Developmental Review 76 (2025) 101189

Contents lists available at ScienceDirect

Developmental Review

journal homepage: www.elsevier.com/locate/dr

Review

A systematic review of modern measures for capturing children's ethnic and racial attitudes, stereotypes, and discrimination

Eren Fukuda [a,*], Katharine E. Scott [a,b], Katherine L. Swerbenski [a], Nicole Huth [a,c], Kierin C. Barnett [a], Natalie Sarmiento [a], Madeline A. Henkel [a], Kristin Shutts [a]

[a] University of Wisconsin-Madison, USA
[b] Wake Forest University, USA
[c] Boston University, USA

ARTICLE INFO

Keywords:
Race
Ethnicity
Social biases
Children
Measurement

ABSTRACT

In recent years, there have been accelerated efforts among developmental scientists to understand and address children's ethnic and racial attitudes, stereotypes, and discrimination. For such efforts, using high-quality and context-appropriate measures is critical. However, focused discussions and investigations of measures for capturing children's ethnic and racial attitudes, stereotypes, and discrimination are scant. Accordingly, we conducted a systematic review of 1,001 measures that were used in 403 journal articles published between 2010 and 2022. Our review was guided by four questions: (1) What types of measures of children's ethnic and racial attitudes, stereotypes, and discrimination are being used by researchers?; (2) How do measures represent target groups?; (3) In which geographic and demographic contexts are measures being used?; and (4) What evidence do we have about some of the psychometric properties of commonly used scales/tasks? In seeking answers to these questions, we found both strengths and problems with our field's toolkit of measures. Taken together, our review provides an overview of modern measures for capturing children's ethnic and racial attitudes, stereotypes, and discrimination; offers initial insights about the characteristics and psychometric properties of those measures; and makes recommendations for future efforts in the field. We argue that measurement evaluation is a fertile avenue for future work in our field and that widespread discussions about measurement are necessary to advance the science of how children feel, think about, and behave toward members of different social groups.

Introduction

There are growing efforts among developmental scientists to elucidate the nature of children's ethnic and racial attitudes, stereotypes, and discrimination. Such efforts support the creation of both descriptive and theoretical accounts of how common social group phenomena (e.g., in-group favoritism) emerge and develop (Baron & Banaji, 2006; Fitzgerald et al., 2019; Kinzler & Spelke, 2011; Raabe & Beelmann, 2011; Skinner & Meltzoff, 2018). Studies of children's ethnic and racial attitudes, stereotypes, and discrimination also address children's experiences as targets of social biases—including whether and which interventions can reduce harm experienced by members of stigmatized groups (Aboud et al., 2012; Marcelo & Yates, 2019; Pachter et al., 2010; Trent et al.,

* Corresponding author.
E-mail address: efukuda@wisc.edu (E. Fukuda).

https://doi.org/10.1016/j.dr.2025.101189
Received 23 April 2024; Received in revised form 30 January 2025;
Available online 30 March 2025

2019; Umaña-Taylor, 2016; Vittrup & Holden, 2011).

A critical tool for these growing efforts is *measurement* of children's ethnic and racial attitudes, stereotypes, and discrimination. Measures are key to discovering the prevalence, magnitude, and quality of children's attitudes, stereotypes, and discrimination; measures are also necessary for capturing changes in children's social biases resulting from interventions. Thus, any researcher hoping to study or intervene on social group phenomena in childhood and adolescence needs to think about measurement: what tools are available, how different tools have been deployed, and what we know about the quality of those tools.

A promising way to understand the availability, usage, and quality of measures in the field is via systematic review. Systematic review allows one to capture all available tools and determine which constructs have been measured, how, and where; such information is critical to those planning new empirical and intervention work. Systematic review can also identify issues that need to be addressed in the field going forward (e.g., What characteristics of measures need to be considered? Which populations have been ignored?). To our knowledge, although there are existing narrative reviews (e.g., Tredoux et al., 2009) and empirical evaluations focused on select measures (e.g., Clark et al., 2017; deMayo & Olson, 2024; Guerrero et al., 2010; Rae & Olson, 2018; Williams & Steele, 2016, 2019), no work thus far has provided a systematic review aimed specifically at taking stock of the field's measurement toolkit for capturing children's ethnic and racial attitudes, stereotypes, and discrimination.

Systematic review can also shed light on the evidence base we have for determining how different measures perform (i.e., psychometric properties). Developmental scientists need to know about and use psychometrically sound methods for elucidating phenomena and for testing the effectiveness of interventions (American Educational Research Association, 2014). For example, when examining individual differences in out-group attitudes or testing the effects of a bias intervention program, it is crucial to use measures that capture stable attitudes or true change in the construct rather than error (i.e., test–retest reliability). Additionally, to draw generalizable inferences about children's behaviors or thought processes outside of the study context, developmental scientists need to create and use measures with a careful eye toward the validity of conclusions that can be drawn (i.e., construct and predictive validity). Unfortunately, there is a lack of widespread discussions and examinations across the field that focus on measures' psychometric properties. Without careful consideration of these issues, the field's ability to make theoretical and practical contributions is severely constricted.

Taken together, a systematic review is necessary to assess what we know about measuring children's ethnic and racial attitudes, stereotypes, and discrimination; identify gaps; spur critical discussions about measurement; and prompt more careful examinations of and considerations for measure properties and generalizability of results. To this end, we conducted a systematic review to capture the landscape of measurement usage in recent years and produce information about the psychometric properties of common measures.

Scope of Review

In the current paper, we focus on reviewing quantitative measures of children's attitudes, stereotypes, and discrimination involving ethnic and racial groups; these three constructs are often conceptualized as the core components of intergroup bias (Dovidio et al., 2010; Dovidio & Gaertner, 2010). We define ethnic and racial groups as ascribed or voluntary groups of people with shared physical features or common history, nationality, geography, language, and culture (Graham & Echols, 2018). Other types of social groups such as gender and novel or "minimal" groups (i.e., fictional groups created in the lab; Dunham, 2018) are outside the scope of this review.

We define attitudes as affective responses to a group, including negative evaluations which are often referred to as "prejudice" in the literature (Correll et al., 2010; Dovidio et al., 2010; Dovidio & Gaertner, 2010). We use the term "attitudes" rather than "prejudice" because we intended to include not only measures that capture negative affect but also those that capture positive affect toward groups (e.g., liking). We define ethnic and racial stereotypes as associations of attributes with groups, regardless of the valence of those attributes (Correll et al., 2010; Dovidio et al., 2010; Dovidio & Gaertner, 2010). Lastly, we define ethnic and racial discrimination as behavior toward or treatment of a group that has positive or negative consequences for the groups involved (Correll et al., 2010). Although authors of manuscripts were often inconsistent in articulating which of these three constructs their measure(s) captured (see further discussion in section Q1), the current review focuses on measures that our study team agreed probed attitudes, stereotypes, and/or discrimination.

In this review, we also considered measures of anti-racism to be an important tool for assessing children's ethnic and racial attitudes, stereotypes, and discrimination. Anti-racism captures efforts to eradicate racism (Aldana et al., 2019); therefore, we conceptualized children's anti-racist judgments and actions to be part of children's affective response toward and treatment of groups. Focus on anti-racism is rapidly intensifying in our field, as developmental scientists have started to recognize the importance of not only understanding and reducing expressions of children's biases but also understanding and promoting children's ability to actively work against interpersonal and systemic discrimination (e.g., Aldana et al., 2019; Cooper et al., 2022; Hazelbaker et al., 2022). Thus, we judged that we could contribute to the emerging scholarship on the development of anti-racism by assessing the use of relevant measures. We assessed measures that were labeled as capturing "anti-racism" as well as those that were not explicitly labeled as such, but captured critical reflection and critical action (i.e., evaluations and actions surrounding discrimination and inequalities), which are argued to be important components of youths' anti-racism (Aldana et al., 2019; Hazelbaker et al., 2022; Heberle et al., 2020; Watts et al., 2011). Note that even though the label "anti-racism" only refers to race, the concept applies to the context of ethnicity as well (e.g., one can work to eradicate anti-Semitism); thus, in our current paper, we use the term anti-racism to include efforts to eradicate ethnicity-based as well as race-based discrimination.

Systematic Review

With the goal of capturing the modern landscape of measure usage, we asked:

Q1. What types of measures of children's ethnic and racial attitudes, stereotypes, and discrimination are being used by researchers?
Q2. How do measures represent target groups?
Q3. In which geographic and demographic contexts are the measures being used?
Q4. What evidence do we have about some of the psychometric properties of commonly used scales/tasks?

We answered these guiding questions by conducting a systematic review of measures of children's ethnic and racial attitudes, stereotypes, and discrimination used in peer-reviewed journal articles published between 2010 and 2022. Our decision to review measures used in articles published during and after 2010 came from our goal of providing useful insights for current practices and future research in the field. We reasoned that the types of measures featured in roughly the last decade are likely to be re-used or adapted in upcoming research. Furthermore, developmental scientists tend to weigh and cite recent studies in their own work, making it especially important to critically evaluate the tools that have been used in recent years.

We are not providing a meta-analysis, but instead a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). Our review may also be characterized as a scoping review which maps out an entire field of research to identify priorities for future research (Levac et al., 2010; Tricco et al., 2018) or an evidence gap map which highlights areas in which we do and do not have evidence regarding a particular topic (Polanin et al., 2023; Snilstveit et al., 2016). By conducting such a review, we hope to spur more careful evaluations of measures and further measure development. We also aim to provide recommendations to the field regarding best practices in measuring and reporting on children's ethnic and racial attitudes, stereotypes, and discrimination.

Method Overview

To identify relevant measures to review in detail, we searched for articles, filtered for articles containing relevant measures, and listed all the measures we found (see Fig. 1 for a diagram of this process).

Article Search

We searched for relevant articles published between 2010 and 2022 by using PsycInfo on January 3rd, 2023. These were the search terms: terms relating to age (*child** or *adolescen**), terms describing attitudes, stereotypes, and discrimination or related constructs (*prejudice**, *bias**, *discriminat**, *attitude**, *stereotyp**, *preference**, *favor**, *affiliat**, *antiracis**, *antibias**, *anti-racis**, or *anti-bias**), and terms describing our social categories of interest (*race*, *ethnic**, or *racial*). The truncation "*" allowed us to search for all forms of the search term roots listed (e.g., child* finds children, childhood, etc.). The search scanned entire articles for these terms, rather than just the abstracts or titles. We only considered articles in peer-reviewed journals. We limited our search to publications available in English

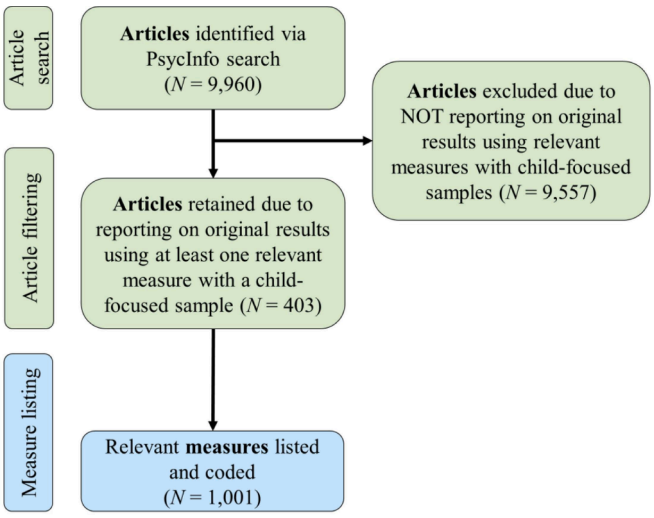

Fig. 1. Flow Diagram Outlining the Process of Listing Relevant Measures. The count of measures ($N = 1{,}001$) represents the number of instances of measure usage rather than the number of unique measures.

because English is the common language of our authorship team. It is also the case that publications in English are likely to be the most representative of work impacting developmental science at an international scale. That said, in our discussion following Q3 below, we consider the implications of limiting our search to English language publications. A total of 9,960 articles were returned based on this search.

Our PsycInfo search did not use terms specifically focused on ethnic and racial identity or acculturation, and we excluded measures that assessed those constructs from our review. Measures focused on *being* a member of a particular group (such as pride or shame) have a different focus than measures probing evaluations and actions *toward* a group. Further, measures of acculturation and identity have been reviewed extensively in other outlets, and there is generally more consensus about how such constructs should be measured as well as more information about the psychometric properties of available measures (e.g., see Byrd, 2012; Casey-Cannon et al., 2011; Cokley, 2007; Helms, 2007; Phinney, 1992; Phinney & Ong, 2007; Satterthwaite-Freiman et al., 2023; Schwartz et al., 2014). That said, there is a conceptual overlap between identity and acculturation measures and the measures we review here, and future research programs that bridge the two currently disconnected areas of literature could help advance our understanding of children's attitudes, beliefs, and behaviors surrounding ethnicity and race (Byrd, 2012; Hazelbaker et al., 2022; Hazelbaker & Mistry, 2022; Karras et al., 2021; Phinney & Ong, 2007; Williams et al., 2023; Umaña-Taylor et al., 2014).

Article Inclusion Criteria and Filtering Protocol

Authors and six research assistants served as initial coders to filter the returned articles ($N = 9{,}960$). In this filtering process, each article was scanned by two coders to determine whether it should be included or excluded. Coders first examined the abstract of articles and scanned the full text if necessary to make a judgment. An article was included if both coders agreed that it reported on original results (as opposed to reviews and meta-analyses) using at least one *relevant measure* with a *child-focused sample* (described below). An article was excluded if both coders agreed that it did not report on original results using *relevant measures* with *child-focused samples*. If the two coders disagreed, a third coder broke the tie or, if necessary, the authors discussed until we came to an agreement (see Supplemental Materials for more details about this process). Because of our broad search criteria, many articles were returned in the PsycInfo search but excluded during our filtering process (e.g., an article about how adults' perceived racial discrimination from childhood related to current substance use). A total of 403 articles were retained after filtering (see OSF for a catalog of retained articles).

We provide key information about inclusion and exclusion decisions below, with more detailed information in Supplemental Materials.

Relevant Measures

Articles were included if they contained at least one relevant measure (i.e., measures within the scope of our review). In line with our definitions of ethnic and racial attitudes, stereotypes, and discrimination described above, relevant measures probed children's own attitudes, valenced ratings, beliefs, stereotypes, inclusive behavior, discriminatory behavior, or behavioral intentions about/toward a person or group with a particular ethnic or racial group identity (including national, religious, linguistic, and caste-based groups). Moreover, in line with our aim to include measures of anti-racism, other relevant measures probed children's evaluation of ethnic and racial discrimination (e.g., expressing disapproval of race-based exclusion) or behaviors that confront discriminatory treatment.

Tasks that involved evaluating specific figures known to children (e.g., friends, classmates, and public figures) were not considered relevant measures due to ambiguity regarding whether such assessments reflect children's responses to the target group or their prior experience with those particular figures. For example, sociometric measures assessing the racial makeup of children's social networks are outside the scope of this review (for reviews on children's social network analyses, see Poulin & Chan, 2010; Wölfer & Hewstone, 2017). Instead, relevant measures presented unfamiliar members (i.e., individual exemplars) of ethnic and racial groups or presented groups themselves (e.g., "Black people") as targets.

Qualitative analysis methods (e.g., thematic analysis) on data from interviews, focus groups, and observations were also not considered relevant measures for our review. However, when data from interview- and observation-based measures were quantified during coding and analyzed using quantitative analysis methods, we considered those measures to be quantitative.

Sample Age

We included articles only when child-focused samples completed the relevant measures. A sample was determined to be "child-focused" if it: (a) had a mean age under 18 years of age, or (b) did not have a mean age reported but most of the age range was under 18 years of age (e.g., participants ranged in age from 12 to 19 years). If a sample was described as a "college," "undergraduate," or "university" sample, it was determined as *not* "child-focused," regardless of mean sample age or age range.

Listing Measures

Once the article filtering process was complete, we created a list of relevant measures from the included articles. Many articles used more than one relevant measure; we found 1,001 measures from the 403 included articles. This number (i.e., $N = 1{,}001$) is the number

of instances of measure usage rather than the number of unique (i.e., distinct) measures. Details on how we determined which study components constitute an individual measure are described in Supplemental Materials.

Measure Coding

Answering our guiding questions (i.e., Q1 through Q4) required us to code for particular characteristics of measure design and usage based on information provided in articles. The specifics of what we coded are outlined in the following sections; however, in general, two authors independently coded each listed measure, and a final code was determined based on the two authors' codes and discussions between the two coders. When there were disagreements between the two coders, the authors discussed as a group to determine the final code. Table S1 shows the summary of the coding scheme. When we could not find the information necessary to code measures in the article text, we examined supplemental materials and online links (e.g., OSF pages) if the article text clearly directed readers to find such information in those materials and links.

In addition to the codes described below, the authors, in pairs, recorded the name and description of each measure as well as the age of participants who completed the measure. To provide the most comprehensive information about the age of participants who completed each measure, we recorded the participant sample's lower and upper bounds in years when that information was provided (72% of all the measures). When articles reported age ranges only in terms of school grades (e.g., "preschool and kindergarteners", "7th to 9th grade") but not age in years, we recorded those grade ranges (21%); converting grades to age in years was not feasible given the wide variation in grade-to-age mappings across school systems and countries. When an age or grade range was not reported in the article text, we recorded the mean and standard deviation of participants' ages (7%).

Below, we lay out our detailed coding scheme, results, and interim discussion relevant to each guiding question.

Q1. What Types of Measures are Being Used?

We found 1,001 measures of children's ethnic and racial attitudes, stereotypes, and discrimination, and we provide a list of their names and descriptions on OSF. Readers can access the list to learn more details about each of the measures we assessed in our coding process. In this section, however, we group measures into like kinds to provide an overview of the different ways researchers have measured children's ethnic and racial attitudes, stereotypes, and discrimination.

Initially, we had planned to sort measures into three categories according to which construct researchers said they tapped: attitudes, stereotypes, or discrimination. These three constructs align with the three commonly proposed types of social bias, with some researchers using terms such as "prejudice, stereotypes, and discrimination" or "affective, cognitive, and behavioral components" (e.g., Correll et al., 2010; Dovidio & Gaertner, 2010; Raabe & Beelmann, 2011); thus, we reasoned that grouping the extracted measures according to the tripartite model (i.e., attitudes, stereotypes, and discrimination) would be simple to execute and helpful to the field. However, attempting to divide the measures into three clear categories proved difficult for several reasons. First, similar types of measures were often labeled as assessing different constructs. For example, some researchers presented tasks probing children's association of different traits with target groups to measure "stereotypes" (e.g., Benatov et al., 2021; Vezzali et al., 2019), while other researchers used these same tasks to measure "attitudes" (e.g., Aronson et al., 2016; Stathi et al., 2014; this inconsistency has been previously noted by other developmental scientists, and we elaborate on it in our discussion following Q1). Second, even when putting aside the labels that were used in articles, attempting to sort measures based on the tripartite model proved impossible because we as the author team often could not come to a consensus on which single construct each measure tapped (a topic we return to in the discussion following Q1).

Finally, we realized that sorting measures according to the tripartite model would not be helpful for describing the diverse types of measures in our field's toolkit; substantial variability exists among measures labeled as tapping the same construct. For example, "behavior" was examined using a variety of measures, including scales probing children's intentions to interact with target groups (e.g., Rastogi & Juvonen, 2019) and tasks asking children to allocate resources to targets (e.g., Rae & Olson, 2018).

The fact that usage of different types of measures did not align with the way psychological researchers often label the different types of social bias seems to reflect the ambiguous links between measures in our toolkit and researchers' conceptual thinking about their constructs of interest. The ambiguity in measure-construct links may lead to inconsistent language when describing or naming measures, and vice versa. These types of ambiguities and inconsistencies have been long problematized in psychological science, as they hinder communication between researchers and preclude convergence of knowledge about measures and phenomena (e.g., for discussions on "jingle and jangle fallacies," see Gonzalez et al., 2021; Kelley, 1927; Thorndike, 1904). In our General Discussion, we suggest future efforts that could help clarify measure-construct relations.

For the reasons stated above, we determined that it would be difficult and unhelpful to sort measures based on the tripartite model (though, to be clear, all members of the study team agreed that all identified measures focused on attitudes, stereotyping, and/or discrimination). Instead, we categorized measures according to the kinds of things that children were asked to think about and do. These measure categories were not groupings based on statistical relations between measures or a "ground truth," so others could reasonably conceive of other ways to carve up the list of measures. Nonetheless, we created parsimonious categories that we judged to be meaningfully cohesive to summarize the large number and vast variety of measures. In the General Discussion, we expand on our invitation for developmental scientists to use our categories as a starting point for more careful conceptual evaluation and discussions about the existing measures as a field.

Q1 Method

To group measures into categories, we first scanned across the listed measures to observe the kinds of things children were asked to think about and do. Then, through discussion, we created parsimonious groupings of measures to serve as our measure categories. Finally, we classified each measure into a category.

Q1 Results

Below, we describe the eight categories and the measures that were sorted into them. Specific examples of measures in each category can be found in Table 1. As described in the Method Overview, the numbers of measures throughout our review represent the instances of measure usage rather than the number of unique measure types (see Q1 Discussion for challenges associated with counting unique measure types). For example, if there were 12 instances of measures with identical names and designs, we counted them as 12 measures. As such, the number of measures reported for each category is the count of instances in which measures that fall under the category were used.

Trait Attribution

We sorted 311 measures (31% of all extracted measures) into the *Trait Attribution* category. These measures asked children to attribute traits, behaviors, and intentions to targets. The traits, behaviors, and intentions were general positive/negative attributes in some cases (e.g., nice, friendly, naughty, mean) and more culturally specific stereotypes associated with particular ethnic/racial groups in other cases (e.g., good at basketball). Additional measures probed children's causal explanations for group-based differences in cultural stereotypes. Most measures took the form of scales (e.g., Black/White evaluative trait scale [BETS]; Hughes & Bigler, 2007) or forced-choice tasks (e.g., Preschool Racial Attitude Measure [PRAM II]; Williams et al., 1975).

Note that we did not draw a distinction between measures that asked children about general positive/negative attributes vs. those that asked about culturally specific stereotypes, because the two types of attributes were often conflated (e.g., "aggressive" was used as a general negative attribute in some measures but it is also often culturally associated with specific groups such as Black people; Ghavami & Peplau, 2012), making it difficult to differentiate valenced attributions from cultural stereotypes. This conflation reflects ambiguities in what *Trait Attribution* measures precisely capture – an issue we return to in the discussion following Q1.

Social Interactions

A total of 176 measures (18%) were sorted into the *Social Interactions* category. These measures asked children to interact with (or imagine interacting with) individuals belonging to particular ethnic or racial groups, and probed children's feelings, intentions, and motivations surrounding those interactions. Many measures took the form of scales (e.g., Bogardus Social Distance Scale; Bogardus, 1933), while others were forced-choice tasks.

Liking and Feelings

A total of 122 measures (12%) were sorted into the *Liking and Feelings* category. These measures asked children to express their general attitudes and affect (i.e., liking and feelings of warmth) as well as specific emotions (e.g., admiration, resentment) toward ethnic or racial groups. Many measures took the form of scales (e.g., feelings thermometer), while others were forced-choice tasks. Note that we sorted measures into the *Liking and Feelings* category when they probed children's feelings in the absence of actual or imagined contact with targets, whereas we sorted measures into the *Social Interactions* category when children's feelings were probed in the context of imagined or actual contact with targets.

Anti-Racism

A total of 114 measures (11%) were placed in the *Anti-Racism* category. These measures presented children with other people's discriminatory actions or scenarios of inequality, and probed children's evaluation and/or confrontation of those wrongdoings. Evaluation measures assessed children's feelings, emotions, or judgments about others' antisocial or prosocial behaviors in an ethnic or racial context. Confrontation measures assessed whether children took action, expressed intentions to take action, reported that they had taken action, or said that it would be right to take action in response to interpersonal or systemic discrimination.

Societal Functioning

We sorted 69 measures (7%) into the *Societal Functioning* category. These measures asked children to respond to scale items tapping into their beliefs about how a society functions or should function with regard to people of different ethnicities and races. Measures assessed participants' endorsement or rejection of egalitarian policies and practices, prescriptive beliefs such as how target group members should behave, and beliefs about how a target group contributes positively or negatively to society.

Implicit

Sixty-eight measures (7%) were sorted into the *Implicit* category. These measures included speeded reaction time tests (e.g., Implicit Association Task and its adaptations), priming tasks (e.g., Affect Misattribution Measure), and physiological measures, which are commonly thought to capture unconscious or uncontrollable prejudice in the social psychology literature (Axt et al., 2014; Gawronski, 2019; Gawronski et al., 2006; Lai & Wilson, 2020).

Table 1
Examples of Measures in Each Category.

Measure category	Number of measures	Example
Trait Attributions	311	Mesman et al. (2022): Children completed a version of The Multiple Response Racial Attitude measure (Aboud, 2003). This is an attribution task in which children were asked to assign five positive descriptors (e.g., nice, friendly) and five negative descriptors (e.g., naughty, mean) to any number of six children on photographs. This particular measure used just the five positive descriptors. Children were presented with photographs of one boy and one girl from the White, Black, and Middle Eastern groups. A total positive outgroup attribution score consisted of the total number of positive descriptors to the Black and Middle Eastern children separately and combined (potential range 0–10). Berger et al. (2018): Participants were asked to report their expectations about how likely the outgroup was to engage in 5 negative behaviors (e.g., ignore you, insult you, hurt, steal). Ratings were made on a 5-point scale ranging from "no chance at all" (1), to "high chance" (5).
Social Interactions	176	Schuitema & Veugelers (2011): To measure intergroup anxiety (see Stephan and Stephan 1985; Turner et al., 2007), we asked students to imagine that they were in a class only with students from the other ethnic group. The question for the Apollo students was: "How would you feel if you were the only non-Dutch student in a class of only Dutch students?" The students responded on five five-point semantic differential items: happy–unhappy, pleased–worried, secure–insecure, anxious–safe and comfortable– tense. Tropp et al. (2014), Study 1 & Study 2: Participants were asked, "How much would you like to become friends with kids who are [White/Black]?" Participants responded on a scale ranging from 1 (Not at All) to 5 (Very Much). Carraro & Castelli (2015): Participants were shown drawings of a Black/Asian vs. a White child and asked which one they'd like as a playmate.
Liking and Feelings	122	Liebkind et al. (2014): Out-group attitudes were assessed by asking the respondents to indicate their overall feeling towards the target out-group (people with immigrant background/Finns) along a feeling thermometer (0 = extremely unfavourable, 100 = extremely favourable). Constantin & Cuadrado (2021), Study 1: Participants are asked to assess a group on eight items corresponding to positive emotions (items adapted from the work of Fiske et al. (2002) by Cuadrado et al. (2016)). The eight items are admiration, understanding, respect, comfort, fondness, pride, inspiration, and security. Responses to the eight items are on a 5-point Likert scale ranging from 1 (not at all) to 5 (very much).
Anti-Racism	114	Brenick et al. (2010): Children were read three vignettes focused on issues of exclusion (based on cultural membership), inclusion (regarding societal conventions and language barriers), and stereotypes. For each vignette, children were first asked to make a judgment (e.g., "How good or bad is it to exclude the child?"). They were then asked to explain their judgment, referred to as justification (e.g., "Why is it OK or not OK to exclude?"). Aldana et al. (2019), Study 1: Four items are as follows: "Attended a meeting on an issue related to race, ethnicity, discrimination, and/or segregation"; "Joined a club or group working on issues related to race, ethnicity, discrimination, and/or segregation"; "Tried to get into a leadership role or committee (i.e. student council, etc)"; "Participated in a leadership group or committee working on issues related to race, ethnicity, discrimination, and/or segregation (i.e. youth organizing group), etc." Each item was answered "Yes" or "No." Scores were summed so that higher scores indicated more action.
Societal Functioning	69	Reijerse et al. (2015): Participants were asked to what extent they agree with 4 statements about migration issues: (1) "Our government should make more of an effort at integrating immigrants into our society", (2) "I think immigrants should be offered a naturalization programme, but only on a voluntary basis", (3) "I think our government should actively protect immigrants against discrimination", (4) "Our government should start up programmes, specifically for immigrants, which help them to increase their chances of getting a job." Participants rated on a 1 (strongly disagree) to 7 (strongly agree) scale. A single score was calculated from the 4 items. Miklikowska (2017): Participants reported on their attitudes towards immigrants by rating three items: "Immigrants often come here just to take advantage of welfare in Sweden," "It happens too often that immigrants have customs and traditions that not fit into Swedish society," and "Immigrants often take jobs from people who are born in Sweden." All items were rated on a 4-point Likert scale (1 = don't agree at all to 4 = agree completely) and their mean was used to construct the scale score.
Implicit	68	Gonzalez et al. (2017): This test measures the strength of an association between race (Black/White) and affect (good/bad). The categories Black and White were each represented with four pictures of children. The "good" and "bad" attributes were presented acoustically. Specifically, children heard four words that could be categorized as "good" (happy, fun, good, nice) or "bad" (yucky, sad, mad, mean). Participants were instructed that any time they saw an image in the middle of the screen or heard a word to determine with which category it belonged (White or Black; good or bad) and to press the corresponding button. A D score was calculated, which is a variation of Cohen's d, and represents the magnitude of a participant's implicit preference for one group relative to a comparison group. Williams & Steele (2019), Study 2: Images of eight White and eight Black boys, as well as eight gray squares, were used as the race and neutral primes, respectively. Target images were inkblots that were pretested to be neutral in

(*continued on next page*)

Table 1 (*continued*)

Measure category	Number of measures	Example
		valence. Children were told that they would briefly see inkblots and their task was to indicate whether the inkblot was "nice" or "not so nice" by pressing one of two computer keys. The proportion of inkblots judged as pleasant following each type of prime was calculated separately, resulting in distinct White, neutral, and Black priming indices.
Prosociality	42	Renno & Shutts (2015), Study 1 & Study 2: Participants gave out coins to either/both/neither of the photographed White or Black child across 6 trials. A "giving score" was calculated by subtracting coins given to Black children from coins given to White children. Taylor & McKeown (2021): Participants were asked rate the extent to which they would be willing to help a Syrian refugee by responding to 6 statements on a 7-point Likert scale from 0 (absolutely not) to 6 (absolutely yes). Helping intention statements included, "Help Mohammad/Fatima with English?" Higher scores indicate higher helping intentions.
Modeling and Testimony	13	Chen et al. (2018): A Hong Kong Chinese experimenter presented a novel object to the participant and said in Cantonese, "In this video, two people will show you how to use this toy. Do you know what this toy does? I don't know what this toy does, so let's see what they think." Children watched the informants silently perform different functions for the object. Informants differed by racial group. Next, the experimenter paused the video, repeated the actions, and asked, "Can you show me how you would use this toy?" This occurred for eight trials. Scores on the task represented the number of trials where the participant endorsed the Chinese informant's use of the toy. Kinzler & Spelke (2011), Experiment 1: In each of four test trials, infants saw a toy choice event, with both White and Black individuals pictured simultaneously offering a toy to the infant. Just at the moment when the toys disappeared off screen, two real toys "magically" appeared from behind the table for the infant to grasp, giving the illusion that they emerged from the screen. This occurred for 4 trials. The number of choices selected for the White individual and the number of choices selected for the Black individual were each counted.

Prosociality

Forty-two measures (4%) were sorted into the *Prosociality* category. In these measures, children themselves were given the opportunity to act, or express their intentions to act, prosocially or antisocially in an ethnic or racial context. For example, children could allocate/remove resources or help someone in need. Note that measures in which children made resource allocations in response to a depicted inequality between groups (i.e., as a rectification response) were sorted in the *Anti-Racism* category.

Modeling and Testimony

Thirteen measures (1%) were placed in the *Modeling and Testimony* category. These measures used a distinctive type of paradigm in which a model (who varied in race or ethnicity) first provided information or modeled a behavior, and then children's subsequent behaviors were assessed. Children were assessed on their preference of objects or labels endorsed by the model, or their actions based on the model's behavior or claims.

Uncategorized

A total of 86 measures (9%) could not be sorted into a category. One measure did not fit the task type of any of the eight categories, and the rest did not cleanly fit into a single category. Measures did not cleanly fit into a single category when they used a mix of items that could each be sorted into a different category. For example, Gómez et al. (2013) used a scale in which participants had to rate their agreement to items such as "I would cooperate with immigrants from these countries to solve problems that affect all Europeans" and "We should provide social programs that help these immigrants face the problems of our society;" the former item fits into the *Social Interactions* category, whereas the latter item fits into the *Societal Functioning* category.

Q1 Discussion

Taken together, we found that many types of measures have been used to assess children's ethnic and racial attitudes, stereotypes, and discrimination. We identified eight categories, which served as a useful way to organize and summarize the measures that exist in our field's toolkit.

Diversity of Measures

Although we were able to categorize most measures into one of the eight categories we created, it is important to recognize that there is diversity within each category. For example, some measures in the *Social Interactions* category assessed the extent to which children wanted to become friends with out-group peers (e.g., on a scale ranging from 1 = strongly disagree to 7 = strongly agree; Tropp et al., 2016) while others assessed the frequency with which children indicated they would select out-group vs. in-group children as friends (e.g., Dore, 2022). Readers who are interested in gaining a more granular sense of the measure usage captured within each of our eight categories can refer to the "clusters" in our list of measures (OSF; measure "clusters" are described below in Q4 Method).

It is also worth noting that even when considering a particular named measure (e.g., Tolerance and Xenophobia Scale in the *Societal Functioning* category; van Zalk et al., 2013; Bayram Özdemir et al., 2016, 2021), we observed variation across different uses of the

measure (e.g., different numbers of items; changes to the wording of items). Unfortunately, these variations were not always clearly detailed in the text, and this lack of clarity resulted in our strategy of reporting on instances of measure usage rather than counts of unique measures in the field. In the General Discussion, we return to problems related to (lack of) clarity about when one is and is not using the same measures, and offer suggestions to the field.

Having a large variety of measures means that researchers who are interested in studying children's ethnic and racial attitudes, stereotypes, and discrimination have a diverse toolkit at their disposal. However, the field cannot benefit from this diversity if researchers do not understand how different measures are conceptually and statistically overlapping with or distinct from each other. In other words, researchers need clarity on whether different types of measures can be used for different or similar purposes, and whether results from different measures are comparable to each other (e.g., can the *Social Interactions* measures from Tropp et al., 2016; Dore, 2022 be used interchangeably in studies?). In our General Discussion, we elaborate on this issue and provide suggestions for future efforts to understand the measure-to-measure relations in our toolkit.

Anti-Racism Measures

We were surprised to find that many measures fell into our *Anti-Racism* category, given that our field's heightened focus on anti-racism has only recently emerged (Aldana et al., 2019; Cooper et al., 2022; Hazelbaker et al., 2022). Only in 19 instances were measures explicitly labeled by authors as assessing "anti-racism" per se (e.g., Anti-Racism Action Scale; Aldana et al., 2019), and these measures were used only with children who were 13 years and older. However, because we defined our *Anti-Racism* category to include measures of children's evaluation and confrontation of discriminatory actions or scenarios of inequality broadly (capturing critical reflection and critical action), many measures used with younger children were categorized as *Anti-Racism*. Specifically, 61 measures asked children (as young as 3 years) to evaluate the wrongfulness or permissibility of acts of interpersonal exclusion ($n = 30$; e.g., Ruck et al., 2011), unequal structural resource distribution ($n = 6$; e.g., Elenbaas & Killen, 2017), or other mean interpersonal acts such as making racist jokes ($n = 25$; e.g., Niwa et al., 2016). Five other measures assessed whether 3- to 11-year-old children would take action to rectify inequalities and punish/reward transgressors in hypothetical interpersonal ($n = 3$; e.g., Olson et al., 2011) and structural scenarios ($n = 2$; e.g., Elenbaas et al., 2016).

Our findings suggest that the use of measures labeled as "anti-racism" may still be sparse and limited to older children; yet when considering measures that tap children's evaluations and actions surrounding discrimination and inequalities more broadly, we see many uses of measures for younger children. Most of these measures for younger children do not assess their actions that directly confront structural bias and racism, but they do assess children's abilities to identify and negatively evaluate interpersonal discrimination, which developmental scientists have proposed to be a foundational step toward recognizing and confronting racism at the structural level (Hazelbaker et al., 2022). Creating more measures that tap young children's abilities to actively confront discrimination both at the interpersonal and structural levels may help elucidate developmental trajectories of anti-racism. Taken together, our finding suggests that there are existing measures in our field's toolkit which we can use and build upon to gain a deeper understanding of how children's anti-racist attitudes and behaviors develop starting in early childhood and how we can promote them.

Relations Between Measures and Constructs

As described above, even though we categorized measures according to the kinds of things that children were asked to think about and do (instead of the constructs that measures tapped), we still considered each measure to be operationalizations of one or more of the three constructs in the tripartite model (i.e., attitudes, stereotypes, and/or discrimination). Given the importance of thinking carefully about the conceptual connections between measures and constructs for effective communication and convergence of knowledge, we lay out here proposed conceptual relations between existing measures and constructs.

When considering how exactly measures under our review may relate to attitudes, stereotypes, and discrimination, some measures seemed to clearly tap a single construct. For example, we as an author team came to a consensus that measures in the *Liking and Feelings* category probe children's affective responses to target groups, which we define as attitudes. However, we saw overlaps and ambiguity in most other relations between measures and constructs. First, measures in our *Social Interactions* and *Societal Functioning* categories may tap attitudes, but they also seem to reflect children's beliefs about target groups (i.e., stereotypes; e.g., immigrants take away jobs) and behavioral responses to target groups (i.e., discrimination; e.g., choosing to sit next to a racial out-group peer).

Second, it is debatable whether children's responses to *Trait Attribution* and *Implicit* measures capture their knowledge or awareness of cultural stereotypes (i.e., widely held beliefs about particular groups in a society), their endorsement of (i.e., agreement with) those cultural stereotypes, or their more general evaluations of groups (Sierksma et al., 2022; see Signorella et al., 1993 for a discussion on the distinction between children's knowledge and attitudes). In a similar vein, social psychologists have long debated whether "implicit" measures capture cultural associations that people have been exposed to or their endorsement of those evaluative associations (i.e., attitudes; Dovidio et al., 2010; Dovidio & Gaertner, 2010; Karpinski & Hilton, 2001; Lai & Wilson, 2020). Thus, measures in our *Trait Attribution* and *Implicit* categories could tap attitudes, stereotypes, or both.

Lastly, measures in our *Prosociality*, *Modeling and Testimony*, and *Anti-Racism* categories could be probing children's behavioral responses to ethnic/racial groups (i.e., discrimination). Here again, however, many measures seemed to tap multiple constructs. For example, measures assessing young children's desire to learn something from a member of their in-group versus their out-group (in *Modeling and Testimony*) could be tapping attitudes (e.g., affective favoritism toward the in-group), stereotypes (e.g., association of competence to the in-group), and/or discrimination (e.g., treating in-group members with higher respect than out-group members).

Overall, there are many conceptual ambiguities in our existing measures toolkit, but the field lacks widespread discussions about the conceptual grounding of measures for capturing children's attitudes, stereotypes, and discrimination. In the General Discussion, we further discuss the issue of measure-construct relations and offer our recommendations for future efforts in the field.

Q2. How do Measures Represent Target Groups?

Characteristics of stimuli in measures can impact children's responses to laboratory measures and the conclusions that we draw from them (see Stengelin et al., 2023 for a discussion). For example, presenting drawings versus photographs of individuals influences how strongly children encode ethnic and racial categories during tasks (Guerrero et al., 2010), and the salience of ethnicity/race in tasks impacts children's activation of implicit attitudes (Williams & Steele, 2019). Thus, the stimuli that are used to represent the targets of evaluation (i.e., ethnic or racial groups and members of those groups) in measures are key aspects of measure designs that should be carefully evaluated.

In this section, we report on the types of stimuli that were used to represent target groups in measures of children's ethnic and racial attitudes, stereotypes, and discrimination. Specifically, we coded whether measures presented verbal descriptors, visual exemplars, live targets, and/or other types of cues to convey group information in measures.

We had also initially planned to code for detailed characteristics of the visual stimuli, including the format (i.e., drawings, photographs, or videos), the gender and age of the targets depicted in stimuli (i.e., whether the gender was matched to the participant and whether they were adults or children), and how those stimuli were created and normed. However, these details were not reliably reported, so we could not review them systematically. In our General Discussion, we offer some ideas about the implications of using stimuli with different characteristics and provide recommendations about reporting practices.

Q2 Method

We first coded whether measures used verbal descriptors, visual exemplars, live targets, or other cues to represent target groups. Verbal descriptors included adjectives and nouns referring to specific groups (e.g., Black, Dutch, Mexicans), as well as more general verbal descriptions of ethnic/racial group distinctions (e.g., "ethnic group," "skin color," "from a different country"). Visual exemplars were photographs, drawings, pre-recorded videos, or dolls representing individuals belonging to ethnic/racial groups. Note that if a measure employed verbal instructions to accompany visual exemplars (e.g., "look at this person") but did not reference ethnic/racial groups verbally, we coded the measure as using visual exemplars but not verbal descriptors. Live targets were unfamiliar children or adults (i.e., confederates or models) who appeared to participants in real time via live video call or in the same physical space. Other representations included: accents in speech, names of individuals that are stereotypical of the target group, and national flags.

In the coding process, we noted whether each of the four types of representation occurred for each measure, attending both to measure instructions that children heard (when described in the article) as well as information provided about/within the measure itself. Thus, each measure could receive one or more codes for representation type. Additionally, to examine whether use of the four representation types differed by the age of participants who completed the measures, we calculated each measure's sample age "midpoint" in years. For measures whose sample age ranges (in years) were reported, we defined the midpoint as the median of the age range and for measures whose age ranges were not reported, we defined the midpoint as the sample mean (see the Measure Coding section above for details on our age coding). Because we could not calculate the midpoint in years for sample age ranges that were reported in terms of school grades (given the variability in the age of children in particular grades across and within countries), measures using those samples (21%) were not included in our representation type-by-age review.

Measures sorted into the *Anti-Racism* category ($n = 114$) were not coded on these criteria because *Anti-Racism* measures typically asked children to evaluate or respond to discriminatory incidents rather than assess children's attitudes, stereotypes, or discriminatory behavior toward particular target groups.

Q2 Results

Verbal descriptors were used in 575 measures (65%), visual exemplars were used in 394 measures (44%), live targets were used in 21 measures (2%), and other cues were used in 74 measures (8%). Thus, both the use of verbal descriptors and visual exemplars to represent target groups were prevalent.

Broken down another way, 469 measures (53%) used only verbal descriptors, 251 measures (28%) used only visual exemplars, 7 measures (1%) used only live targets, 7 measures (1%) used only other representations, and 153 measures (17%) used a combination of those representation types. Among measures that used a combination of representation types, most used a combination of verbal descriptors and visual exemplars ($n = 100$).

When looking at the sample age midpoint for measures using each representation type, we found that the average midpoint for measures using verbal descriptors was 12.89 years, while the average midpoint for measures using visual exemplars was 7.73 years. Furthermore, the average midpoints for measures using live targets and other cues were 4.55 years and 9.45 years, respectively. These findings suggest that measures with verbal descriptors were frequently used with older samples, on average, than measures with visual exemplars. It is also evident that live targets tended to be used with very young children.

Q2 Discussion

Our coding results reveal that target groups are most frequently represented in measures by using verbal descriptors, visual exemplars, or a combination of these two approaches. Measures of children's ethnic and racial attitudes, stereotypes, and discrimination rarely present children with live people (see Table S2 for a list of all measures that used live people as targets). Our results also suggest that the ways in which targets were represented differed based on children's age, such that verbal descriptors were used more in

measures for older children while visual exemplars and live targets were used more in measures for younger children. This age-based pattern is not surprising considering that older children have more advanced verbal skills. However, researchers should more critically evaluate the implications of using different types of stimuli in measures and systematically examine how children of different ages interpret different representations of ethnic/racial groups (as we further discuss in our General Discussion).

The finding that live targets are rarely used in measures of children's ethnic and racial attitudes, stereotypes, and discrimination, and that they are used almost exclusively with children under school age (see Table S2), raises an intriguing point: Even though many researchers presumably aim to better understand children's consideration and treatment of real-life people in real-time settings, laboratory measures employ live people as targets of evaluation in very limited contexts. The scarcity of measures with live targets is likely due to practical and ethical concerns. Employing real-life people with specific ethnic/racial identities to serve as targets (e.g., confederates) for each participant requires time and resources and may lead to less controlled methodologies. Moreover, involving real-life people requires significant ethical considerations, especially if confederates would be subjected to discrimination.

Given that practical and ethical considerations may limit developmental scientists' ability to use live targets, it is important for the field to understand whether existing laboratory measures that do not use live people as targets (e.g., measures that use hypothetical or photographed people) predict children's real-world behaviors toward people with the target ethnic and racial identities. If the field has established that such laboratory measures do indeed predict children's real-world ethnic and racial behaviors, then we may not need concern ourselves with employing live targets in measures. As we reveal in Q4 Results, however, we currently have little evidence to suggest the predictive validity of measures under our review. Our General Discussion offers recommendations regarding how the field can establish a better understanding of the existing measure toolkit, such as knowing how to use laboratory measures to appropriately make inferences about children's consideration and treatment of real-life people in real-time settings.

Q3. In Which Geographic and Demographic Contexts are the Measures Being Used?

Q1 and Q2 focused on exploring the types of tools that have been used in our field to measure children's ethnic and racial attitudes, stereotypes, and discrimination. To enrich our understanding of measure usage, we now turn to reviewing the contexts in which those measures have been deployed. Specifically, we examined the geographic locations where the measures were used, the participant sample's ethnic/racial makeup, and the ethnic/racial groups that were presented in measures. Reviewing these patterns of measure usage is key to understanding how much information we have about the way measures perform in different contexts. Identifying geographic or demographic biases and gaps is important for guiding future efforts to better understand the properties and generalizability of measures in our field's toolkit. Additionally, assessing contexts of measure use can shed light on locations and populations where children's ethnic and racial attitudes, stereotypes, and discrimination have been well-studied vs. understudied.

Q3 Method

We first coded for the countries (or regions recognized by the United Nations or the U.S. Department of Homeland Security; Office of Homeland Security Statistics, 2023; United Nations, n.d.) of the samples based on descriptions in articles. Additionally, to summarize patterns of measure usage on a larger geographic scale, we coded for the continent in which each country/region is located. Continents were determined based on the list of countries/regions for each continent provided by the United Nations (United Nations, n.d.).

Next, we recorded the ethnic/racial groups presented in measures and the ethnicities/races represented in the participant sample. To record the ethnic/racial groups that were presented (i.e., groups that children were asked about), we noted the groups that were targets of evaluation (i.e., "target groups") for all but *Anti-Racism* measures. As mentioned in Q2 Method, *Anti-Racism* measures typically asked children to evaluate or respond to discriminatory incidents rather than particular target groups; therefore, we instead recorded the groups involved in *Anti-Racism* measures (e.g., "involved groups" would be "Black" and "White" for *Anti-Racism* measures where children were asked to evaluate a discrimination incident involving a White perpetrator and a Black victim).

To record the ethnic/racial groups represented in participant samples (i.e., "participant groups"), we searched each article for participant demographic information that was relevant to the ethnic/racial groups that were presented in measures. That is, although participants could be described along multiple racial/ethnic dimensions (e.g., someone could be White, Jewish, American, and an English speaker; see the Scope of Review section for our definition of race and ethnicity), we recorded the participant groups that were most relevant to the dimension of interest in the measure (i.e., dimension of the target/involved groups). For example, for a measure assessing U.S. children's attitudes toward Muslim people, we recorded information about the participants' religious identities (rather than recording whether participants were Asian, Black, White, etc.).

In our coding process, because we did not wish to impose our interpretation of the reported racial/ethnic groups and because of our limited knowledge about the specific ethnic/racial categorizations in every demographic context, we took the names of ethnicities/races directly from the language that articles used to describe their samples and measures. That is, articles used varying category boundaries and language to describe their groups (e.g., "non-immigrants," "British," "Jews," "European American"), and our coding of race/ethnicity names stayed true to those reports rather than forcing those groups into categories that we created (e.g., "White").

Readers can refer to our list of measures (OSF) to learn the ethnicities/races that were represented in samples and measures across all the reviewed measures. That said, it is impossible to summarize across all measures used in all the countries/regions that contributed measures without glossing over important differences in ethnic/racial dynamics across geographic locations. Therefore, in the current paper, we provide synthesis that focuses on a few select countries/regions to help readers learn about some of the measure usage patterns in specific ethnic/racial contexts across a range of geographic locations. To provide such a synthesis, we conducted

additional coding for the country in each continent that contributed the most measures to the current review. Specifically, we recorded the number of participants belonging to each race/ethnicity and coded the type of sample racial/ethnic makeup (e.g., all majority children, more minority children than majority children; see Table 2).

Q3 Results

Measures were used across 53 countries/regions (for a full list of countries/regions, see Table S3) across six continents. Most measures were used in North America ($n = 451$; 45%) and Europe ($n = 366$; 36%). A relatively small proportion of the measures were used in Asia ($n = 139$; 14%), South America ($n = 21$; 2%), Africa ($n = 19$; 2%), and Oceania ($n = 9$; 1%). See Fig. 2 for a graph showing the percentage of measures by continent.

The samples spanned similar age ranges across continents, and different types of measures (i.e., the eight measure categories described in Q1) did not map clearly onto different continents (see Table S4). In other words, the wide variation in measure types cannot be attributed to different types of measures being created and used in different locations.

In each continent, the country(ies) that contributed the most measures was (were) the U.S. ($n = 401$; 89% of measures in North America), the Netherlands ($n = 87$; 24% of measures in Europe), China ($n = 41$; 29% of measures in Asia), Israel ($n = 41$; 29% of measures in Asia, tied in number of measures with China), Brazil ($n = 16$; 76% of measures in South America), South Africa ($n = 12$; 63% of measures in Africa), and Australia ($n = 5$; 56% of measures in Oceania).

Below, we summarize the ethnicities/races of participants and target/involved groups in measures used in each continent's top country. Altogether, these summaries cover 602 measures (60% of all measures in the current review). As readers will see in the results and Table 2, we only provide participant group summaries for measures that focused on the most common racial/ethnic dimension in each country; this is because the participant groups were often too variable to summarize across all measures within a country/region (see the Q3 Method above for details on how we matched participant group codes to target/involved group dimension). Information on the participant groups represented in the rest of the measures can be found in our list of measures (OSF).

Table 2
Number of Measures by Target/Involved Groups and Participant Sample Makeup in Top Countries.

Country	Target/Involved Groups	Participant Sample Makeup					
		All majority children	**More majority children**	**Equal majority/ minority**	**More minority children**	**All minority children**	**Cannot determine from text**
United States ($N = 401$)	Groups that align with or fall under ethnic/racial groups commonly recognized by U.S. institutions ($n = 357$)	78 (22%)	96 (27%)	7 (2%)	84 (24%)	69 (19%)	23 (6%)
	Other groups ($n = 44$)	See list of measures for specific participant groups					
Netherlands ($N = 87$)	Groups with and without modern history of migration to the Netherlands ($n = 83$)	50 (60%)	27 (33%)	0	2 (2%)	1 (1%)	3 (4%)
	Other groups ($n = 4$)	See list of measures for specific participant groups					
China ($N = 41$)	Asian/Chinese, Black, White ($n = 36$)	35 (97%)	1 (3%)	0	0	0	0
	Other groups ($n = 5$)	See list of measures for specific participant groups					
Israel ($N = 41$)	Jewish, Arab ($n = 35$)	8 (23%)	13 (37%)	0	8 (23%)	6 (17%)	0
	Other groups ($n = 6$)	See list of measures for specific participant groups					
Brazil ($N = 16$)	Black, Pardo, White ($n = 16$)	0	6 (38%)	0	4 (25%)	6 (38%)	0
South Africa ($N = 12$)	Black, Coloured, White ($n = 12$)	5 (42%)	2 (17%)	0	5 (42%)	0	0
Australia ($N = 5$)	Various groups	See list of measures for specific participant groups					

Note. Numbers indicate count of measures, not count of participants. Percentages indicate the proportion of measures in each participant sample makeup bin for each row. List of measures can be accessed on OSF.

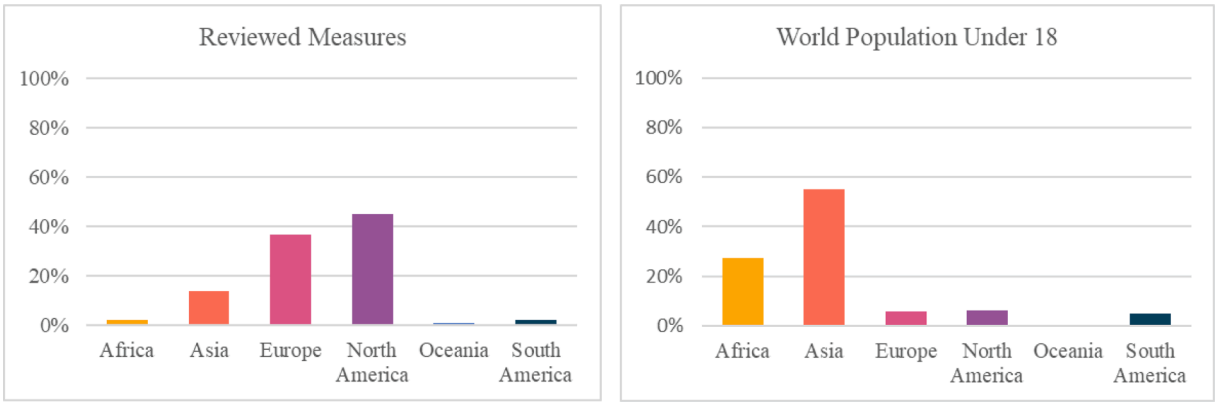

Fig. 2. Graphs of the Percentages of Reviewed Measures and 2021 World Population by Continent.

United States

Most measures ($n = 357$ of 401) in the U.S. asked children about groups that align with or fall under ethnic/racial groups commonly recognized by U.S. institutions (e.g., the U.S. Census offers these groups as racial/ethnic identification options: White, Black or African American, American Indian and Alaska Native, Asian, Native Hawaiian and other Pacific Islander, and Hispanic or Latino). Measures invoked those groups by labeling or depicting them directly (e.g., most commonly Asian, Black, and White people), by asking children about more specific ethnic groups that fall under those broader groups (e.g., "East Asian," "Mexican"), by asking children about people with different skin colors (e.g., "people with a different skin color than me"), or by referring to ethnicity or race itself (e.g., "people of a different racial/ethnic background," "is it alright or not alright to exclude based on race?"). Out of the 357 measures, 258 measures (72%) presented Black people, 241 measures (68%) presented White people, 81 measures (23%) presented Asian people, 55 measures presented Hispanic or Latino people (15%), 5 measures presented Native American people (1%), and 2 measures presented multiracial people (0.5%). Across the 357 measures in the U.S., children were rarely asked about only a single ethnic/racial group ($n = 36$, 10%); most measures asked children about two or more groups ($n = 321$, 90%), with the combination of Black and White people being the most common ($n = 167$, 47%). Furthermore, 7 measures (2%) asked children about people with different skin colors and 65 measures (18%) referred to ethnicity or race itself. Overall, these patterns of target/involved groups highlight that U.S children's evaluations of Black and White people are heavily studied, whereas their evaluations of other ethnic/racial groups, particularly Native American and multiracial people, are severely understudied in comparison (for discussions, see Fryberg & Eason, 2017; Gaither, 2015; Nishina & Witkow, 2020).

According to the United States Census Bureau (2020), 75% of people living in the U.S. identify as White, 14% identify as Black or African American, 6% identify as Asian, 3% identify as multiracial, 1% identify as American Indian and Alaska Native, and 0.3% identify as Native Hawaiian and other Pacific Islander (and 20% identify as Hispanic or Latino in addition to one of these racial groups). Therefore, we considered White children (as well as children labeled as "European American" and "Caucasian") as the majority children and children of all other races/ethnicities as minority children in our summary of study samples. Of the 357 measures, over half of them were used with samples that consisted of a mix of majority and minority children, and approximately equal numbers of measures were used with entirely majority and entirely minority samples (see Table 2 for exact numbers of measures by sample type).

Forty-four other measures in the U.S. asked children about ethnic/racial groups defined by religion (e.g., "Jewish"), immigration status (e.g., "immigrants," "people who move to America from other countries"), nationality (e.g., "American"), or accent in speech (e.g., "French accented"). Our list of measures (OSF) contains information about the participants who completed these measures.

The Netherlands

Most measures ($n = 83$ out of 87) used in the Netherlands asked children about groups with no modern history of migration ("Dutch," "Dutchmen," "Dutch majority," "Dutch natives") or groups with a more recent history of migration to the Netherlands. The latter were diverse, but most commonly included people/groups with origins in Suriname or the Middle East (most often, Morocco and Turkey). Although authors often labeled their stimuli (in their articles and/or for their participants) with specific country of origin names (e.g., "Antillean," "Moroccans," "Surinamese Dutch"), some measures referred to groups more generally (e.g., "majority," "immigrants," "refugees," "from another country," "ethnic minorities") or used racial and regional names (i.e., Black, Middle Eastern, White). Out of the 83 measures, many measures ($n = 57$, 69%) asked children about both groups with and without recent history of migration, whereas fewer measures asked children about just groups with modern history of migration ($n = 22$, 27%) or just those without modern history of migration ($n = 4$, 5%).

According to Statistics Netherlands (2022), 75% of people in the Netherlands have a Dutch background while 25% have a migration background (11% European but not Dutch, 6% Asian, 4% African, 4% American, and 0.1% Oceanian background). Therefore, we considered Dutch children to be majority children and considered children of all other population groups as minority children in our summary of study samples. As is evident from Table 2, almost all measures were used with samples that consisted entirely of or mostly

of majority children.

Beyond the already discussed measures, four additional measures asked children about religious groups (i.e., "Muslims" or the "Muslim minority"). Our list of measures (OSF) contains information about the participants in these measures.

China

Most measures ($n = 36$ out of 41) used in China asked children about Asian, Black, and White people. In many of these measures, additional specificity was given beyond these labels; specifically, the label "Chinese" was used to describe Asian targets. Out of the 36 measures, 34 measures presented Black people (94%), 30 measures presented Asian people (83%), and 14 measures presented White people (39%), indicating that measures almost always asked about Black people (alone or in addition to Asian and White people). Note that, in comparison to measures in the other top countries, which probed ethnic/racial groups that participants had likely encountered in their lives, measures in China often asked children about racial groups participants were unlikely to have encountered (e.g., Black or White people; Jizhe, 2021; National Bureau of Statistics, 2021; Qian et al., 2016).

According to the Seventh National Population Census (Jizhe, 2021), the population in China consists almost exclusively of Han Chinese (91%) and 56 other Chinese state-recognized nationalities or ethnic identities (including Zhuang, Hui, Manchu, and Uyghur). Therefore, we considered participants described as "Chinese" or "Han" as the majority children and children of all other ethnicities as minority children in our summary of study samples. Of the 36 measures, 35 were used with samples that consisted entirely of majority children; only one measure used a sample that included minority children.

Among the five other measures used in China, four measures asked children about Han and Uyghur people, and one measure asked children about "other ethnic groups." Our list of measures (OSF) contains information about the participants in these measures.

Israel

Most measures ($n = 35$ out of 41) used in Israel asked children about Jewish and Arab people, which are the two major ethnic groups that are commonly recognized by the Israel Central Bureau of Statistics (2022). Measures invoked these groups by labeling or depicting these groups directly (e.g., Arabs, Jews) or referencing or depicting specific subsets of these two groups (e.g., Jewish: "Israeli Jewish," "Russian Jews;" Arab: "Israeli Arab," "Palestinian"). Out of the 35 measures, many measures ($n = 25$, 71%) asked children about both Jewish and Arab people, whereas fewer measures asked children about just Jewish people ($n = 7$, 20%) or just Arab people ($n = 3$, 9%).

According to the Israel Central Bureau of Statistics (2022), 74% of people living in Israel are Jewish, 21% are Arab, and 5% are non-Arab Christian or people with other classifications. Therefore, we considered Jewish children (as well as children labeled as "Israeli Jewish" and "secular Jewish Israeli") as the majority children in Israel and children of all other ethnicities/races as minority children in our summary of study samples. Of the 35 measures that asked children about Jewish and Arab people, over half of them were used with samples that consisted of a mix of majority and minority children, and approximately equal numbers of measures were used with entirely majority and entirely minority samples (see Table 2).

Six other measures in Israel asked children about people from Ethiopia in addition to Jewish and Arab people ($n = 4$) or people who speak different languages or those from different countries ($n = 2$). Our list of measures (OSF) contains information about the participants in these measures.

Brazil

All 16 measures in Brazil asked children about racial/ethnic groups that align with the ethnic/racial groups commonly recognized by The Brazilian Institute of Geography and Statistics (IBGE; Instituto Brasileiro de Geografia e Estatística, 2022). The IBGE offers these groups as racial/ethnic identification options: Multiracial (Pardo), White, Black, Asian, and Indigenous. Fourteen measures asked just about Black and White people (88%), whereas two measures asked about Black, Pardo, and White people (13%).

According to the IBGE (2022), as of 2022, Pardo people have become the largest group in Brazil (45%), closely followed by White people (44%), and the other groups are Black people (10%), indigenous people (0.6%), and East Asian people (0.4%). More than half of the measures were used with samples consisting of a mix of Pardo participants and participants of other races, and the remaining measures were used with samples of Black and White participants (see Table 2 and our list of measures [OSF]).

South Africa

All 12 measures used in South Africa asked children about Black, Coloured, and/or White people. In two of these measures, additional specificity was given beyond these labels; specifically, the labels "Xhosa" and "Foreign Black" were used (in addition to Coloured and White) to describe targets. Xhosa people are a Bantu ethnic group native to South Africa; under apartheid and in the modern South African census, Xhosa people would be classified as African, Black, or Black African (Statistics South Africa, 2022). Of the 12 measures, 5 asked about Black and White people (42%), 2 asked about Black and Coloured people (17%), and 5 asked about all three target/involved groups (42%).

According to the 2022 South African census, Black Africans are the majority group in the country (81%), followed by Coloured people (8%), White people (7%), and Indian/Asian people (3%; Statistics South Africa, 2022). Of the 12 measures used in South Africa, most measures were used with samples consisting entirely of Black participants or a mix of Black and Coloured participants (see Table 2 and our list of measures [OSF]).

Australia

The five measures used in Australia asked children about "kids from your ethnic group" ($n = 2$), "people from other cultural groups"

($n = 1$), Asian and Caucasian people ($n = 1$), and Black and White people with French or Australian accents ($n = 1$). Given the variation in target groups, and therefore the variation in relevant ethnic/racial dimensions (see Q3 Method for what dimension refers to), participant groups are just reported in our list of measures (OSF).

Q3 Discussion

In sum, our findings reveal that most of the measures we extracted were used in North America and Europe. The fact that measures of children's ethnic and racial attitudes, stereotypes, and discrimination have been used in limited parts of the world has implications for establishing the appropriateness of these measures to be used across different ethnic, racial, and cultural contexts. Specifically, researchers studying children in European and North American contexts have prior research they can consult when selecting measures for their research. In contrast, we do not know how most children in the world would respond to existing measures in the field and we do not have enough use cases to evaluate whether existing measures can be used in valid ways in regions where we have limited reports.

The numbers reported above not only represent the frequency of measure usage, but they also reflect the amount of research on children's ethnic and racial attitudes, stereotypes, and discrimination that has been conducted and made visible to (i.e., published in English language journals in) our field. Consonant with the concerns that psychological researchers have continued to voice (e.g., Henrich et al., 2010; Moriguchi, 2023; Nielsen et al., 2017; Roberts & Mortenson, 2022), our results suggest that non-Western populations are underrepresented in published studies of children's ethnic and racial attitudes, stereotypes, and discrimination. Even though ethnic and racial biases are not unique to North American and European contexts, little research has been reported from outside of those contexts and populations. This is especially stark when considering that most children under 18 years of age live in Asia and Africa (see Fig. 2; United Nations, 2022).

As mentioned above, our search for articles was limited to peer-reviewed journal articles that were published in the English language. Therefore, we acknowledge that we would have missed any relevant measures that were used and reported in non-English publications. We do not claim that important work on children's ethnic and racial attitudes, stereotypes, and discrimination is completely absent or that relevant measures have not been used in parts of the world that we stated are "underrepresented" above. Rather, the point is that research and measure usage in those regions are underrepresented in English language publications. This underrepresentation is problematic, given that discussions and societies in developmental science at the international scale are conducted in the English language (e.g., "high impact" journals such as *Child Development* and organizations such as the Society for Research in Child Development), and those discussions and societies are powerful forces in decision-making that impacts children and communities across the world.

Further exploring the countries that contributed the largest number of measures in each continent, we found that the measures under review were used much less with ethnic/racial minority participants relative to majority participants in the Netherlands, China, and South Africa. In contrast, measures were used with minority participants almost as much as with majority participants in the U.S., Israel, and Brazil. These findings suggest that children's ethnic and racial attitudes, stereotypes, and discrimination may be understudied among minority children in some, but not all, countries/regions in the world. The results from the U.S., Israel, and Brazil are surprising given the commonly voiced concerns about the underrepresentation of ethnic/racial minority people in psychological research (e.g., Roberts & Mortenson, 2022; Rowley & Camacho, 2015); however, these results align with a pattern found by Roberts and colleagues (2020) showing that published developmental psychology studies that focused on topics of race included more participants of color than those that did not focus on topics of race. Our and Roberts et al.'s (2020) findings may reflect developmental scientists' increasing motivation to examine ethnic/racial minority children's thoughts and behaviors surrounding ethnicity and race in increasingly diverse societies (Rowley & Camacho, 2015).

Q4. What Evidence do we Have About Some of the Psychometric Properties of Commonly Used Scales/Tasks?

Thus far, we have reported on the landscape of measure design and usage: how and where researchers have measured children's ethnic and racial attitudes, stereotypes, and discrimination. Readers may also be curious to learn how much we know about the psychometric properties of measures that currently exist in our toolkit. In the last part of our review (Q4), we endeavored to review the landscape of evidence on psychometric properties of existing measures. We recognized that reviewing all the psychometric properties reported in articles for all 1,001 measures would not allow us to provide a comprehensible overview. As such, we decided to select the most commonly used types of measures (defined in Q4 Method) and assess the availability of psychometric reports and provide some initial evidence supporting their psychometric properties. We focused on psychometric properties that are likely to be critical for reliable and valid usage of measures among most developmental scientists—namely, test–retest reliability, predictive validity, and responsiveness to interventions. Below, we describe why we focus on these three psychometric properties.

Test-retest reliability, or the consistency and absence of errors between two or more measurements of the same individual when using the same measure under the same conditions (Aldridge et al., 2017; American Educational Research Association, 2014), is an important property of measures in the context of the current review for multiple reasons. First, although some theories in social psychology argue that stereotypes and prejudice are temporary mental associations or states that reflect currently present cues and contexts (e.g., Blair, 2002; Galvan & Payne, 2024; Payne et al., 2017), developmental scientists commonly consider children's ethnic and racial attitudes, stereotypes, and discrimination as stable attributes (in short developmental timespans, in the absence of interventions; e.g., Raabe & Beelmann, 2011; Rae & Olson, 2018). Therefore, variability in children's responses to the same measure over time are considered measurement errors that hinder our ability to accurately capture those constructs (American Educational

Research Association, 2014; Polit, 2014). Second, because measures are often used in longitudinal or intervention studies where children complete the same measure multiple times, we want to ensure that observed changes across time can be attributed to genuine changes among children rather than to the unreliability of the measure itself (Aldridge et al., 2017). Third, given that measures of ethnic and racial attitudes, stereotypes, and discrimination are often used to compare between groups of children (e.g., intervention groups, age groups), they need to capture individual children accurately so that we can be confident that any group differences we find are genuine and not an artifact of measurement error (Aldridge et al., 2017). In sum, it is important for developmental scientists to understand the extent to which children respond consistently to particular types of measures over time.

Next, the consideration of predictive validity is an important way to evaluate whether the use of a measure is warranted for particular study purposes (Kimberlin & Winterstein, 2008). Here, we focused on whether measures of children's ethnic and racial attitudes, stereotypes, and discrimination predict (i.e., statistically relate to) measures of children's real-world ethnic or racial behaviors (e.g., whom they decide to befriend, how they treat people in their communities). Significant statistical relations between these two measurements can serve as evidence that the measure of children's ethnic and racial attitudes, stereotypes, and discrimination can indeed be used to draw inferences about children's behaviors outside of the context of that measure. Assessing the extant amount of such evidence for different types of measures is important, as it can tell us which measures need further evaluation to understand what they truly capture.

Lastly, responsiveness, or the capacity of a measure to detect change over time when the construct has indeed changed, is another crucial quality of measures (Kimberlin & Winterstein, 2008). Finding that children's responses to a measure change due to an event or a manipulation that is expected to change the construct of interest is evidence that using the measure to capture the construct is highly valid. Moreover, given that measures of children's ethnic and racial attitudes, stereotypes, and discrimination are often used to evaluate the effects of interventions, it is especially important that those measures are responsive to interventions that would be expected to change the constructs captured by those measures (i.e., attitudes, stereotypes, and discrimination).

In the context of intervention studies, responsiveness of measures is consequential to how results can be interpreted. For example, consider an intervention study that tests children on a racial prejudice measure before and after an intervention program. If we found that children's prejudice on the measure decreased from pre- to post-test in the intervention (but not control) condition, then we have evidence both that the intervention was effective and that the measure is responsive. On the other hand, if we fail to find such a result, it could either be that the intervention is ineffective, or the measure is unresponsive; in this scenario, both the intervention and the measure need to be reconsidered. Thus, given the importance of considering measure responsiveness when choosing measures to use in studies, developmental scientists should know which types of existing measures tend to be responsive to interventions (i.e., yield significant intervention effects) and which ones have less evidence of responsiveness across existing studies.

This part of our review is intended to be an "evidence gap map" (Polanin et al., 2023; Snilstveit et al., 2016), or an overview that establishes where we have and do not have evidence telling us about the psychometric properties of existing measures. Therefore, our report should serve as a useful first step for developmental scientists who hope to conduct further research testing psychometric properties of measures (see General Discussion for more details about this future direction) or those who are choosing measures for their new study. For example, researchers who are choosing measures to use in their intervention study can learn which types of measures have frequently been responsive to interventions and may pick measures for which we have more existing cases where the measure was responsive.

Our intention is not to impose judgments on which types of measures are "high quality" relative to others, or to provide concrete values of psychometric qualities. As articulated by the American Educational Research Association (2014), there are no absolute criteria that make measures good or bad quality; rather, the reliability and appropriateness of using a measure for a particular purpose should be evaluated in the context of its intended use. In addition, the wide variability in study designs and analysis methods in the reviewed measures makes summary statistics collapsed across measures uninformative (Cooper, 2017; Cooper et al., 2019; see General Discussion for further discussion on this point). Therefore, we instead focus on the ranges of test–retest reliability metrics as well as the number of reports on test–retest reliability, predictive validity, and responsiveness that developmental scientists have at their disposal for each type of commonly used measure.

Q4 Method

Identifying Commonly Used Scales/Tasks

To identify commonly used types of measures, we worked within each of our measure categories (see Q1 Results) to first cluster measures together by methodological design. For example, in the *Implicit* category, measures that assessed children's associations between target groups and negative/positive concepts based on reaction times were grouped together in one cluster (*speeded reaction tasks* described below), while measures that assessed similar associations by using priming tasks were grouped together in a different cluster. We then selected clusters of measures that had been used more than twice per year on average over the past 13 years (i.e., more than 26 times). Selecting methodologically coherent measures within the broader measure categories allowed us to draw more concrete conclusions about the availability of psychometric evidence for specific measures. These clusters were highly methodologically coherent such that we treated measures within those clusters as the same "task" or "scale." In this way, we first identified nine commonly used scales/tasks. Then, we coded each instance of their use on test–retest reliability, predictive validity, and responsiveness to interventions.

Coding for Test-retest Reliability, Predictive Validity, and Responsiveness to Interventions

To find information about test–retest reliability, predictive validity, and responsiveness to intervention in each case of measure use,

we followed the coding procedures detailed below. For each psychometric property, our main goal was to report: the number of opportunities in which the indicators of the property could have been reported, the actual number of reports of those indicators, and the evidence supporting the property. Additionally, we coded for other relevant information that helps contextualize and further interpret the evidence; these additional pieces of information are reported in Tables S5-1 to S7-9.

Test-retest Reliability. Test-retest reliability is most often reported using correlation coefficients between timepoints (e.g., Pearson's correlation coefficient). It is also often reported using intraclass correlation coefficients (ICC) between timepoints (Aldridge et al., 2017; Polit, 2014). Therefore, we considered reports of correlation coefficients and ICC between timepoints (hereon referred to simply as "correlation coefficients") as indicators of test–retest reliability in the current review (for a discussion on the difference between and limitations of the two metrics, see Aldridge et al., 2017).

For gathering indicators of test–retest reliability, we first coded whether a scale/task was repeated by the same participants within a study (i.e., presence of an opportunity to report on correlations between timepoints). If a scale/task was repeated, we coded why and how many times the scale/task was repeated, the interval (i.e., time lapse) between repeated measurements, whether correlations between those repeated measurements were reported (i.e., reports of test–retest reliability indicators), and if so, the metric and value of the correlations (i.e., evidence of test–retest reliability).

Predictive Validity. We used the presence of any statistically significant correlations or effects relating the scales/tasks to children's actual ethnic or racial behavior outside the study context (hereon referred to as "real-world behavior measure") as evidence of predictive validity. Because the purposes for which the two measures were used varied across studies, these effects included main or simple effects as well as interaction effects and effects reported as part of more complex models.

To gather indicators of predictive validity, we first coded whether the study containing each scale/task also used real-world behavior measures (i.e., presence of an opportunity to report on predictive validity). If they did, then we further coded for: what the real-world behavior measure assessed, who completed the real-world behavior measure (e.g., the participant themselves, parents, or teachers), when the real-world behavior measure was completed relative to the scale/task, whether the statistical relations between the two measures were reported (i.e., reports of predictive validity indicators), and whether those relations were statistically significant (i.e., evidence of predictive validity). To note, we had initially hoped to assess relations between scales/tasks and direct reports of children's actual treatment of target group members in their real lives outside of the study context. However, we found only a single case in which such a real-world behavior measure was used alongside a commonly used scale/task (in Taylor et al., 2014; see description under *general affect scales* in Q4 Results). As such, we expanded our definition of real-world behavior measures to include friendship nomination measures, which could serve as indicators of children's voluntary intergroup behavior in the real world. Although we acknowledge that children's friendship nominations could be driven by forces other than children's own behavior (e.g., with whom their parents allow them to regularly play), for the purposes of coding, we assumed that children have some control over their own friendship choices.

Responsiveness to Interventions. We used the presence of any statistically significant intervention effects on scales/tasks as evidence of measures being responsive to interventions. To gather indicators of responsiveness to interventions, we first coded whether the study containing the scale/task implemented an intervention meant to change participants' ethnic or racial attitudes, stereotypes, or discrimination (i.e., presence of an opportunity to report on intervention effects). For the purposes of our review, we only considered intervention designs where the researchers assigned participants to two or more conditions (e.g., a control condition and an intervention condition; "arms" in Tables S7-1 to S7-9). If the study did involve such an intervention design, we recorded the nature of the intervention, whether the effect of the intervention on the scale/task was reported (i.e., reports of responsiveness indicators), and whether there was a significant effect of condition on the scale/task (i.e., evidence of responsiveness). We did not consider intervention designs in which there was only one condition, because contrasting conditions were necessary for us to evaluate whether responses on the scale/task could change due to the intervention manipulation (e.g., rather than due to time passage alone).

Note that in our results, we rely on different types of metrics for test–retest reliability versus predictive validity and responsiveness to offer maximally informative overviews of the existing evidence on each psychometric property using the results that were available in published articles. Test-retest reliability indicators were always reported as correlation coefficients or intraclass correlation coefficients (ICC) between repeated measurements, as we described above. Thus, we were able to report a clear range of r values; however, we could not determine whether test–retest correlations were significant as the significance of the correlations was rarely reported. In contrast, when considering predictive validity and responsiveness, we broadened our criteria to look beyond simple correlations as basic correlations were rarely reported but we still wished to gather as much information as possible. As such, we looked for any reports of statistical relations/effects between variables to use as indicators of predictive validity and responsiveness. Given the variety of metrics (e.g., regression coefficients, correlation coefficients) and analysis models (e.g., regression models that included our variables of interest as part of interactions, structural equation models that treated our variables of interest as outcome variables), we could not summarize a simple range. Instead, we examined the presence or absence of significant results to assess whether there was any signal to suggest the measure's predictive validity or responsiveness; counts of those signals represent the amount of evidence that exists.

Q4 Results

Below, descriptions and reports on the evidence supporting psychometric properties of the nine commonly used scales/tasks are organized by the measure categories introduced in section Q1. Table 3 provides a summary of the number of cases reviewed for each commonly used scale/task, and Fig. 3 is a graph (or an evidence gap map) depicting the total number of measures and the number of reports on each psychometric property for each scale/task.

Some caution is warranted when interpreting the results that we provide here. First, for test–retest reliability, we provide the range of correlation coefficient values because the coefficient ranges could be useful for gleaning the variability that exists in the available evidence. However, readers should practice caution when interpreting those ranges because the contexts in which repeated measurements occurred varied across studies. Although we reviewed all cases in which correlation coefficients between repeated measurements were reported in articles, the informativeness of those correlation coefficients for evaluating the test–retest reliability of the measure may vary depending on the context in which the repeated measurements occurred. For example, if the repeated measurements were taken two days apart, during which no interventions or notable events took place, the correlation between those measurements is likely a strong indicator of reliability of the measure. On the other hand, if the repeated measurements were taken across a long time span or if children participated in an intervention between measurements, then the correlation between those measurements may be a weaker indicator of the reliability of the measure itself. Moreover, given that the correlation between measurements can be expected to differ depending on events and the amount of time that passes between measurements, it would be inappropriate to directly compare between correlation coefficients across studies to draw conclusions about which scales/tasks have stronger test–retest reliability than others. Most importantly, there are no absolute correlation coefficient values that deem measures reliable or unreliable (American Educational Research Association, 2014), so the values themselves cannot be interpreted without considering the context in which the repeated measurements occurred. Readers who are interested in learning more about the test–retest reliability of measures should utilize Tables S5-1 to S5-9 to evaluate individual correlation coefficients.

As another point of caution, although we use the presence of significant statistical results as evidence of predictive validity and responsiveness, statistical significance is dependent on various methodological characteristics. Similar to our review of test–retest reliability indicators, we extracted the predictive validity and responsiveness indicators across studies that differed in many ways, including the types of real-world behavior measures used and the types of interventions that were implemented. Relying on information about statistical significance across a variety of study designs is useful for gaining a coarse overview of whether there are any signals that the measures predict real-world behaviors or that they are responsive to interventions. However, readers who wish to learn more about particular results should refer to Tables S6-1 to S7-9 where we provide more contextual information that is necessary for interpreting individual study results.

Trait Attributions Category

Application of Valenced Traits Tasks. *Application of valenced traits tasks* were used 57 times in the *Trait Attribution* category. *Application of valenced traits tasks* included the Preschool Racial Attitude Measure II (PRAM II; Williams et al., 1975), the Multi-Response Racial Attitudes Measure (MRA; Doyle & Aboud, 1995), and their adaptations. The tasks asked children to attribute positively- or negatively-valenced traits to targets in an alternative-choice format. For example, participants heard, "Some children are mean. They say and do nasty things to other children. Who is mean?" and had to choose which children in the presented pictures fit the trait. Each of these tasks presented multiple traits to probe children's general positive or negative attitudes about the targets rather than their beliefs about the targets regarding one trait. Specifically, the tasks tapped multiple dimensions, including: integrity/trustworthiness, niceness/prosociality, intelligence/capability, appearance/hygiene, playfulness/cheerfulness, and/or humility. *Application of valenced traits tasks* were used with children ages 3 to 17.99 (our maximum inclusion age).

As is apparent from Table 3, evidence supporting test–retest reliability and predictive validity of *application of valenced traits tasks* was nonexistent, and evidence supporting responsiveness to intervention was almost nonexistent (only $n = 2$). The two cases suggesting the responsiveness of *application of valenced traits tasks* revealed that children who experienced imagined intergroup contact showed less biased attribution of valenced traits to racial out-groups versus racial in-groups, compared to those who did not experience such an intervention. Further details such as reasons for repeated measurement (e.g., pre- and post-design, longitudinal study) and type of interventions used are provided in Tables S5-1, S6-1 and S7-1.

Degree of Valenced Traits Scales. *Degree of valenced traits scales* were used 53 times in the *Trait Attributions* category. These measures included scales such as the Black/White Evaluative Trait Scale (BETS; Hughes & Bigler, 2007) and its adaptations. These scales tapped children's beliefs about the degree to which positively- or negatively-valenced traits applied to the targets or their level of agreement with statements which applied those traits to the targets. For example, children were presented with the statement "I think most Moroccan children are honest," and they were asked to rate their agreement on a 5-point scale ranging from 1 = totally disagree to 5 = totally agree. Each of these scales presented multiple traits to tap children's general positive or negative attitudes about the targets rather than their beliefs about the targets regarding one trait. Specifically, the scales tapped multiple dimensions, including: integrity/trustworthiness, niceness/prosociality, smartness/capability, appearance/hygiene, funness, and/or humbleness. These scales were used with children ages 4 to 17.99 (our maximum inclusion age).

As shown in Table 3, a handful of reports provided evidence supporting the test–retest reliability, predictive validity, and responsiveness of *degree of valenced traits scales*. Specifically, seven cases of repeated use of *degree of valenced traits scales* showed that the correlation between repeated measurements ranged from $r = 0.31$ to 0.74. Eight reports showed that higher quantities of friends with a target group or intergroup friendships in general were associated with less biased attribution of traits to target groups, supporting the predictive validity of *degree of valenced traits scales*. Moreover, 10 reports demonstrated that children who experienced direct, indirect, or imagined intergroup contact and had classroom discussions ($n = 6$), went through trainings that aim to increase cognitive, emotional, mindfulness, or compassion skills ($n = 3$), or watched racially diverse television programs and had discussions about race with their parents ($n = 1$), attributed fewer negative traits and more positive traits to target groups compared to children who did not experience such interventions, providing evidence for the responsiveness of *degree of valenced traits scales*. See Tables S5-2, S6-2, and S7-2 for more details on these reports.

Table 3
Reports by Commonly used Scales/Tasks.

Common Scale/Task	Total use	Test-retest reliability			Predictive validity			Responsiveness to interventions		
		Measure repeated	Cases of reported correlations	Correlation coefficient range	Real-world behavior measured	Cases of reported relations	Cases of significant relations	Used in intervention study	Cases of reported intervention effects	Cases of significant intervention effects
Category: Trait Attributions										
Application of Valenced Traits Tasks	**57**	10 (18%) (S5-1)	0	NA	0 (S6-1)	0	0	7 (12%) (S7-1)	7 (12%)	2 (4%)
Degree of Valenced Traits Scales	**53**	17 (32%) (S5-2)	7 (13%)	0.31 to 0.74	9 (17%) (S6-2)	8 (15%)	8 (15%)	14 (26%) (S7-2)	14 (26%)	10 (19%)
Category: Social Interactions										
Interaction Intention Scales	**62**	18 (29%) (S5-3)	6 (10%)	−0.19 to 0.58	20 (32%) (S6-3)	20 (32%)	18 (29%)	15 (24%) (S7-3)	15 (24%)	12 (19%)
Interaction Partner Selection Tasks	**52**	11 (21%) (S5-4)	4 (8%)	0.08 to 0.35	3 (6%) (S6-4)	2 (4%)	2 (4%)	6 (12%) (S7-4)	6 (12%)	2 (4%)
Affective Reaction to Interaction Scales	**35**	12 (34%) (S5-5)	6 (17%)	0.26 to 0.79	16 (46%) (S6-5)	14 (40%)	11 (31%)	7 (20%) (S7-5)	7 (20%)	4 (11%)
Category: Liking and Feelings										
General Affect Scales	**72**	13 (18%) (S5-6)	5 (7%)	0.30 to 0.65	10 (14%) (S6-6)	7 (10%)	6 (8%)	10 (14%) (S7-6)	9 (13%)	7 (10%)
Category: Anti-Racism										
Evaluation of Exclusion Tasks	**28**	5 (18%) (S5-7)	0	NA	3 (11%) (S6-7)	0	0	5 (18%) (S7-7)	5 (18%)	5 (18%)
Category: Implicit										
Speeded Association Tasks	**49**	15 (31%) (S5-8)	4 (8%)	−0.17 to 0.39	3 (6%) (S6-8)	1 (2%)	1 (2%)	11 (22%) (S7-8)	10 (20%)	8 (16%)
Category: Prosociality										
Resource Allocation Tasks	**29**	1 (3%) (S5-9)	1 (3%)	0.62	0 (S6-9)	0	0	4 (14%) (S7-9)	4 (14%)	2 (7%)

Note. Under each psychometric property are: the number of opportunities in which the property could have been reported, the actual number of reports of the property, and the evidence supporting the property (i.e., correlation coefficient ranges for test–retest reliability and number of statistically significant cases for predictive validity and responsiveness to intervention). Percentages indicate the proportion of cases out of total use of each scale/task. Labels S5-1 to S7-9 indicate the supplemental tables where details are reported.

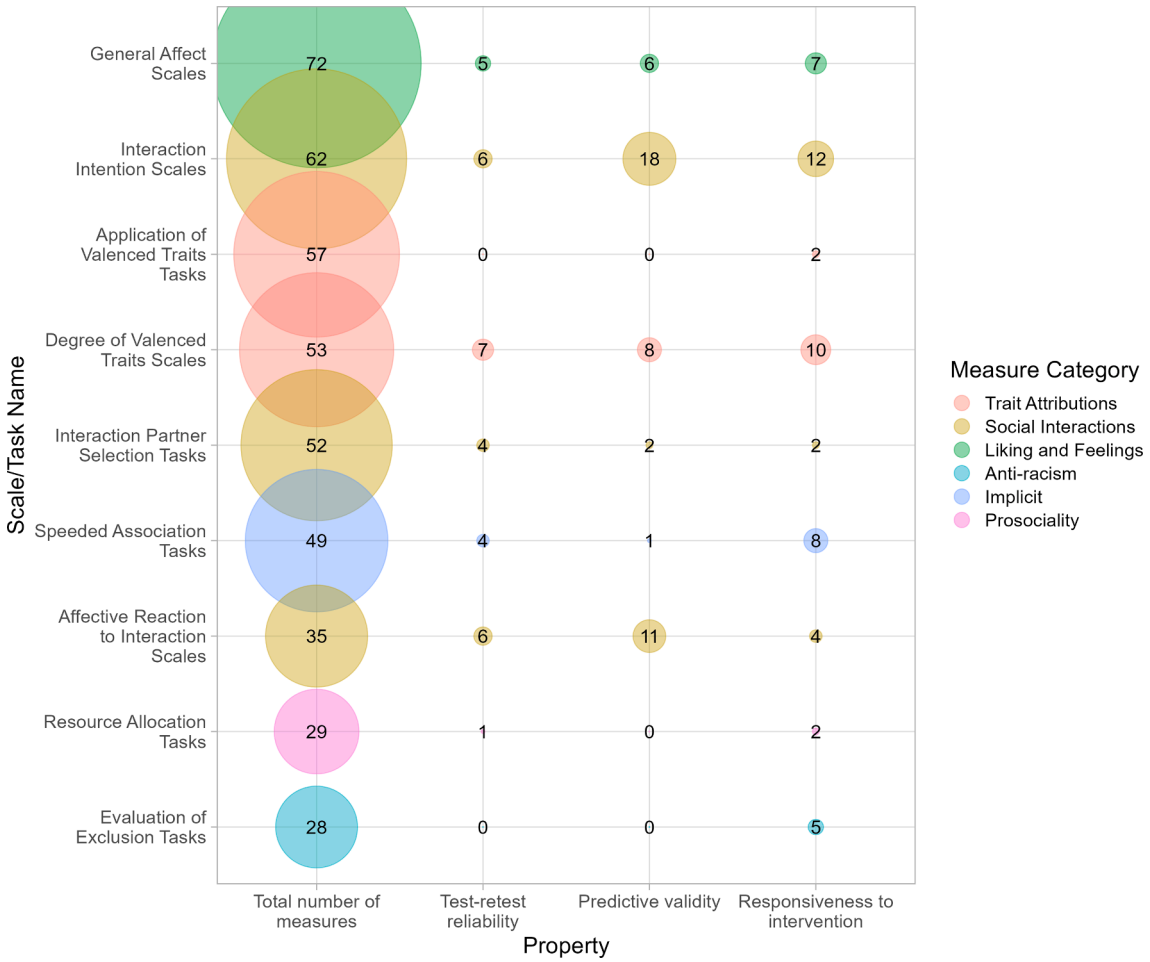

Fig. 3. Number of Cases by Commonly used Scales/Tasks. The commonly used scales/tasks are presented in descending order of total number of measures. The sizes of the circles and the numbers in the graph represent the total number of measures, number of reported correlations (for test-retest reliability), and number of significant relations/effects (for predictive validity and responsiveness). The colors of the circles indicate the measure categories to which the scales/tasks belong.

Social Interactions Category

Interaction Intention Scales. *Interaction intention scales* were used 62 times in the *Social Interactions* category. *Interaction intention scales* included adaptations of Bogardus's (1933) Social Distance Scale. In these scales, children expressed their willingness to, interest in, and intentions to interact with different targets in a variety of contexts. The interaction contexts varied widely between scale items, even within a single measure. For example, children were asked to rate how much they would like to sit next to a peer from the target group on the school bus, have them as neighbors, have them as a friend, have them over for a sleepover, and so on. However, if there were multiple items, children's responses were collapsed across items during analyses. Scales also varied in number of points and what the end points signified (e.g., "no way" to "for sure yes," "strongly disagree" to "strongly agree"). Despite the variation, these scales all tapped children's willingness to, interest in, or intention to interact with targets. These scales were used with children ages 3 to 17.99 (our maximum inclusion age).

As shown in Table 3, we found more reports to support the psychometric properties of focus for *interaction intention scales* relative to the other commonly used scales/tasks. Six cases of repeated *interaction intention scales* showed that the correlation between repeated measurements ranged widely, from $r = -0.19$ to 0.58. Eighteen reports showed that having more friends belonging to the target group or intergroup friendships in general was associated with expressions of more positive interaction intentions, such as lower avoidance and stronger interest in interacting with target group members, providing evidence for predictive validity of *interaction intention scales.* Lastly, 12 reports revealed that children who experienced direct or extended intergroup contact ($n = 10$), received messages that prejudice is malleable ($n = 1$), or went through trainings that aimed to increase cognitive, emotional, mindfulness, or compassion skills ($n = 1$), showed more positive attitudes on *interaction intention scales* than those who did not experience such interventions, demonstrating the responsiveness of those scales. See Tables S5-3, S6-3, and S7-3 for more details on these reports.

Interaction Partner Selection Tasks. *Interaction partner selection tasks* were used 52 times in the *Social Interactions* category. The tasks asked children to select interaction partners from options of targets to tap children's relative willingness, interests, and intentions to interact with different targets. The probes varied from asking children which child target they preferred to interact with (e.g., play with, sit next to) to asking children to select between adult targets that would choose to interact with them in different capacities (e.g., teacher, swim coach, doctor). *Interaction partner selection tasks* were used with children ages 3 to 17.99 (our maximum inclusion age).

A few reports provided evidence supporting the test–retest reliability, predictive validity, and responsiveness of *interaction partner selection tasks* (see Table 3). Four cases of repeated use of *interaction partner selection tasks* showed that the correlation between repeated measurements ranged from $r = 0.08$ to 0.35. Just two reports provided evidence supporting the predictive validity of *interaction partner selection tasks*: they showed that children's friendships with ethnic/racial out-group peers were associated with less biased partner selection and increased likelihood of choosing out-group targets during the task. Two other reports provided evidence for the responsiveness of *interaction partner selection tasks*: they showed that children who experienced intergroup contact through cooperative or physical activities had reduced in-group favoritism in their interaction partner selection compared to those who did not experience such interventions. See Tables S5-4, S6-4, and S7-4 for more details on these reports.

Affective Reaction to Interaction Scales. *Affective reaction to interaction scales* were used 35 times in the *Social Interactions* category. *Affective reaction to interaction scales* included single- or multiple-item scales that probed children's emotional or affective reactions when interacting or imagining interacting with targets. The scales assessed intensity of emotions or agreement to statements about having those emotions. Emotions ranged from positive (e.g., comfortable) to negative (e.g., anxious). These scales were used with children 5 to 17.99 years (our maximum inclusion age).

Several reports provided evidence supporting the test–retest reliability, predictive validity, and responsiveness of *affective reaction to interaction scales* (see Table 3). Specifically, six cases of repeated use of *affective reaction to interaction scales* showed that the correlation between repeated measurements ranged from $r = 0.26$ to 0.79. Eleven reports provided evidence supporting the predictive validity of *affective reaction to interaction scales*: results from 10 cases showed that higher quantities of friends with a target group or intergroup friendships in general were associated with lower anxiety and higher comfort regarding interacting with target out-group members; results from one case showed that higher quantities of friends with the target group were associated with higher anxiety with target group members. Lastly, four reports provided evidence supporting the responsiveness of *affective reaction to interaction scales*: children who experienced extended intergroup contact ($n = 2$), or were exposed to positive exemplars from target groups ($n = 1$) showed less anxiety and more comfort around interactions with target groups compared to those who do not experience such interventions; one case showed that children who experienced extended intergroup contact showed more anxiety interacting with the target group compared to those who did not experience such contact. See Tables S5-5, S6-5, and S7-5 for more details on these reports.

Liking and Feelings Category

General Affect Scales. *General affect scales* were used 72 times in the *Liking and Feelings* category. *General affect scales* included scales such as "feeling thermometers" and "seven faces scales" (Yee & Brown, 1992; Wilcox et al., 1989). Scales varied in number of points (range: 3- to 100-point scales) and what the end points signified (e.g., "cold" to "warm", "negative" to "positive", "really dislike" to "really like"), but they all probed children's overall general positive or negative affect toward a verbally labeled target group or exemplars belonging to a target group. These scales were used with children ages 4 to 17.99 (our maximum inclusion age).

As shown in Table 3, we found several reports to support the test–retest reliability, predictive validity, and responsiveness of *general affect scales.* Five cases of repeated *general affect scales* showed that the correlation between repeated measurements ranged from $r =$ 0.30 to 0.65. Six reports provided evidence supporting the predictive validity of *general affect scales*: five reports showed that having more friends belonging to the target group was associated with more positive general affect toward that group; one report showed that the more children behaved prosocially toward the out-group (as reported by children's mothers), the more they expressed positive general affect toward their religious out-group (Taylor et al., 2014). Lastly, seven reports provided evidence supporting the

responsiveness of *general affect scales*: six cases showed that children who experienced extended or imagined intergroup contact interventions ($n = 3$), were exposed to positive exemplars of target groups ($n = 2$), or heard about the unique challenges of a racial outgroup member ($n = 1$) expressed more positive attitudes on *general affect scales* than those who did not go through such experiences; one case showed responsiveness in an unexpected direction, such that children who did not experience extended intergroup contact showed more positive attitudes on *general affect scales* than those who did. See Tables S5-6, S6-6, and S7-6 for more details on these reports.

Anti-Racism Category

Evaluation of Exclusion Tasks. *Evaluation of exclusion tasks* were used 28 times in the *Anti-Racism* category. These tasks presented children with intergroup contact scenarios and probed them to evaluate the extent to which acts of exclusion are permissible. For example, in one case, children were presented with a scenario depicting a European American child (Michael) and an African American child (Doug), and asked "what if Michael doesn't invite Doug to lunch because he thinks Doug won't fit in? How good or bad is that?" children responded on a scale ranging from 1 = "very, very good" to 8 = "very, very bad." Particular scenarios and exact probes differed between cases, and the response options were typically good-bad or okay-not okay. *Evaluation of exclusion tasks* were used with children ages 3 to 17.99 years (our maximum inclusion age).

As is apparent from Table 3, evidence supporting test–retest reliability and predictive validity of *evaluation of exclusion tasks* was nonexistent, while there was some evidence supporting responsiveness to intervention ($n = 5$). The five cases suggesting the responsiveness of *evaluation of exclusion tasks* revealed that children who experienced direct or indirect intergroup contact and received cognitive/emotional training or had classroom discussions showed increased rejection of intergroup exclusion compared to those who did not experience such interventions. Further details on these reports can be found in Tables S5-7, S6-7, and S7-7.

Implicit Category

Speeded Association Tasks. *Speeded association tasks* were used 49 times in the *Implicit* category. They included traditional IATs (Greenwald et al., 2003), child-adapted IATs (i.e., modified versions of the traditional IAT appropriate for children, who have limited reading abilities and attention spans; Baron & Banaji, 2006; Rutland et al., 2005), Implicit Racial Bias Tests (IRBT; Qian et al., 2016), and single-target IATs (Mähönen et al., 2010). In these measures, children saw faces of target group exemplars or verbal labels (e.g., stereotypical names) on a computer or tablet screen and matched (typically using a keyboard) those targets to negatively- or positively-valenced stimuli as quickly as possible. The difference in reaction time between matchings (i.e., associations) was taken as indicators of implicit prejudice. *Speeded association tasks* were used with children ages 3 to 17.99 (our maximum inclusion age).

As shown in Table 3, some reports provided evidence supporting the test–retest reliability, predictive validity, and responsiveness of *speeded association tasks*. Specifically, four cases of repeated use of *speeded association tasks* showed that the correlation between repeated measurements ranged from $r = $ -0.17 to 0.39. Only one report provided evidence to support the predictive validity of *speeded association tasks*: it revealed a counterintuitive result, such that the more Italian 7- to 9-year-olds listed immigrant peers as their friends, the more prejudiced implicit attitudes they showed against immigrant children on the immigrant-Italian Child-IAT (Vezzali et al., 2012). Lastly, eight reports provided evidence to support the responsiveness of *speeded association tasks*: children who were exposed to positive exemplars from target groups ($n = 2$), went through trainings that increase empathy and perspective-taking ($n = 1$), or went through trainings that increase individuation of target group members ($n = 3$), showed less prejudice on *speeded association tasks* than those who did not experience such interventions; one study showed that children who experienced direct contact with an out-group member demonstrated more prejudice on *speeded association tasks* than those who did not experience such contact; another experiment showed that children who were experimentally exposed to essentialist beliefs about ethnicity demonstrated heightened prejudice compared to those who were not exposed to such beliefs. See Tables S5-8, S6-8, and S7-8 for more details on these reports.

Prosociality Category

Resource Allocation Tasks. *Resource allocation tasks* were used 29 times in the *Prosociality* category. In these tasks, children were asked to allocate resources between the targets and themselves, decide how many resources to give to targets, give resources to targets as a reward, or take away resources from targets as punishment. For example, children were given a total of 10 stickers that they could take home, but they were also told that the laboratory did not have enough stickers for other children who were allegedly coming to the laboratory the next day. Children were then shown pictures of those children and asked to put stickers in envelopes to share with them if they wished. The number of stickers placed in the envelope for each target was later counted. *Resource allocation tasks* were used with children ages 1 to 17.99 years (our maximum inclusion age).

As shown in Table 3, evidence supporting test–retest reliability and responsiveness of *resource allocation tasks* was almost nonexistent (only $n = 1$ and $n = 2$, respectively), while evidence supporting predictive validity was entirely nonexistent. The only case of repeated use of *resource allocation tasks* showed that the correlation between repeated measurements was $r = 0.62$. In the two reports supporting the responsiveness of *resource allocation tasks*, children who experienced direct or imagined intergroup contact showed less biased resource allocation decisions compared to those who did not experience such interventions. Further details on these reports are provided in Tables S5-9, S6-9 and S7-9.

Q4 Discussion

In sum, we examined nine commonly used scales/tasks to provide an overview of the evidence that is currently available on their test–retest reliability, predictive validity, and responsiveness to interventions. On the one hand, we found hopeful hints that the

commonly used scales/tasks have properties that make them useful tools for our field: numerous results indicated that the scales/tasks are statistically significantly related to number of friends children have who belong to the target group (see Tables S6-1 to S6-9), and that the scales/tasks are responsive to interventions (most evidently to interventions that involve intergroup contact; see Tables S7-1 to S7-9). On the other hand, we observed that evidence for test–retest reliability, predictive validity, and responsiveness is sparse relative to the frequency of usage (as is clear from Fig. 3). For example, the *application of valenced traits tasks* were used 57 times, yet we found no reports of its test–retest reliability or predictive validity, and only two cases suggesting its responsiveness to interventions. Even when taking the most commonly used *general affect scales* (used 72 times), there were only 5 reports of test–retest reliability, 6 cases showing their predictive validity, and 7 cases showing their responsiveness to interventions. Thus, studies frequently present these scales/tasks to children, yet we know comparatively little about their consistency, ability to predict children's real-world behaviors, or whether they are responsive to interventions that should cause changes in children's ethnic and racial attitudes, stereotypes, and discrimination.

The fact that we have so little evidence about the psychometric properties of the most popular measures in our toolkit is problematic, as it indicates that researchers are frequently using measures for which reliability and validity are unestablished. For example, researchers may find that extraverted preschoolers express stronger desires to interact with racial out-group peers in *interaction intention scales* compared to their introverted counterparts; but if we do not know whether *interaction intention scales* predict children's behaviors in the real world, we cannot be sure that the conclusion that extraverted preschoolers behave more inclusively than introverted preschoolers would be warranted. Relatedly, we do not have enough evidence as a field for researchers to make informed decisions about which measures to select for intervention studies or for making inferences about children's real-world behaviors. Thus, as we raise in the General Discussion, future efforts focused on examining the psychometric properties of measures reviewed here are critical for ensuring a productive and rigorous science for understanding and addressing children's ethnic and racial attitudes, stereotypes, and discrimination.

Lastly, to provide substantial overarching insights on the availability and nature of psychometric properties of commonly used scales/tasks, we had to assess instances of scale/task use collapsed across geographic contexts, participant characteristics (e.g., age, ethnicity, race), and specific target groups (i.e., ethnicity/race of the targets being evaluated). However, assessments of psychometric properties and measure selection should ideally be conducted with careful consideration toward context and purpose of measure use. This is because measure quality and appropriateness will depend on the particular context, and the validity of using a particular type of measure will depend on the purpose of using that measure. For example, it is possible that *interaction partner selection tasks* are highly predictive of playmate choice on a real playground among preschoolers but not among older children; if this is the case, then drawing conclusions about real-world behavior based on responses to *interaction partner selection tasks* are warranted when the study is about preschoolers, but not when the study is about middle schoolers. Therefore, we invite readers to utilize the list of measures on OSF as a starting point for evaluating (or knowing whether there is enough information to evaluate) the feasibility, reliability, and validity of different types of measures for their purposes. For example, one could look at the 16 measures used with early school-aged children in Brazil to evaluate which type of measure is feasible or can be used reliably to draw useful conclusions in a correlational study in the Brazilian context. In our General Discussion, we provide more suggestions for utilizing our list of measures to further the field's understanding of measures in our existing toolkit.

General Discussion

Summary

The goal of the current work was to review modern measures of children's ethnic and racial attitudes, stereotypes, and discrimination. By conducting a systematic review of measures used in articles published between 2010 and 2022, we identified a wide variety of measures in our field's toolkit, which we summarized into eight broad measure types. In the measures we identified, targets of evaluation were typically represented using verbal descriptors (e.g., labels of ethnic/racial groups) and visual exemplars (e.g., drawings of individual group members). Furthermore, measures were used in studies conducted across five continents, but most often in Europe and North America. Analysis on the measures used within each continent's top country(ies) suggested that children's ethnic and racial attitudes, stereotypes, and discrimination may be understudied among ethnic/racial minority children in some, but not all, countries/regions in the world. Lastly, upon further assessing the types of scales/tasks that were most frequently used, we found that evidence for test–retest reliability, predictive validity, and responsiveness was sparse relative to measure usage.

The advantage of the scoping approach of the current review was that we were able to provide a broad overview of measures focused on children's ethnic and racial attitudes, stereotypes, and discrimination. Such an overview is necessary to elucidate trends in measure design and usage, which is a critical step for identifying directions for future discussions and research in the field. Below, we describe some of the issues that we believe should be prioritized in future efforts, and offer recommendations for tackling those issues through examining our list of measures (see OSF), conducting new empirical studies, and improving reporting practices. A summary of our recommendations appears in Table 4.

Clarify how Existing Measures Relate to Constructs of Interest

As mentioned in our introduction of and discussion following Q1, it is important for the field to have clarity on how the different types of measures in our field's toolkit relate to constructs that researchers are interested in capturing. However, existing efforts have not focused on elucidating these relations. A critical starting point for clarifying measure-construct relations is establishing how

Table 4
Summary of Recommendations.

Current problem: Measure-construct relations are unclear. **Future goal:** Clarify how existing measures relate to constructs of interest.	
Action item:	**Resources/models:**
Pick a measure type of interest. Then, carefully evaluate and discuss what construct it seems to capture, how its creators/users have operationalized that construct, and how one could draw inferences about that construct based on children's responses to it.	- Measure list (OSF) - Measure category (Q1 Results) - Proposed measure-construct links (Q1 Discussion)
Current problem: Measure-to-measure relations are unclear. **Future goal:** Clarify how existing measures relate to each other.	
Action items:	**Resources/models:**
Examine articles to find existing evidence showing statistical relations between measures. Find out whether a hypothesized conceptual link can be supported by existing data.	- Article catalog (OSF) - Measures list (OSF)
Conduct new studies to test how the same children respond to different types of measures and examine the statistical relations between those measures.	- Rae & Olson (2018) - deMayo & Olson (2024)
Current problem: Implications of using different types of stimuli are unclear. **Future goal:** Examine how stimulus type impacts children's responses.	
Action items:	**Resources/models:**
Conduct new studies to investigate how children interpret and respond to verbal and visual representations of ethnic/racial groups.	- Guerrero et al. (2010) - Williams & Steele (2019)
Conduct new studies to test how specific characteristics of visual stimuli impact children's perception of and responses to measures.	- Dunham et al. (2015) - Lei et al. (2024)
Current problem: Evidence for psychometric properties of commonly used measures is sparse. **Future goal:** Conduct studies focused on psychometric properties of measures.	
Action items:	**Resources/models:**
Conduct new studies to examine the reliability and validity of measures while holding other factors constant.	- Rae & Olson (2018) - Williams & Steele (2016)
Conduct new studies that test whether measures predict children's behaviors in the real world by using adult reports, observations, and automated technologies.	- Martin & Fabes (2001) - Elbaum et al. (2024)
Current problem: Reports often do not share enough rationales or details of measures. **Future goal:** Increase consistency, transparency, and details in reports.	
Action items:	**Resources/models:**
Correctly reference sources of measures, and clearly describe and justify changes to the original measure.	- Yu et al. (2022) - Dhont & Van Hiel (2012) - Williams et al. (2021)
Justify measure choice and articulate what construct or phenomena the measures are meant to tap.	- Aldana et al. (2019) - Clark et al. (2017)
Make all items, participant instructions, detailed characteristics of visual stimuli, and visual stimuli themselves available to readers.	- Mandalaywala et al. (2021) - Ghavami et al. (2020) - Renno & Shutts (2015)

researchers are thinking conceptually about measures within our existing toolkit. As such, we invite researchers to view our list of measures (see OSF) and use our measure categorizations to closely examine existing measures by evaluating what construct each measure seems to capture, how the measure creators/users have operationalized that construct, and how one could draw inferences about that construct based on children's responses to the measure.

We also suggest developmental scientists think critically about the constructs themselves. The inconsistent usage of the words "attitudes," "stereotypes," and "discrimination," as well as ambiguity in measure-construct relations in our existing toolkit (see introduction of Q1), could be attributed to the lack of careful label usage among developmental scientists; however, these practices could also reflect the weak conceptual grounding of our field's thinking about social biases among children. Although using the tripartite model (i.e., thinking of attitudes, stereotypes, and discrimination as the three types of social bias; Correll et al., 2010; Dovidio & Gaertner, 2010) can be useful for describing the different types of social group phenomena out in the world, developmental scientists may need to establish a more specific framework that describes the manifestation of those phenomena within children's thoughts and actions. Establishing such a framework should in turn lead to a widespread consensus on measure-construct relations and increase the consistent use of construct labels.

Clarify how Existing Measures Relate to Each Other

After researchers establish their thinking about how measures map onto constructs, they should consider how different types of measures are conceptually similar to or different from each other. To examine these measure-to-measure relations, researchers need to develop theory-driven ideas about how different types of measures relate conceptually and then test the statistical relations among measures. That is, conceptual relations among measures inform how those measures *should* relate statistically. Measures that are

conceptualized as tapping similar constructs should be strongly statistically related (e.g., correlated), while those that are conceptualized as tapping dissimilar constructs should be more weakly correlated (for discussions on the "nomological net," see Hoyle et al., 2002; Judd & Kenny, 1981). Therefore, future efforts to clarify the statistical relations among measures should be guided by our understanding of conceptual relations among those measures. If statistical tests support the conceptual link between measures, we can be more confident that those measures indeed capture similar constructs. On the other hand, if statistical tests do not concur with the conceptual link, then measures may need to be re-evaluated as to whether they capture the constructs they are meant to capture.

One way to examine statistical relations among measures is to utilize the list of articles that were retained in our filtering process (i.e., article catalog) and the list of measures that we have shared publicly (see OSF). Because there are myriad combinations of measure types that could be examined, we did not analyze measure-to-measure relations in the current review. However, researchers could use our article catalog and measure list to answer their own questions about whether and how particular measures related to one another. For example, if readers conceptualize children's prosocial behaviors toward a group as closely tied to children's general affect toward that group and wish to examine the statistical relation between *Prosociality* measures and *Liking and Feelings* measures, they can access our article catalog to find all the articles which used measures from both of those categories. Then, readers can find out the names of measures from the measure list and review the articles to see if the statistical relation between measures was reported, and if so, answer whether the hypothesized conceptual link can be supported statistically.

Another way to examine the statistical links between measures is to conduct new empirical studies which test the same children on different types of measures. Such studies will allow us to directly explore the statistical relations between a variety of existing measures and test whether conceptual similarities or overlaps between measures can be supported by data. As one example, developmental scientists could follow the manner in which social psychologists have examined relations between "implicit" and "explicit" measures; these examinations have provided useful and complex insights regarding whether those two groups of measures capture distinct constructs (for studies with adults, see Cameron et al., 2012; Greenwald et al., 2009; Hofmann et al., 2005; Kurdi et al., 2019; for a study with children, see Rae & Olson, 2018).

As another model, consider deMayo & Olson (2024), who conducted a within-subjects study where 5- to 6-year-old U.S. children responded to a forced-choice racial preference task (i.e., child chooses which of the two photographed targets they like the most) and a rating scale (i.e., child rates how much they like the photographed target on a 6-point smiley scale). deMayo & Olson (2024) tested the statistical relations between the two measures and concluded that they are correlated with each other, but that children's in-group preference was stronger in the forced-choice task than in the rating scale. These results suggest not only that the two types of measures tap the same (or similar) concepts, but also that the forced-choice task pulls the measure scores toward more extreme values compared to the rating scale which allows children to express more graded preferences. Thus, future studies that compare results from the same children on multiple measures could help confirm whether measures tap overlapping or distinct constructs. Furthermore, such studies have the additional benefit of informing how particular characteristics (e.g., response format, single or simultaneous presentation of target photographs) of measures impact the strength (i.e., effect sizes) or presence of expressed biases, which in turn could help the field explain discrepancies in conclusions across different studies using slight variations of similar measures (for more examples on methodological differences impacting results, see Hofmann et al., 2005; Rae & Olson, 2018; Reyes-Jaquez et al., 2021).

Examine how Stimulus Type Impacts Children's Responses

In the introduction of Q2, we mentioned that information about detailed characteristics of stimuli (e.g., exact format, gender and age of the targets depicted in visual stimuli, information on stimuli creation and norming) were often missing from reports. Such trends in reporting seem to reflect the fact that our field has not focused thus far on the implications of using different types of stimuli in measures of children's ethnic and racial attitudes, stereotypes, and discrimination. Some existing studies hint that children's responses can be influenced by the way in which target people or groups are represented in measures (Guerrero et al., 2010; Stengelin et al., 2023; Williams & Steele, 2019). Thus, different ways of representing targets and the details of stimuli could have significant consequences for the conclusions that we draw. Moreover, although measures for younger vs. older children seem to use visual stimuli and verbal labels at different frequencies (see Q2 Results), reflecting researchers' intuitions about what children of different ages understand, we argue that more formal investigations are needed to elucidate the impact of stimulus choices. Future research should examine how children in different age groups and ethnic/racial contexts comprehend and perceive different types of stimuli, and how those different perceptions in turn impact children's responses to measures of ethnic and racial attitudes, stereotypes, and discrimination.

Because using verbal descriptors and visual exemplars are the most popular ways of representing target groups, researchers should prioritize clarifying how children comprehend and respond to verbal and visual representations of ethnic and racial groups. For example, studies should test whether children understand verbal descriptors of target groups (e.g., labels like "Black people," "White people," and "immigrants") and what comes to their mind when hearing them. Relatedly, researchers should conduct new empirical studies that examine how verbally labeling ethnic/racial groups (vs. only using visual exemplars of group members) impacts young children's responses in measures of attitudes, stereotypes, and discrimination. Investigating these topics is critical for addressing whether and how researchers should use labels for ethnicity/race in measures for young children. Such investigations would also help researchers design stimuli that minimize disconnects between how researchers want children to interpret the measure and how children actually interpret the measure.

Furthermore, future research should address how specific characteristics of visual stimuli influence children's perception of and responses to measures. For example, it is important to examine which visual features of targets (e.g., skin color and hair texture) children attend to when evaluating them. Research by Dunham et al. (2015) showed that U.S. 4- to 9-year-old children attended to skin

color when asked to identify the race of photographed faces, whereas U.S. adults attended to skin color and other facial features (e.g., fullness of lips) in the same task. A recent study also suggests that children attend to hair texture when making judgments about people in photographs (Lei et al., 2024). Thus, children and adults may base their judgments and evaluations of targets on different types of visual information.

Accumulating evidence on children's perceptions of visual stimuli could guide investigations focused on how specific characteristics of visual stimuli impact children's responses to the types of measures that we reviewed. Researchers could then conduct new empirical studies to systematically test how children's responses change when stimulus characteristics change. For example, a study could examine whether preschoolers show similar levels of biased selection in the *interaction partner selection task* when the contrasting target groups are represented using drawings of individuals with similar vs. dissimilar skin tones. Studies in this direction could help us, for example, develop measures that are sensitive enough to capture very early racial attitudes, identify measures that strongly predict children's discriminatory behavior in real-world contexts, or uncover characteristics of stimuli that lead children to encode ethnicity or race information at different rates.

Conduct Studies Focused on Psychometric Properties of Measures

In Q4, we offered a landscape of evidence on the test–retest reliability, predictive validity, and responsiveness of popular measures for capturing children's ethnic and racial attitudes, stereotypes, and discrimination. We found that available evidence is sparse, and more evidence is needed to gain a better understanding of the psychometric properties of measures that our field uses very often. Moreover, as mentioned in the introduction of Q4, the wide variability in specific measure designs and study setup makes it difficult to summarize statistics across existing uses of the same/similar types of measures and directly compare different scales/tasks based on any given psychometric property. Rae and Olson (2018) have made a similar point, arguing that it is impossible to examine psychometric properties of measures across multiple studies when the studies differ in many ways, including age of participants, exact number of trials, time interval between measurements, and so on. Therefore, future research programs focused on producing psychometric evidence of these common measures are crucial for informing the field on which ones are reliable and best fit for different purposes. New empirical studies may present different tasks to the same sample of children, and test each of the tasks on its test–retest reliability, predictive validity, and responsiveness to an intervention. By testing various measures under the same condition (e.g., same sample, same metric of real-world behaviors, same intervention, and same time interval between repeated measurement), we can directly compare the psychometric properties between those measures (for limited but excellent examples, see Rae & Olson, 2018; Williams & Steele, 2016).

Given that many researchers in our field are interested in making inferences about children's real-world behaviors, the fact that there is limited evidence for the predictive validity of commonly used measures of children's ethnic and racial attitudes, stereotypes, and discrimination is especially notable. As demonstrated in Q4 Results, our insights into the relations between commonly used scales/tasks and real-world behaviors were based almost exclusively on studies using friendship nomination measures alongside commonly used scales/tasks. However, children's real-world ethnic and racial behaviors are not limited to picking friends; for example, they interact with other people in prosocial or antisocial ways in their school or other community settings, and they choose with whom to spend time. Future research should assess whether laboratory measures of children's ethnic and racial attitudes, stereotypes, and discrimination predict children's behaviors in the real world as measured through parent or teacher reports, observations of children in naturalistic settings such as the playground (see Martin & Fabes, 2001), or automated sensing technologies in classrooms (see Elbaum et al., 2024).

Increase Consistency, Transparency, and Details in Reports

The last issue we list as a priority for our field's future efforts centers on improving reporting practices. Specifically, we make three recommendations that we believe will promote better communication about measures as a field and help address some of the issues that we have already discussed.

First, when researchers adapt existing measures, they should clearly describe and justify adjustments they made to the original measures. During our coding process, we observed that researchers often do not clarify how or why they made changes when using existing measures, making the exact design and items of each measure instance unclear (we briefly referred to this clarity issue in the discussion following Q1). To increase transparency in replications and adaptations of existing measures, researchers should correctly reference the original source and clearly describe and justify any changes to the original measure. Additionally, for ease of referring to existing measures correctly and consistently across the field, researchers should consider naming any new measures (i.e., with novel tasks or items) that they create. In addition to helping the field understand the methodological similarities and variations between measures, these reporting practices could help researchers evaluate whether results across similar measures can be compared to each other.

Second, researchers should justify their measure choice and articulate what construct or phenomena the measures are meant to capture. Such reporting practices would not only ensure that children's responses are interpreted in ways that the authors intended, but they would also elucidate how researchers are thinking conceptually about their measures. As we discussed above, understanding the conceptualization of measures is an important step toward clarifying the conceptual and statistical relations between different types of measures. In turn, clarifying those relations is important for understanding how different types of measures can be used for different or similar purposes, or whether results from different measures are comparable to each other. Thus, better articulation of thought processes about measurement across the field may prompt more careful considerations of measure-construct relations and a

better understanding of measure-to-measure relations.

Third, researchers should make details of measure designs available. While searching for relevant measures and coding them, we often faced challenges extracting key information about measures from articles. We frequently had to track down referenced articles or contact article authors (see Supplemental Materials for more information) to obtain scale items or other information needed to code the measures (see Table S1 for our coding scheme). Moreover, as mentioned in Q2, we could not code for specific characteristics of visual stimuli used in measures (e.g., the gender and age of depicted individuals) as we had hoped, because those details were often not reported. The fact that we could not easily access these types of key information hints that other researchers may also find it difficult to evaluate and replicate existing tools, hindering the field's abilities to replicate studies and accumulate evidence of how existing measures perform and to compare between results.

Based on our observations, we suggest that authors make all scale/task items, participant instructions, and visual stimuli available to readers. If only a subset of them can appear in the main text of an article, authors should make them accessible in an appendix, supplemental materials, or on an online platform such as Databrary or the Open Science Framework (OSF), and clearly direct readers to those locations in the main text. If full items, instructions, and stimuli cannot be made public (e.g., because they are copyrighted), authors should be transparent about how readers can obtain them or find out more about them if they wish.

Because visual stimuli (e.g., photographs and drawings) are very frequently used but detailed descriptions are often lacking from reports (as mentioned in Q2), we make further recommendations. Descriptions of visual stimuli should include detailed characteristics of the targets (e.g., age, gender, and facial expressions), how the stimuli were created or obtained, and how they were validated. To provide a model, Mandalaywala et al. (2021) stated that their visual stimuli were obtained from the Child Affective Facial Expression (CAFE) dataset (LoBue, 2014; LoBue & Thrasher, 2015), described the age, gender, and facial expressions of depicted targets, and directed readers to a spreadsheet on OSF that lists the ID numbers of the photographs they took from the CAFE dataset. Other models by Ghavami et al. (2020) and Renno and Shutts (2015) show how authors can describe the way in which their stimuli were validated (e.g., rated for perceived attractiveness and race/ethnic stereotypicality by pilot participants prior to the reported studies). These types of details allow researchers to closely replicate measures and help the field gain a better sense of the specific materials used in measures, which can have implications for study conclusions, as we discussed above.

Conclusion

Together, the current review provides a landscape of modern measures for capturing children's ethnic and racial attitudes, stereotypes, and discrimination; offers initial insights about the characteristics of those measures; provides an evidence gap map to guide focused evaluation of psychometric properties of those measures; and makes recommendations for future efforts in the field. We argue that evaluating measures is a fertile avenue for future research. Moreover, focusing on measurement is an imperative step for accumulating knowledge and advancing our science of intergroup issues. Our field can collectively gain a deeper understanding of the development of social group phenomena by accelerating investigations and discussions surrounding measurement.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Amelia Dow, Brianna Epping, Amara Schiffman, Addie Westman, Mia Williams, and Nihan Zhou for their assistance on the article filtering process. We also thank Daniel Bolt for offering expert insights on psychometrics, James Pustejovsky for advising us on synthesizing information across publications, and Ashley Jordan for providing feedback on a previous version of the manuscript. Preparation of this article was supported by an NIH grant (R01HD106970) to K. Shutts, an NSF grant (1941756) to K. Shutts, and a core grant to the Waisman Center from the National Institute of Child Health and Human Development (P50HD105353). Preparation of this article was also supported by the Institute of Education Sciences, U.S. Department of Education, through Award #R305B200026 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dr.2025.101189.

References

Aboud, F. E., Tredoux, C., Tropp, L. R., Brown, C. S., Niens, U., & Noor, N. M. (2012). Interventions to reduce prejudice and enhance inclusion and respect for ethnic differences in early childhood: A systematic review. *Developmental Review, 32*(4), 307–336. https://doi.org/10.1016/j.dr.2012.05.001

Aldana, A., Bañales, J., & Richards-Schuster, K. (2019). Youth anti-racist engagement: Conceptualization, development, and validation of an Anti-Racism Action Scale. *Adolescent Research Review, 4*(4), 369–381. https://doi.org/10.1007/s40894-019-00113-1

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. *European Psychologist, 22*(4), 207–218. https://doi.org/10.1027/1016-9040/a000298

American Educational Research Association (Ed.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Aronson, K. M., Stefanile, C., Matera, C., Nerini, A., Grisolaghi, J., Romani, G., Massai, F., Antonelli, P., Ferraresi, L., & Brown, R. (2016). Telling tales in school: Extended contact interventions in the classroom. *Journal of Applied Social Psychology, 46*(4), 229–241. https://doi.org/10.1111/jasp.12358

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science, 25*(9), 1804–1815. https://doi.org/10.1177/0956797614543801

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes. Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science, 17*(1), 53–58. https://doi.org/10.1111/j.1467-9280.2005.01664.x

Bayram Özdemir, S., Giles, C., & Özdemir, M. (2021). Why do immigrant and swedish adolescents engage in ethnic victimization? Common and distinct underlying factors. *Journal of Youth and Adolescence.*. https://doi.org/10.1007/s10964-021-01485-1

Bayram Özdemir, S., Özdemir, M., & Stattin, H. (2016). What makes youth harass their immigrant peers? Understanding the risk factors. *The Journal of Early Adolescence, 36*(5), 601–624. https://doi.org/10.1177/0272431615574887

Benatov, J., Berger, R., & Tadmor, C. T. (2021). Gaming for peace: Virtual contact through cooperative video gaming increases children's intergroup tolerance in the context of the Israeli–Palestinian conflict. *Journal of Experimental Social Psychology, 92*. https://doi.org/10.1016/j.jesp.2020.104065

Berger, R., Brenick, A., Lawrence, S. E., Coco, L., & Abu-Raiya, H. (2018). Comparing the effectiveness and durability of contact- and skills-based prejudice reduction approaches. *Journal of Applied Developmental Psychology, 59*, 46–53. https://doi.org/10.1016/j.appdev.2018.04.002

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*(3), 242–261. https://doi.org/10.1207/S15327957PSPR0603_8

Bogardus, E. S. (1933). A social distance scale. *Sociology & Social Research, 17*, 265–271.

Brenick, A., Killen, M., Lee-Kim, J., Fox, N., Leavitt, L., Raviv, A., Masalha, S., Murra, F., & Al-Smadi, Y. (2010). Social understanding in young Israeli-Jewish, Israeli-Palestinian, Palestinian, and Jordanian children: Moral judgments and stereotypes. *Early Education and Development, 21*(6), 886–911. https://doi.org/10.1080/10409280903236598

Byrd, C. M. (2012). The measurement of racial/ethnic identity in children: A critical review. *Journal of Black Psychology, 38*(1), 3–31. https://doi.org/10.1177/0095798410397544

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Carraro, L., & Castelli, L. (2015). On the generality of children's racial attitudes across target groups. *Psicologia Sociale, 10*(1), 71–80. psyh.

Casey-Cannon, S. L., Coleman, H. L. K., Knudtson, L. F., & Velazquez, C. C. (2011). Three ethnic and racial identity measures: Concurrent and divergent validity for diverse adolescents. *Identity: An International Journal of Theory and Research, 11*(1), 64–91. https://doi.org/10.1080/15283488.2011.540739

Central Bureau of Statistics (2022, December 29). *Population of Israel on the Eve of 2023*. Retrieved October 8, 2024 from https://www.cbs.gov.il/he/mediarelease/DocLib/2022/426/11_22_426e.pdf.

Chen, E. E., Corriveau, K. H., Lai, V. K. W., Poon, S. L., & Gaither, S. E. (2018). Learning and socializing preferences in Hong Kong Chinese children. *Child Development, 89*(6), 2109–2117. https://doi.org/10.1111/cdev.13083

Clark, K. D., Yovanoff, P., & Tate, C. U. (2017). Development and psychometric validation of a child Racial Attitudes Index (RAI). *Behavior Research Methods, 49*(6), 2044–2060. https://doi.org/10.3758/s13428-016-0841-y

Cokley, K. (2007). Critical issues in the measurement of ethnic and racial identity: A referendum on the state of the field. *Journal of Counseling Psychology, 54*, 224–234. https://doi.org/10.1037/0022-0167.54.3.224

Constantin, A. A., & Cuadrado, I. (2021). The effect of imagined contact valence on adolescents' and early adults' stereotypes, emotions, and behavioral intentions toward ethnic groups. *Social Development, 30*(3), 697–712. https://doi.org/10.1111/sode.12510

Cooper, H. (2017). *Research synthesis and meta-analysis*. SAGE Publications, Inc, 10.4135/9781071878644.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Cooper, S. M., Hurd, N. M., & Loyd, A. B. (2022). Advancing scholarship on anti-racism within developmental science: Reflections on the special section and recommendations for future research. *Child Development, 93*(3), 619–632. https://doi.org/10.1111/cdev.13783

Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes and discrimination. *The SAGE Handbook of Prejudice, Stereotyping and Discrimination, 45–62*. https://doi.org/10.4135/9781446200919.n3

deMayo, B., & Olson, K. R. (2023). Comparing methods of social preference assessment in childhood. *Social Development, e12736*. https://doi.org/10.1111/sode.12736

Dhont, K., & Van Hiel, A. (2012). Intergroup contact buffers against the intergenerational transmission of authoritarianism and racial prejudice. *Journal of Research in Personality, 46*(2), 231–234. https://doi.org/10.1016/j.jrp.2011.12.008

Dore, R. A. (2022). The effect of character similarity on children's learning from fictional stories: The roles of race and gender. *Journal of Experimental Child Psychology, 214*, Article 105310. https://doi.org/10.1016/j.jecp.2021.105310

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. *Handbook of Social Psychology, 1084–1121*. https://doi.org/10.1002/9780470561119.socpsy002029

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. *The SAGE Handbook of Prejudice, Stereotyping and Discrimination, 3–28*. https://doi.org/10.4135/9781446200919.n1

Doyle, A. B., & Aboud, F. E. (1995). A longitudinal study of white children's racial prejudice as a social-cognitive development. *Merrill-Palmer Quarterly (1960), 41*(2), 209–228.

Dunham, Y. (2018). Mere Membership. *Trends in Cognitive Sciences, 22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Dunham, Y., Stepanova, E. V., Dotsch, R., & Todorov, A. (2015). The development of race-based perceptual categorization: Skin color dominates early category judgments. *Developmental Science, 18*(3), 469–483. https://doi.org/10.1111/desc.12228

Elbaum, B., Perry, L. K., & Messinger, D. S. (2024). Investigating children's interactions in preschool classrooms: An overview of research using automated sensing technologies. *Early Childhood Research Quarterly, 66*, 147–156. https://doi.org/10.1016/j.ecresq.2023.10.005

Elenbaas, L., & Killen, M. (2017). Children's perceptions of social resource inequality. *Journal of Applied Developmental Psychology, 48*, 49–58. https://doi.org/10.1016/j.appdev.2016.11.006

Elenbaas, L., Rizzo, M. T., Cooley, S., & Killen, M. (2016). Rectifying social inequalities in a resource allocation task. *Cognition, 155*, 176–187. https://doi.org/10.1016/j.cognition.2016.07.002

Fitzgerald, H. E., Johnson, D. J., Baolian, D., & Francisco, Q. (2019). *Handbook of children and prejudice*. *Handbook of Children and Prejudice*. https://doi.org/10.1007/978-3-030-12228-7

Fryberg, S. A., & Eason, A. E. (2017). Making the invisible visible: Acts of commission and omission. *Current Directions in Psychological Science, 26*(6), 554–559. https://doi.org/10.1177/0963721417720959

Gaither, S. E. (2015). "Mixed" results: Multiracial research and identity explorations. *Current Directions in Psychological Science, 24*(2), 114–119. https://doi.org/10.1177/0963721414558115

Galvan, M. J., & Payne, B. K. (2024). Implicit bias as a cognitive manifestation of systemic racism. *Daedalus, 153*(1), 106–122. https://doi.org/10.1162/daed_a_02051

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15*(3), 485–499. https://doi.org/10.1016/j.concog.2005.11.007

Ghavami, N., Kogachi, K., & Graham, S. (2020). How Racial/ethnic diversity in urban schools shapes intergroup relations and well-being: Unpacking intersectionality and multiple identities perspectives. *Frontiers in Psychology*, 11. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.503846.
Ghavami, N., & Peplau, L. A. (2013). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly, 37*(1), 113–127. https://doi.org/10.1177/0361684312464203
Gómez, Á., Dovidio, J. F., Gaertner, S. L., Fernández, S., & Vázquez, A. (2013). Responses to endorsement of commonality by ingroup and outgroup members: The roles of group representation and threat. *Personality and Social Psychology Bulletin, 39*(4), 419–431. https://doi.org/10.1177/0146167213475366
Gonzalez, A. M., Steele, J. R., & Baron, A. S. (2017). Reducing children's implicit racial bias through exposure to positive out-group exemplars. *Child Development, 88*(1), 123–130. https://doi.org/10.1111/cdev.12582
Graham, S., & Echols, L. (2018). Race and ethnicity in peer relations research. In W. M. Bukowski, B. Laursen, & K. H. Rubin (Eds.), *Handbook of peer interactions, relationships, and groups* (2nd ed., pp. 590–616). New York, NY: The Guilford Press.
Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197
Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17. https://doi.org/10.1037/a0015575
Guerrero, S., Enesco, I., Lago, O., & Rodríguez, P. (2010). Preschool children's understanding of racial cues in drawings and photographs. *Cognitive Development, 25*(1), 79–89. https://doi.org/10.1016/j.cogdev.2009.07.001
Hazelbaker, T., Brown, C. S., Nenadal, L., & Mistry, R. S. (2022). Fostering anti-racism in white children and youth: Development within contexts. *American Psychologist*. https://doi.org/10.1037/amp0000948
Hazelbaker, T., & Mistry, R. S. (2022). Negotiating Whiteness: Exploring White elementary school-age children's racial identity development. *Social Development, 31*(4), 1280–1295. https://doi.org/10.1111/sode.12602
Heberle, A. E., Rapa, L. J., & Farago, F. (2020). Critical consciousness in children and adolescents: A systematic review, critical assessment, and recommendations for future research: Psychological Bulletin. *Psychological Bulletin, 146*(6), 525–551. https://doi.org/10.1037/bul0000230
Helms, J. E. (2007). Some better practices for measuring racial and ethnic identity constructs. *Journal of Counseling Psychology, 54*(3), 235–246. https://doi.org/10.1037/0022-0167.54.3.235
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X
Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613
Hoyle, H., Harris, M. J., & Judd, C. M. (2002). *Research methods in social science* (7th Edition). Fort Worth, TX: Wadsworth.
Hughes, J. M., & Bigler, R. S. (2007, March). Development and validation of new measures of racial stereotyping and prejudice. Poster presented at the biennial meeting of the Society for Research in Child Development, Boston.
Instituto Brasileiro de Geografia e Estatística. (2022). *Destaques*. Censo 2022. Retrieved October 1, 2024 from https://censo2022.ibge.gov.br/panorama/.
Jizhe, N. (2021, May 11). *Main Data of the Seventh National Population Census*. National Bureau of Statistics of China. Retrieved September 25, 2024 from https://www.stats.gov.cn/english/PressRelease/202105/t20210510_1817185.html.
Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social intervention*. Cambridge University Press.
Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. *Journal of Personality and Social Psychology, 81*(5), 774–788. https://doi.org/10.1037/0022-3514.81.5.774
Karras, J. E., Niwa, E. Y., Adesina, F., & Ruck, M. D. (2021). Confronting whiteness: Conceptual, contextual, and methodological considerations for advancing ethnic-racial socialization research to illuminate white identity development. *Journal of Social Issues*. https://doi.org/10.1111/josi.12485
Kelley, T. L. (1927). *Interpretation of educational measurements. World*.
Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364
Kinzler, K. D., & Spelke, E. S. (2011). Do infants show social preferences for people differing in race? *Cognition, 119*(1), 1–9. https://doi.org/10.1016/j.cognition.2010.10.019
Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*(5), 569. https://doi.org/10.1037/amp0000364
Lai, C. K., & Wilson, M. E. (2020). Measuring implicit intergroup biases. *Social and Personality Psychology Compass*, 1–16. https://doi.org/10.1111/spc3.12573. September 2019.
Lei, R. F., Cohen, A. J., Wong, P., & Hudson, S.-K.-T.-J. (2024). Investigating hair cues as a mechanism underlying Black women's intersectional invisibility. *Developmental Psychology, 60*(10), 1928–1934. https://doi.org/10.1037/dev0001729
Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science, 5*(1), 69. https://doi.org/10.1186/1748-5908-5-69
Liebkind, K., Mähönen, T. A., Solares, E., Solheim, E., & Jasinskaja-Lahti, I. (2014). Prejudice-reduction in culturally mixed classrooms: The development and assessment of a theory-driven intervention among majority and minority youth in Finland. *Journal of Community & Applied Social Psychology, 24*(4), 325–339. https://doi.org/10.1002/casp.2168
LoBue, V. (2014). The Child Affective Facial Expression (CAFE) set. *Databrary*. 10.17910/B7301K.
LoBue, V., & Thrasher, C. (2015). The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.01532
Mähönen, T. A., Jasinskaja-Lahti, I., Liebkind, K., & Finell, E. (2010). Perceived normative pressure and majority adolescents' implicit and explicit attitudes towards immigrants. *International Journal of Psychology, 45*(3), 182–189. https://doi.org/10.1080/00207590903487412
Mandalaywala, T. M., Benitez, J., Sagar, K., & Rhodes, M. (2021). Why do children show racial biases in their resource allocation decisions? *Journal of Experimental Child Psychology, 211*. https://doi.org/10.1016/j.jecp.2021.105224
Marcelo, A. K., & Yates, T. M. (2019). Young children's ethnic–racial identity moderates the impact of early discrimination experiences on child behavior problems. *Cultural Diversity and Ethnic Minority Psychology, 25*(2), 253–265. https://doi.org/10.1037/cdp0000220
Martin, C. L., & Fabes, R. A. (2001). *The stability and consequences of young children's same-sex peer interactions*. *Developmental Psychology, 37*(3), 431. https://doi.org/10.1037/0012-1649.37.3.431
Mesman, J., de Bruijn, Y., van Veen, D., Pektas, F., & Emmen, R. A. G. (2022). Maternal color-consciousness is related to more positive and less negative attitudes toward ethnic-racial outgroups in children in White Dutch families. *Child Development, 93*(3), 668–680. https://doi.org/10.1111/cdev.13784
Miklikowska, M. (2017). Development of anti-immigrant attitudes in adolescence: The role of parents, peers, intergroup friendships, and empathy. *British Journal of Psychology, 108*(3), 626–648. https://doi.org/10.1111/bjop.12236
Moriguchi, Y. (2023). Beyond bias to Western participants, authors, and editors in developmental science. *Infant and Child Development, 32*. https://doi.org/10.1002/icd.2430
National Bureau of Statistics. (2021, May 11). *Communiqué of the Seventh National Population Census (No. 8)*. National Bureau of Statistics of China. Retrieved September 25, 2024 from https://www.stats.gov.cn/english/PressRelease/202105/t20210510_1817193.html.
Nielsen, 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. https://doi.org/10.1016/j.jecp.2017.04.017
Nishina, A., & Witkow, M. R. (2020). Why developmental researchers should care about biracial, multiracial, and multiethnic youth. *Child Development Perspectives, 14*(1), 21–27. https://doi.org/10.1111/cdep.12350

Niwa, E. Y., Boxer, P., Dubow, E., Huesmann, L. R., Shikaki, K., Landau, S., & Gvirsman, S. D. (2016). Growing up amid ethno-political conflict: Aggression and emotional desensitization promote hostility to ethnic outgroups. *Child Development, 87*(5), 1479–1492. https://doi.org/10.1111/cdev.12599
Office of Homeland Security Statistics (2023, November 17). Geographic Regions. https://www.dhs.gov/ohss/about-data/geographic-regions.
Olson, K. R., Dweck, C. S., Spelke, E. S., & Banaji, M. R. (2011). Children's responses to group-based inequalities: Perpetuation and rectification. *Social Cognition, 29*(3), 270–287. https://doi.org/10.1521/soco.2011.29.3.270
Pachter, L. M., Bernstein, B. A., Szalacha, L. A., & Coll, C. G. (2010). Perceived racism and discrimination in children and youths: An exploratory study. *Health & Social Work, 35*(1), 61–69. https://doi.org/10.1093/hsw/35.1.61
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery, 88*, Article 105906. https://doi.org/10.1016/j.ijsu.2021.105906
Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568
Phinney, J. S. (1992). *The Multigroup Ethnic Identity Measure*.
Phinney, J. S., & Ong, A. D. (2007). Conceptualization and measurement of ethnic identity: Current status and future directions. *Journal of Counseling Psychology, 54*(3), 271–281. https://doi.org/10.1037/0022-0167.54.3.271
Polanin, J. R., Zhang, Q., Taylor, J. A., Williams, R. T., Joshi, M., & Burr, L. (2023). Evidence gap maps in education research. *Journal of Research on Educational Effectiveness, 16*(3), 532–552. https://doi.org/10.1080/19345747.2022.2139312
Polit, D. F. (2014). Getting serious about test–retest reliability: A critique of retest research and some recommendations. *Quality of Life Research, 23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9
Poulin, F., & Chan, A. (2010). Friendship stability and change in childhood and adolescence. *Developmental Review, 30*(3), 257–272. https://doi.org/10.1016/j.dr.2009.01.001
Qian, M. K., Heyman, G. D., Quinn, P. C., Messi, F. A., Fu, G., & Lee, K. (2016). Implicit racial biases in preschool children and adults from Asia and Africa. *Child Development, 87*(1), 285–296. https://doi.org/10.1111/cdev.12442
Raabe, T., & Beelmann, A. (2011). Development of ethnic, racial, and national prejudice in childhood and adolescence: A multinational meta-analysis of age differences. *Child Development, 82*(6), 1715–1737. https://doi.org/10.1111/j.1467-8624.2011.01668.x
Rae, J. R., & Olson, K. R. (2018). Test-retest reliability and predictive validity of the implicit association test in children. *Developmental Psychology, 54*(2), 308–330. https://doi.org/10.1037/dev0000437
Rastogi, R., & Juvonen, J. (2019). Interminority friendships and intergroup attitudes across middle school: Quantity and stability of Black-Latino ties. *Journal of Youth and Adolescence, 48*(8), 1619–1630. https://doi.org/10.1007/s10964-019-01044-9
Reijerse, A., Vanbeselaere, N., Duriez, B., & Fichera, G. (2015). Accepting immigrants as fellow citizens: Citizenship representations in relation to migration policy preferences. *Ethnic and Racial Studies, 38*(5), 700–717. https://doi.org/10.1080/01419870.2014.916812
Renno, M. P., & Shutts, K. (2015). Children's social category-based giving and its correlates: Expectations and preferences. *Developmental Psychology, 51*(4), 533–543. https://doi.org/10.1037/a0038819
Reyes-Jaquez, B., Escala, M. J., & Bigler, R. S. (2021). When multiracial individuals are the numerical majority: Children's racial attitudes in the Dominican Republic. *Developmental Psychology, 57*(5), 662–677. https://doi.org/10.1037/dev0001052.supp
Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science, 15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709
Roberts, S. O., & Mortenson, E. (2022). Challenging the white = neutral framework in psychology. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916221077117
Rowley, S. J., & Camacho, T. C. (2015). Increasing diversity in cognitive developmental research: Issues and solutions. *Journal of Cognition and Development, 16*, 683–692. https://doi.org/10.1080/15248372.2014.976224
Ruck, M. D., Park, H., Killen, M., & Crystal, D. S. (2011). Intergroup contact and evaluations of race-based exclusion in urban minority children and adolescents. *Journal of Youth and Adolescence, 40*(6), 633–643. https://doi.org/10.1007/s10964-010-9600-z
Rutland, A., Cameron, L., Milne, A., & McGeorge, P. (2005). Social norms and self-presentation: Children's implicit and explicit intergroup attitudes. *Child Development, 76*(2), 451–466. https://doi.org/10.1111/j.1467-8624.2005.00856.x
Satterthwaite-Freiman, M., Sladek, M. R., Wantchekon, K. A., Rivas-Drake, D., & Umaña-Taylor, A. J. (2023). Examining ethnic-racial identity negative affect, centrality, and intergroup contact attitudes among white adolescents. *Journal of Youth and Adolescence, 52*(1), 61–75. https://doi.org/10.1007/s10964-022-01680-8
Schuitema, J., & Veugelers, W. (2011). Multicultural contacts in education: A case study of an exchange project between different ethnic groups. *Educational Studies, 37*(1), 101–114. https://doi.org/10.1080/03055691003729252
Schwartz, S. J., Syed, M., Yip, T., Knight, G. P., Umaña-Taylor, A. J., Rivas-Drake, D., & Lee, R. M. (2014). Methodological issues in ethnic and racial identity research with ethnic minority populations: Theoretical precision, measurement issues, and research designs. *Child Development, 85*(1), 58–76. https://doi.org/10.1111/cdev.12201
Sierksma, J., Brey, E., & Shutts, K. (2022). Racial stereotype application in 4-to-8-year-old white american children: Emergence and specificity. *Journal of Cognition and Development, 23*(5), 660–685. https://doi.org/10.1080/15248372.2022.2090945
Signorella, M. L., Bigler, R. S., & Liben, L. S. (1993). Developmental differences in children's gender schemata about others: A meta-analytic review. *Developmental Review, 13*(2), 147–183. https://doi.org/10.1006/drev.1993.1007
Skinner, A. L., & Meltzoff, A. N. (2018). Childhood experiences and intergroup biases among children. *Social Issues and Policy Review, 13*(1), 211–240. https://doi.org/10.1111/sipr.12054
Snilstveit, B., Vojtkova, M., Bhavsar, A., & Gaarder, M. (2013). Evidence gap maps-a tool for promoting evidence-informed policy and prioritizing future research. *World Bank Policy Research Working Paper, 6725*.
Stathi, S., Cameron, L., Hartley, B., & Bradford, S. (2014). Imagined contact as a prejudice-reduction intervention in schools: The underlying role of similarity and attitudes. *Journal of Applied Social Psychology, 44*(8), 536–546. https://doi.org/10.1111/jasp.12245
Statistics Netherlands. (2022, May 31). *Population; sex, age, generation and migration background, 1 Jan; 1996-2022*. StatLine. Retrieved September 25, 2024, from https://opendata.cbs.nl/statline/#/CBS/en/dataset/37325eng/table?ts=1726841667989.
Statistics South Africa (2022). *Population*. Statistics South Africa. Retrieved October 1, 2024 from https://census.statssa.gov.za/#/statsbytheme.
Stengelin, R., Haun, D. B. M., & Kanngiesser, P. (2023). Simulating peers: Can puppets simulate peer interactions in studies on children's socio-cognitive development? *Child Development, 94*(5), 1117–1135. https://doi.org/10.1111/cdev.13913
Taylor, L. K., & McKeown, S. (2021). Adolescent outgroup helping, collective action, and political activism in a setting of protracted conflict. *International Journal of Intercultural Relations, 85*, 37–46. https://doi.org/10.1016/j.ijintrel.2021.09.001
Taylor, L. K., Merrilees, C. E., Goeke-Morey, M. C., Shirlow, P., Cairns, E., & Cummings, E. M. (2014). Political violence and adolescent out-group attitudes and prosocial behaviors: Implications for positive inter-group relations. *Social Development, 23*(4), 840–859. https://doi.org/10.1111/sode.12074
Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Teachers College, Columbia University.
Tredoux, C. G., Noor, N. M., & de Paulo, L. (2009). Quantitative measures of respect and social inclusion in children: Overview and recommendations. *Effective Education, 1*(2), 169–186. https://doi.org/10.1080/19415530903522568
Trent, M., Dooley, D. G., Dougé, J., Trent, M. E., Cavanaugh, R. M., Lacroix, A. E., Fanburg, J., Rahmandar, M. H., Hornberger, L. L., Schneider, M. B., Yen, S., Chilton, L. A., Green, A. E., Dilley, K. J., Gutierrez, J. R., Duffee, J. H., Keane, V. A., Krugman, S. D., McKelvey, C. D., & Wallace, S. B. (2019). The impact of racism on child and adolescent health. *American Academiy of Pediatrics, 144*(2). https://doi.org/10.1542/peds.2019-1765

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

Tropp, L. R., O'Brien, T. C., Gutierrez, R. G., Valdenegro, D., Migacheva, K., Tezanos-Pinto, P., Berger, C., & Cayul, O. (2016). How school norms, peer norms, and discrimination predict interethnic experiences among ethnic minority and majority youth. *Child Development, 87*(5), 1436–1451. https://doi.org/10.1111/cdev.12608

Tropp, L. R., O'Brien, T. C., & Migacheva, K. (2014). How peer norms of inclusion and exclusion predict children's interest in cross-ethnic friendships. *Journal of Social Issues, 70*(1), 151–166. https://doi.org/10.1111/josi.12052

Umaña-Taylor, A. J. (2016). A post-racial society in which ethnic-racial discrimination still exists and has significant consequences for youths' adjustment. *Current Directions in Psychological Science, 25*(2), 111–118. https://doi.org/10.1177/0963721415627858

Umaña-Taylor, A. J., Quintana, S. M., Lee, R. M., Cross, W. E., Jr., Rivas-Drake, D., Schwartz, S. J., Syed, M., Yip, T., & Seaton, E. (2014). Ethnic and racial identity during adolescence and into young adulthood: An integrated conceptualization. *Child Development, 85*(1), 21–39. https://doi.org/10.1111/cdev.12196

United Nations. (n.d.). *Standard country or area codes for statistical use (M49)*. United Nations Statistics Division. Retrieved March 12, 2023, from https://unstats.un.org/unsd/methodology/m49/.

United Nations (2022). World Population Prospects 2022. https://population.un.org/wpp/.

United States Census Bureau. (2020, April 1). *QuickFacts*. United States Census Bureau. Retrieved September 25, 2024, from https://www.census.gov/quickfacts/fact/table/US/POP010220.

van Zalk, M. H. W., M, K., van Zalk, N., & Stattin, H. (2013). Xenophobia and Tolerance Toward Immigrants in Adolescence: Cross-Influence Processes Within Friendships. *Journal of Abnormal Child Psychology*, *41*, 627–639.

Vezzali, L., Di Bernardo, G. A., Stathi, S., Visintin, E. P., & Hewstone, M. (2019). Using intercultural videos of direct contact to implement vicarious contact: A school-based intervention that improves intergroup attitudes. *Group Processes & Intergroup Relations, 22*(7), 1059–1076. https://doi.org/10.1177/1368430218809885

Vezzali, L., Giovannini, D., & Capozza, D. (2012). Social antecedents of children's implicit prejudice: Direct contact, extended contact, explicit and implicit teachers' prejudice. *European Journal of Developmental Psychology, 9*(5), 569–581. https://doi.org/10.1080/17405629.2011.631298

Vittrup, B., & Holden, G. W. (2011). Exploring the impact of educational television and parent-child discussions on children's racial attitudes. *Analyses of Social Issues and Public Policy, 11*(1), 82–104. https://doi.org/10.1111/j.1530-2415.2010.01223.x

Watts, R. J., Diemer, M. A., & Voight, A. M. (2011). Critical consciousness: Current status and future directions. *New Directions for Child and Adolescent Development, 2011*(134), 43–57. https://doi.org/10.1002/cd.310

Wilcox, C., Sigelman, L., & Cook, E. (1989). Some like it hot: Individual differences in responses to group feeling thermometers. *Public Opinion Quarterly, 53*, 246–257. https://doi.org/10.1086/269505

Williams, A. D., Bigler, R. S., & Ramirez, M. C. (2023). Latinx children's race- and ethnicity-related identities, beliefs, and preferences. *Group Processes & Intergroup Relations, 26*(1), 120–139. https://doi.org/10.1177/13684302211050553

Williams, C. D., Umaña-Taylor, A. J., Updegraff, K. A., & Jahromi, L. B. (2021). Measuring 5-year-old Mexican-heritage children's ethnic-racial identity attitudes, centrality, and knowledge. *Journal of Applied Developmental Psychology, 75*, Article 101290. https://doi.org/10.1016/j.appdev.2021.101290

Williams, J. E., Best, D. L., Boswell, D. A., Mattson, L. A., & Graves, D. J. (1975). Preschool racial attitude measure II. *Educational and Psychological Measurement, 35*(1), 3–18. https://doi.org/10.1177/001316447503500101

Williams, A., & Steele, J. R. (2016). The reliability of child-friendly race-attitude implicit association tests. *Frontiers in Psychology, 7*. https://ezproxy.library.wisc.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=psyh&AN=2016-52709-001&site=ehost-live&scope=site.

Williams, A., & Steele, J. R. (2019). Examining children's implicit racial attitudes using exemplar and category-based measures. *Child Development, 90*(3), e322–e338. https://doi.org/10.1111/cdev.12991

Wölfer, R., & Hewstone, M. (2017). Beyond the dyadic perspective: 10 Reasons for using social network analysis in intergroup contact research. *British Journal of Social Psychology, 56*(3), 609–617. https://doi.org/10.1111/bjso.12195

Yee, M. D., & Brown, R. (1992). Self-evaluations and intergroup attitudes in children aged three to nine. *Child Development, 63*(3), 619–629. https://doi.org/10.2307/1131350

Yu, C., Qian, M., Amemiya, J., Fu, G., Lee, K., & Heyman, G. D. (2022). Young children form generalized attitudes based on a single encounter with an outgroup member. *Developmental science, 25*(3), Article e13191. https://doi.org/10.1111/desc.13191