Socio-emotional learning (SEL) at scale: Developing measurement tools for successful interventions

Catarina Vales, PhD. CMU Psychology Department.
Israa Belbaita. International Relations and Politics, Statistics.
Catherine Dundon. Technical Writing, HCI.
Molly Niehaus, BS. CMU Psychology Department.
CMU Cognitive and Social Development Lab (CSDL)

Problem:

Many decades of research in developmental science show that young children hold ideas and beliefs about themselves and others based on characteristics associated with racial/ethnic groups (e.g., physical features, culture, religion, language; see Raabe & Beelman, 2011, Waxman, 2021 for reviews). This rapid development of racialized beliefs is in conflict with one critical goal of socio-emotional learning (SEL), which is to *foster children's positive attitudes towards themselves and others.*

At the same time, there is a *lack of validated tools* to assess children's beliefs about different social categories (e.g., gender, race, class, disability) that can be used at scale by researchers and educators (Byrd, 2011; Eddie et al., 2025; Fukuda et al., 2025). This lack of valid and reliable measurement tools makes it challenging to appropriately assess whether a specific SEL intervention or instructional approach is effective.

Opportunity:

We have a unique opportunity to fill this gap by combining the expertise of the CSDL research team and the TGDS teaching and student teams to develop and evaluate game-based assessment approaches that produce high-quality data. We will focus on racial/ethnic categories as a first case study. Developing these assessment tools will facilitate future intervention delivery and assessment at scale, by researchers and educators alike.

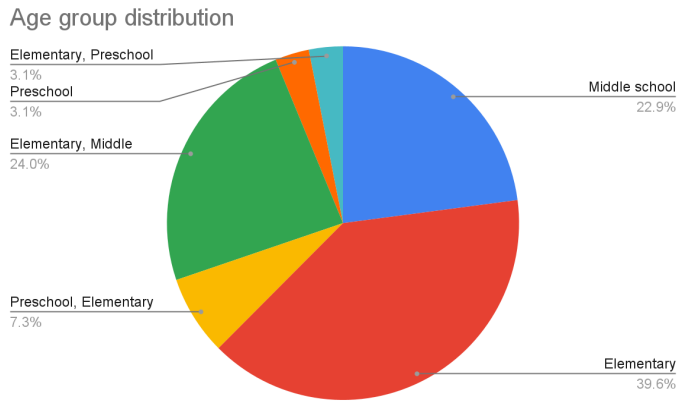
Approach:

The CSDL team conducted a scooping literature review and coding of existing assessments to evaluate racial/ethnic biases in young children, including documenting existing gaps (presented below). The goal for the Fall 2025 is to create multiple digital game prototypes that the CSDL team can then test with preschool-aged children in the Spring 2026.

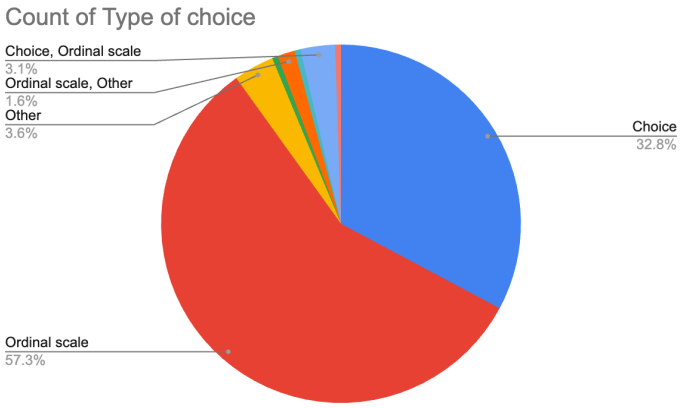
Literature review: Findings & Identified gaps

We reviewed a total of 193 measures (~80 papers) used to assess racial/ethnic bias in children and coded these measures along the dimensions described below. For some of the dimensions, only a subset of the measures were coded (we note the findings for which this is the case).

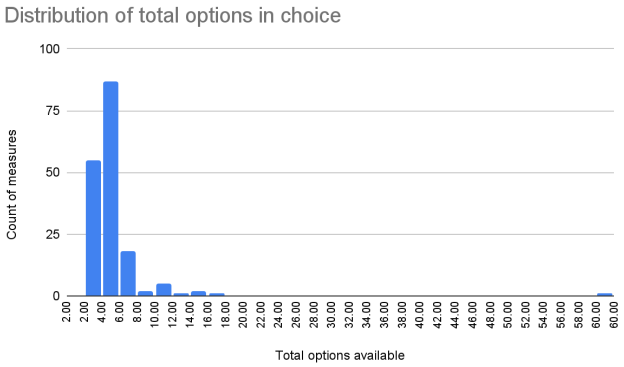
- The majority of the measures reviewed were used with children in the elementary and middle school ages. Less than 10% of the measures were used specifically with preschool-aged children. This is an important gap because prior work suggests that by the elementary school age there are already measurable biases (see Raabe & Beelman, 2011), meaning that the preschool years are an important period for measurement and intervention. Thus, developing better measures for *children 6 and under* is critical.

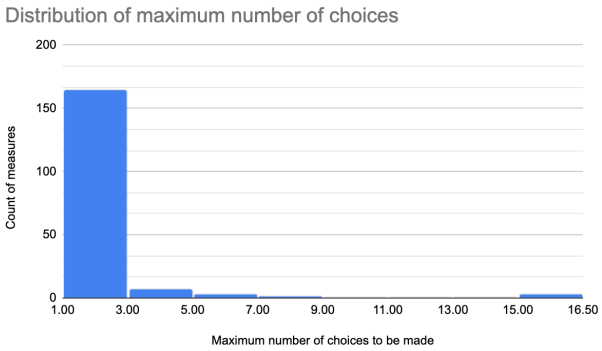


- The measures tended to show participants predetermined options to choose from. The most common approach included options that can be interpreted in an ordinal manner ("ordinal scale" i.e., a scale in which the order has inherent meaning – e.g., none to all; close to far; strongly dislike to strongly like) while the second most common approach included response options without an inherent order ("choice"; e.g., selecting from different people or from different names). Having predetermined response options often simplifies the data analysis process, but also only allows for analysis of those predetermined options. It would be helpful to develop more mixed methods approaches including age-appropriate open-ended questions that can still be used to conduct quantitative analyses.

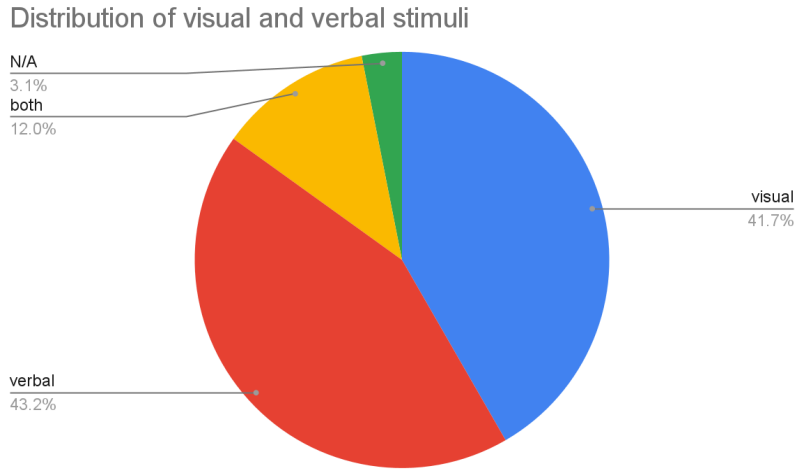


- Most measures included a small number of options to choose from in each question and asked participants to select 1-2 options in each question. This is likely a good idea with young children.





- Because these measures aim to understand what children know/think about racial/ethnic categories, they need to communicate information about racial/ethnic groups. This can be tricky for young children because many young children do not (yet) know the labels for racial/ethnic groups; from an ethical standpoint, researchers should also aim to prevent the introduction of stereotypes – therefore, teaching children about racial/ethnic categories for the sole purpose of assessing racial/ethnic biases would be unethical. On aggregate, information about racial/ethnic groups was communicated either visually (e.g., asking participants to select from an array depicting White and Black people, with the racial group communicated via visual phenotypical characteristics) or verbally (e.g., using labels for racial/ethnic groups; presenting a written or spoken story involving different racial/ethnic groups).



Importantly, when we examined only the data for the youngest children, we see a preponderance of visual stimuli (likely for the reasons mentioned above). <u>This will be an important methodological and ethical consideration for the purposes of this collaboration.</u>

- For visual stimuli depicting people, we also coded (for a subset of the measures) whether those depictions tended to be *realistic* (a real photo, real video, or photorealistic visualization of a person) or *cartoon-like* (a visual which is drawn, animated, or otherwise presented as a non-photorealistic person or caricature). We found that most visual depictions were realistic (59% only realistic, 66% realistic + other); a subset of the

measures used cartoon-like depictions (9% only cartoon; 30% cartoon + other). Realistic photos are more likely to allow respectful representation of visible racial/ethnic differences (e.g., hair texture, eye shape, religious markers, etc), while drawn stimuli can be more child-friendly and can also allow for more controlled stimuli. Balancing these needs will be an important methodological consideration for this collaboration.

- For a subset of the measures, we coded for the presence of the gamified aspects described below. About 20% of the coded measures did not include any of these elements – an important gap because these elements are likely to allow for a better user experience and potentially increase the amount of data we can collect from each participant. We found that the most common elements were: character assignment/interaction, story, categorization; and the least common were: levels, rewards, and points.
    - Character assignment/interaction (38%): a participant is "placed in the shoes" of a character within a story/hypothetical scenario and asked how they might interact with a certain person/people (given the situation).
    - Story (33%): a participant was exposed to, primed with, or asked to interact with an established story (fictional or nonfictional) as part of the task/measure (note: as opposed to a hypothetical scenario, a story includes more in-depth context, such as other characters' names, characters' backgrounds and motivations, environments, and/or dialogue).
    - Categorization (16%): a participant is asked to sort people, places, and/or things into categories or with each other (e.g., "___ people are good, while ___ people are bad", "[this person] is more like [that person]", "[social group] belongs in [place]").
    - Feedback (11%): a participant receives commentary or feedback (verbal or visual) during or after the completion of a task (e.g., "Good job!", "Why did you choose that?", smiley face, thumbs down).
    - Timed trial (7%): a task/measure has a time limit to complete and/or participants' response times are measured.
    - Points system (3%): completion of a task/measure results in a participant earning "points", potentially towards an end goal of points accumulated.
    - Rewards (1%): a (tangible) reward/gift is provided for completion of a task/measure.
    - Levels (1%): participants can progress to a new "level" or "stage" of a game by completing a task/measure.
    - Competition (0%): the progress, "points", rewards, or alternative performance indicators of a participant are compared with other participants/people/characters.

- For a subset of the measures, we coded whether the participants were asked about feelings (if the tasks asked how a participant feels about a particular person, race, or group), actions (e.g., participant was asked what they would like to do; or the task included observations of participants' behaviors), or both (if the task asked about feelings and actions to the same extent). We found that feelings were the most common type of question (57%), with action being the second most common (28%) and 'both' the least common (15%). Because both feelings and actions are important components of racial/ethnic bias, it will be important to incorporate both aspects into new measures.

- For a small subset of measures (still in progress), we coded the racial/ethnic groups depicted. We found that measures in U.S. studies tended to include only White and Black response choices. While studying anti-Black/pro-White bias is undoubtedly important, it is also critical to broaden the scope of biases that measures can assess; this is an important gap to be addressed.