, 359-386.

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development*, *21*(4), 477-493.

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child Development*, *85*(3), 956-970.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*. Appleton-Century-Crofts.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart and Winston.

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2010). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, *76*(3), 429-444.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511-531.

Rubin, D. L. (2002). Help! My professor (or doctor or boss) doesn't talk English! In J. Martin , T. Nakayama , & L. Flores (Eds.), *Readings in intercultural communication: Experiences and contexts* (pp. 127-137). McGraw-Hill.

Schmale, R., Cristià, A., Seidl, A., & Johnson, E. K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy*, *15*(6), 650-662.

Schmidt, M. F. H., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, *124*(3), 325-333.

Segal, S. C., Reyes, B. N., Gobin, K. C., & Moulson, M. C. (2019). Children's recognition of emotion expressed by own-race versus other-race faces. *Journal of Experimental Child Psychology*, *182*, 102-113.

Shatz, M., & Gelman, R. (1973). The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the Society for Research in Child Development*, *38*(5), 1-38.

Shutts, K., Kinzler, K. D., Katz, R. C., Tredoux, C., & Spelke, E. S. (2011). Race preferences in children: Insights from South Africa. *Developmental Science*, *14*(6), 1283-1291.

Silverman, I. (1974). The experimenter: A (still) neglected stimulus object. *The Canadian Psychologist/Psychologie canadienne*, *15*(3), 258-270.

Sindi, S., Fiocco, A. J., Juster, R.-P., Pruessner, J., & Lupien, S. J. (2013). When we test, do we stress? Impact of the testing environment on cortisol secretion and memory performance in older adults. *Psychoneuroendocrinology*, *38*(8), 1388-1396.

Skinner, A. L., Olson, K. R., & Meltzoff, A. N. (2020). Acquiring group bias: Observing other people's nonverbal signals can create social group biases. *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes*, *119*(4), 824-838.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811.

St. Pierre, T., & Johnson, E. K. (2020). The development of accent-based friendship preferences: Age and language exposure matter. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong

(Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 2635-2641). Cognitive Science Society.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86-100.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, *99*(2), 253-273.

Thames, A. D., Hinkin, C. H., Byrd, D. A., Bilder, R. M., Duff, K. J., Mindt, M. R., Arentoft, A., & Streiff, V. (2013). Effects of stereotype threat, perceived discrimination, and examiner race on neuropsychological performance: Simple as black and white? *Journal of the International Neuropsychological Society*, *19*(5), 583-593.

Tsui, L. (2007). Effective strategies to increase diversity in STEM fields: A review of the research literature. *The Journal of Negro Education*, *76*(4), 555-581.

Van der Velde, B., Junge, C. (2020). Limiting data loss in infant EEG: Putting hunches to the test. *Developmental Cognitive Neuroscience*, *45*, 100809.

van Rooijen, R., Bekkers, E., & Junge, C. (2019). Beneficial effects of the mother's voice on infants' novel word learning. *Infancy*, *24*(6), 838-856.

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation*, *24*(3-5), 124-179.

Wagner, L., Dunfield, K. A., & Rohrbeck, K. L. (2014). Children's use of social cues when learning conventions. *Journal of Cognition and Development*, *15*(3), 479-494.

Weatherhead, D. & White, K. S. (2018). And then I saw her race: Race-based expectations affect infants' word processing. *Cognition*, *177*, 87-97.

Xiao, N. G., Wu, R., Quinn, P. C., Liu, S., Tummeltshammer, K. S., Kirkham, N. Z., Ge, L., Pascalis, O., & Lee, K. (2017). Infants rely more on gaze cues form own-race than other-race adults for learning under uncertainty. *Child Development*, *89*(3), e229-e244.

Young, S. K., Fox, N. A., & Zahn-Waxler, C. (1999). The relations between temperament and empathy in 2-year-olds. *Developmental Psychology*, *35*(5), 1189-1197.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology*, *28*(1), 126-136.

Table 1. Articles explicitly mentioning experimenter and participant attributes

Trait	Experimenters	Participants
Accent	4 (<1%)	12 (1%)
Age	5 (<1%)	1152 (99.8%)
Language	112 (9.7%)	506 (43.8%)
Gender	105 (9.1%)	1094 (94.8%)
Race	14 (1.2%)	618 (53.6%)

Note: Of the 1154 articles including an interaction with an experimenter, we report the number

of studies which report on the language, accent, race, age, and gender of experimenters and/or

participants.

Table 2. Reported experimenter traits for studies focusing on those attributes

Studies focusing on…	Experimenter trait reported
Accent (10)	2 (10%)
Gender (25)	6 (24%)
Language (204)	28 (13.7%)
Race (39)	6 (15.4%)

Note: The left column represents the number of studies which focused on the variable of interest, and the right column indicates the number of papers which reported those attributes in the experimenter(s). Across all attributes, studies appear to underreport on the characteristics of the experimenter.