The Development of Structural Thinking about Social Categories

Nadya Vasilyeva

Alison Gopnik

Tania Lombrozo

University of California, Berkeley

Word count: 5564 (excluding references, abstract, footnotes, figure and table captions)

This work was supported by the Varieties of Understanding Project (funded by the John Templeton Foundation), the McDonnell Foundation, and NSF grant SMA-1730660.

Corresponding author: Nadya Vasilyeva (vasilyeva@berkeley.edu), Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720 USA

Abstract

Representations of social categories help us make sense of the social world, supporting predictions and explanations about groups and individuals. In an experiment with 156 participants, we explore whether children and adults are able to understand category-property associations (such as the association between "girls" and "pink") in *structural* terms, locating an object of explanation within a larger structure and identifying *structural constraints* that act on elements of the structure. We show that children as young as 3-4 years old show signs of structural thinking, and that 5-6 year olds show additional differentiation between structural and non-structural thinking, yet still fall short of adult performance. These findings introduce structural connections as a new type of non-accidental relationship between a property and a category, and present a viable alternative to internalist accounts of social categories, such as psychological essentialism.

Keywords: structural explanation, structural factors, social categories, essentialism, category representation

The Development of Structural Thinking about Social Categories

Imagine that a school introduces a dress code stating that children must dress in solid colors. When school begins, most boys are wearing blue; most girls are wearing pink. What explains the correlation between gender and color? One explanation is that boys naturally prefer blue, and girls pink. But a glance at history reveals that in the 19th century, pink was considered the vigorous, masculine color, whereas girls wore "delicate and dainty" blue (Fausto-Sterling, 2012). If an explanation that appeals to intrinsic preferences is inadequate, an alternative might be to appeal to a structural feature of the environment: stores reliably stock more pink options for girls than for boys. In this case, availability could be a sufficient explanation for the observed correlation.

This example illustrates structural thinking. A hallmark of structural thinking is locating an object of explanation within a larger structure and identifying *structural constraints* that act on components of the structure to shape the distribution of outcomes for each component. In our example, girls occupy a position within larger social and institutional structures that make them more likely than boys to wear pink. A structural approach to social categories differs from *internalist* approaches, which focus on essential or inherent properties of the category itself. In the current paper, we ask whether and when children develop the ability to think about social categories in structural terms.

The most prominent internalist approach to theorizing about the representation of social categories is based on the notion of *psychological essentialism*, which refers to the tendency to represent (some) categories in terms of underlying essences that are constitutive of category membership and/or causally responsible for key category features (Gelman, 2003). Psychological essentialism offers an efficient basis for classification and

inference, but can also lead to unwarranted normative expectations about categories, stereotypical generalizations, and prejudice (Leslie, 2015).

Other approaches to social categories are similarly internalist. For example, Cimpian and Salomon (2014) proposed the *inherence heuristic* (distinct from but compatible with essentialism), defined as the tendency to explain observed patterns in terms of the inherent properties of the objects that instantiate them (see also Cimpian, 2015; Salomon & Cimpian, 2014). If girls wear pink, people might infer that it must be due to something inherent about pink ("it is delicate") and/or girls ("they are attracted to delicate colors"), rather than considering a broader range of external, historical factors. Another approach comes from Prasada and Dilllingham's (2006, 2009) *aspect hypothesis*, according to which some features of a category are viewed as *aspects* of the kind. For example, "fighting crime" is an aspect of being a police officer, so the feature "fighting crime" shares what they call a "principled" connection to the representation of the category, whereas a feature that is associated only statistically (e.g., "eating donuts") does not.

While psychological essentialism, the inherence heuristic, and the aspect hypothesis are importantly distinct in their commitments regarding categorical representations, they all support internalist explanations for associations between a category and a feature (e.g., "she chose pink because girls like delicate colors"), as well as formal explanations that appeal to category membership (e.g., "she chose pink because she is a girl"). By contrast, they lack mechanisms for representing structures as distinct from their elements, i.e. differentiating *kinds* ("girls") from the *structures* in which they are embedded (the social position occupied by girls). As a result, they cannot readily

accommodate the kind of structural thinking supported by a structural approach.

With a structural approach, reliable connections between properties and categories can be represented as a consequence of stable structural constraints acting on categories from the outside. Category-property associations thus support what philosophers of social science call *structural explanations*, which situate the object of explanation in a network of relationships within a larger, organized whole (a structure) (Haslanger, 2015). These explanations identify how relationships to other parts of the whole modify the probability distribution over possible states of the part whose properties are being explained (compared to a hypothetical case outside a structure, to other locations within the structure, or to different structures). For example, an internalist explanation for why many (married, heterosexual) women leave their jobs after having a child might appeal to women's priorities or abilities, whereas a structural explanation would identify constraints that affect women in virtue of their position within the social structure (e.g., unpaid parental leave, a gender wage gap). These structural constraints shift the probability distribution across different outcomes for women versus men. Under different structural constraints (e.g., "if society were organized differently" or "for men or women in a different culture"), the same event (having a child) need not trigger the same outcomes. Rather than pinpoint triggering causes (e.g., the baby's arrival), structural explanations identify constraints that shape the causal relationships between triggering causes and their effects (Dretske, 1988). To use a non-social example, consider whether the accelerator pedal causes the car to go. Under one structural arrangement of car parts, the pedal press triggers the car's movement. However, under a different structural arrangement (e.g., in a car in an autonomous driving mode, or in a neutral gear) this

relationship would no longer hold.

Like essentialist explanations, structural explanations can account for the relative homogeneity within social groups and the rich inductive potential of social categories. Indeed, some advocates for essentialism recognize that external constraints can give rise to these features (Rangel & Keller, 2011; Rhodes & Mandalaywala, 2017). But the structural view does more than acknowledge external factors; it also builds in a distinction between *nodes* (positions within social structures) and *node-occupiers* (categories that occupy those positions; Haslanger, 2015). This distinction brings to light a potential ambiguity in formal explanation (e.g. "Smith quit her job after the baby because she's a woman," where the term "woman" can refer to either the node or the node-occupier). Such explanations could attribute stable properties directly to the node (i.e., women's location in a structure), without necessarily tying them to its inherent nature (i.e., to women themselves). In other words, a formal explanation could support both structural and internalist interpretations, a prediction that our experiments test.

One way to appreciate what constitutes a structural explanation is to consider what it is not. Structural explanations are not merely "situation" explanations from the traditional person-situation dichotomy, such as appealing to unexpected traffic to explain why Mary is late (Ross & Nisbett, 2011), because structural explanations necessarily invoke *stable* constraints acting on a *category* in virtue of its *position in a structure*. Structural explanations also differ from "causal history of reasons" explanations (Malle, 2004), which are narrower in their restriction to intentional behavior, yet broader in allowing for *non-structural* antecedents to reasons. It is useful to think of structural explanations in terms of the ANOVA or "cube model" (Kelley, 1973), in which a

behavior is attributed to co-varying factors (person, situation, or stimulus). The cube model assumes that the data (behaviors) come from an "unconfounded" factorial design in which factors vary independently. Structural thinking is instead sensitive to *confounds* between people and situations; within a social structure, categories are often constrained by their nodes. The category "women" can only occupy the "women" node, which constrains the range of properties the occupier can display. When a social position and a category are thus confounded, a pattern of covariation between a category and property is compatible with (at least) two causal models, internalist and structural (i.e., the property can be caused either by the inherent characteristics of the category, or by the structural position). In this the structural approach departs from Kelley's original model, where situational and internal causes are expected to produce distinct covariation patterns.

The notion of a confound between a category and its social location also helps to position the structural view of categories relative to role-based categories, such as *guest*, which specify roles in relational structures, roughly corresponding to Haslanger's "nodes" (Asmuth & Gentner, 2016; Goldwater, Bainbridge, & Murphy, 2016; Markman & Stilwell, 2001). This research focuses on how people extract a common relation across a taxonomically diverse set of items (a hotel guest, a house guest, a dragonfly visiting your garden; Goldwater, Markman, & Stilwell, 2011) to form such representations. Structural thinking about social categories similarly requires representations of relational positions, but applies to cases where a relational position is *confounded* with membership in a (perceived) taxonomic category (e.g., when a particular position is more likely to be occupied by people of a particular gender) – a condition that need not hold for role-based categories like *guest*. The structural view is thus a genuine departure from prior work on

the representation of social categories.

The development of structural thinking. Structural thinking shares some characteristics with internalist thinking (e.g., supporting category homogeneity and inductive potential), but others with other forms of externalist reasoning (e.g., appealing to features "outside" the category). As a result, research that is not specifically designed to measure the signatures of structural thinking is hard to interpret. Preliminary work suggests that adults are able to engage in structural thinking (Vasilyeva & Lombrozo, in prep), and cross-cultural research on independent versus interdependent construals (Nisbett, 2003) suggests that the reasoning style associated with structural thinking is not as "unnatural" as it may seem. But are children able to engage in structural thinking? And if so, when does this ability develop?

Several findings suggest that young children may lack the conceptual prerequisites and/or knowledge to engage in structural thinking. For instance, prior work demonstrates that children view some social categories (such as gender) as essentialized natural kinds from an early age (Rhodes & Gelman, 2009; Taylor, 1996), even when cultural input suggests otherwise (Astuti et al., 2004). There is also evidence that young children tend not to endorse environmental factors as explanations for category features (Taylor, Rhodes, & Gelman, 2009), although the environmental factors that were examined were primarily non-structural in nature. Finally, as young as 4-5 years of age, children tend to generate and endorse "inherent" explanations of categorical patterns over "extrinsic" ones (Cimpian & Markman, 2011; Cimpian & Steinberg, 2014).

Beyond evidence of early essentialist and inherence-based reasoning, there is evidence that children lack capacities involved in structural thinking. Structural

explanation could rely on structure-wide counterfactual alternatives (i.e., considering how things would be if a structure were different), which may not emerge until age 7-8 (Beck et al., 2006; Rafetseder, Cristi-Vargas, & Perner, 2010). Structural reasoning also relies on representing relations, and research on relational reasoning suggests a developmental shift in relevant capacities throughout and beyond the preschool years (e.g., Gentner, 1983, 1988, 2005; Richland, Morrison, & Holyoak, 2006).

On the other hand, there is evidence consistent with the idea that children might engage in structural (or at least externalist) thinking from an early age. By age 3, children understand that emotions can have situational causes (Harris, 1989; Lagattuta, Wellman, & Flavell, 1997; Lagattuta & Wellman, 2001; Sayfan and Lagattuta, 2008, 2009; Wellman & Lagattuta, 2000), and by age 4, children can use covariation information to make situational over personal attributions (Seiver, Gopnik, and Goodman, 2013). Four-year-olds also recognize moral constraints on their own behavior (Chernyak & Kushnir, 2014) and acknowledge that the behavior of members of a social category can be driven by common norms (Kalish, 2011; Kalish & Shiverick, 2004; Rakoczy, Warneken, & Tomasello, 2008; Smetana, 1981; Turiel, 1983).[1] These findings suggest that children can engage in externalist and norm-based thinking, if not structural thinking per se.

A final and more intriguing possibility is that young children could be *more* open to structural thinking than older children and adults. Young children are more flexible than older children about some social categories, such as race (Rhodes & Gelman, 2009),

[1] Translating research on norms into predictions about structural reasoning is not straightforward. First, moral norms carry deontic content, which distinguishes them from other kinds of structural constraints (such as a wage gap) that do not. Second, category-specific norms can be interpreted in either essentialist or structural terms (e.g., if girls are not allowed to go out after 9 pm, this could stem from inherent characteristics of girls, or structural forces). Existing studies about norms have not made these distinctions, complicating their interpretation with regard to structural reasoning.

and less rigidly dispositional in their explanations for behavior (Gonzales, Zosuls & Ruble, 2010; Kalish, 2002; Rholes & Ruble, 1984; Seiver, Gopnik & Goodman, 2013). There is also evidence that they have weaker assumptions about causal structure, which can translate into superior learning of a structure that older children and adults don't anticipate encountering (Lucas, Bridgers, Griffiths, & Gopnik, 2014). This body of work suggests that relative to older children and adults, young children could have weaker expectations about the causal principles governing social categories, and thus be more willing to entertain a variety of representations. Our experiment tests these possibilities.

Experiment

This experiment had three goals: to determine whether and when children can successfully engage in structural thinking in explaining the association between a category and a property, to determine whether a structural construal can be experimentally induced, and to evaluate the prediction that structural thinking can support formal explanations. To accomplish these goals, we introduced a novel category-property association and we induced either an *internalist* construal (in a non-structural framing condition) or a *structural* construal (in a structural framing condition). We then prompted children to explain the association, coding their explanations as internalist, structural, or other. We predicted that on both open-ended and close-ended measures, the former condition would promote internalist explanations, and the latter would promote structural explanations. We included additional measures to probe other markers of structural thinking and to test our prediction about formal explanations.

For these additional measures, we adopted an approach mirroring Prasada and Dillingham (2006, 2009; see also Haward, Wagner, Carey, & Prasada, in prep.), who

developed a set of tasks that can be used to identify whether people construe the connection between a feature and a category as principled (e.g., "fighting crime" and being a police officer) or statistical (e.g., "eating donuts" and being a police officer). They showed that only principled connections between kinds and features supported *partial definitions* (a police officer is a person who fights crime), and *formal explanations* (this person fights crime because she is a police officer). We employed modified versions of these tasks, as well as a measure of *mutability*, which probed the extent to which a property-category association is perceived to be contingent on the structure within which the category is embedded.

We predicted that a structural construal, relative to an internalist construal, would manifest in higher ratings of property mutability (since the category-property association is contingent on the structure) and lower ratings for partial definitions (since the property is not inherent to the category). We also predicted that both internalist and structural thinking would support formal explanations. Specifically, if both experimental conditions succeed in framing the property-category connection as non-accidental, and if the category label invoked within a formal explanation can be taken to refer either to the category per se (under the internalist construal) or to the structural node (under the structural construal), we would expect the label to support explanations for the property-category association in each case.

Method

Participants. We recruited 41 3-4-year-olds (mean age 4.3 years, range 3.0-4.9; 23 females, 18 males), 48 5-6-year-olds (mean age 5.6 years, range 5.0-6.9; 23 females, 25 males), and 67 adults (mean age 33 years, range 19-71; 33 females, 64 males).

Children were recruited in local museums and preschools and tested in person using an illustrated storybook presented on a laptop; adults were recruited via Amazon Mechanical Turk and tested online.[2]

Materials, Design, and Procedure. Participants were first introduced to a school where girls and boys study in separate classrooms, and presented with fictitious data about students playing different games during recess: girls predominantly played Yellow-Ball while boys predominantly played Green-Ball. Participants were told that the game each child played was determined by tossing a pebble towards two buckets standing side-by-side: if the pebble fell into the yellow bucket, that child played Yellow-Ball that day, and if the pebble fell into the green bucket, that child played Green-Ball that day (Figure 1a); in the end, each child received a ball to play with.

The critical manipulation concerned the sizes of the buckets. In one condition, both buckets were of the same size (Figure 1b); we refer to it as the non-structural framing condition, so-named because it was designed to induce a non-structural, internalist mode of construal, by way of establishing that the structural factors (the bucket sizes in each classroom) did not favor one game over the other. The striking deviations in game choices from the chance pattern of "50% Yellow-Ball + 50% Green-Ball" in this condition thus provided evidence that girls and boys differed in their inherent preferences (see Kushnir, Xu, & Wellman, 2010, for evidence that even younger children infer preferences on the basis of such statistical evidence). In the structural framing condition, so named because it was designed to induce a structural construal, one bucket was much

[2] For adults, participation was restricted to users with an IP address within the US and an approval rating of at least 95% based on at least 50 previous tasks. The study was approved by the University of California Berkeley Committee for Protection of Human Subjects, Causal Learning in Children project, protocol #2010-01-631. The size of developmental sample was determined from power analyses based on effect sizes from pilot studies.

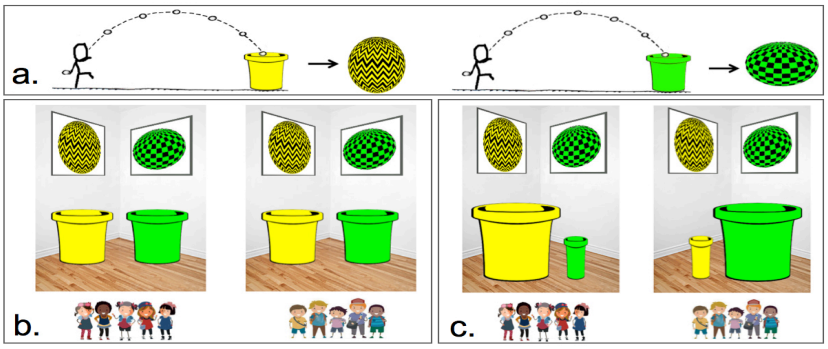

Figure 1. Illustrations of the procedure determining which game each student played in the story (a) and of the different constraints on the probability of outcomes in the non-structural (b) and structural conditions (c).

larger than the other: in the girls' classroom the yellow bucket was larger, with the reverse in the boys' classroom (Figure 1c). The size difference imposed a stable structural constraint on the probability distribution over options available to members of each category, inviting a structural interpretation of the category-property connection.

After comprehension checks, all participants completed a series of measures designed to differentiate an internalist from a structural construal of the property-category association (see Supplementary Materials for the full script and details). First, in the *open-ended explanation* task, participants were asked: "So, the girls in the girls' classroom play Yellow-Ball a lot at their school. Why?". Second, participants completed a causal explanation evaluation task and the three additional measures: mutability, partial definition, and formal explanation.

In the *causal explanation evaluation* task, children evaluated three kinds of causal explanations offered by puppets that "sometimes say things that are smart, and sometimes say things that are silly." The puppets explained that girls tend to play Yellow-Ball "because girls like playing Yellow-Ball" (internalist); "because in the girls' classroom,

it's easier to throw a pebble in the yellow bucket" (structural); or "because they got sprinkled with water" (an incidental explanation invoking an irrelevant fact from the cover story, included to monitor how much young children struggle differentiating the truth of a claim from its status as a good explanation; see Allen, 2008; Amsterlaw, 2006). Participants evaluated each explanation using a two-step, four-point thumb scale: they first chose one of two thumbs representing "good explanation" (up) and "bad explanation" (down), and they then chose between two subsequent options based on their choice: "kind of good/bad" (small thumb) or "really good/bad" (big thumb) – a scale previously shown to work well to measure children's agreement with explanations (Cimpian & Steinberg, 2014; Hussak & Cimpian, 2015).

For the *mutability judgment*, participants were told that after a change in the school's rules allowing children to attend any classroom, Suzy's parents transferred her to the boys' classroom "because they know the teacher there" (suggesting the transfer was not driven by Suzy's preferences). Participants were asked to guess which game Suzy would play the day after transferring, responding on a two-step, four-point scale ranging from "for sure Yellow-Ball" to "for sure Green-Ball." This mutability judgment mirrors more familiar "switched at birth" tasks in the essentialism literature (Gelman & Wellman, 1991), in which children are asked, e.g., whether a cow raised by pigs will moo or go oink. Similarly, our mutability judgment involves a change in environment (structural constraints), and participants are asked to infer whether a property will match the exemplar's category (the node occupier) or the new environment (the node). On an internalist construal of the category-property association, participants should predict that Suzy will play Yellow-Ball. On a structural construal, they should be more inclined to

think she will play Green-Ball. This shift would also show that structural positions are seen as influencing behavior, rather than merely reflecting internal preferences.

For the *partial definition* task, participants rated whether an alien did a good job telling what a girl is to another alien who had never heard about girls: "A girl is a person who plays Yellow-Ball a lot." Participants used a two-step, four-point scale ("really bad job" - "really good job").

In the *formal explanation* task, participants were asked to evaluate a puppet's formal explanation for why Suzy plays Yellow-Ball a lot at her school - "Because Suzy is a girl" - using the two-step, four-point thumb scale ranging from "really bad" to "really good."

Results and Discussion

Due to differing test formats and sample sizes, data from children and adults were analyzed separately. For the open-ended explanation task, participants' explanations were

Table 1

Open-ended explanation coding scheme: sample explanations coded as *internalist*, *structural*, or *miscellaneous*.

Internalist explanations: appeal to category members' liking, wanting, preferring, aiming for one of the games	Structural explanations: make a comparative statement about accessibility of the games for girls vs. boys	Miscellaneous: question restatements, proximal cause explanations & unclassifiable responses
"maybe the girls just like it better, so they always aim to get their pebbles into the yellow ball bucket"	"because the pebble went into the yellow bin, because the yellow one is bigger"	"I don't know"
"'cause they love yellow ball"	"because for the girls, it is easier to get their pebble into the yellow bucket"	"'cause they did"
"because they like the color yellow"		"the yellow ball is brighter than the green one"
		"because they need to get balls for fun"
		"because of the amount of times the pebble went into the yellow bucket"

coded as internalist, structural, or miscellaneous (see Table 1). The explanations were coded by two independent coders, Cohen's kappa=.87, p<.001 (see Appendix for additional details on the coding procedure).

The distribution of internalist and structural explanations was affected by the framing condition for each age group: Fisher's exact tests comparing response distributions as a function of framing (non-structural, structural) x explanation type (internalist, structural) were significant, $p_{younger}$=.032; p_{older}<.001; p_{adults}<.001. As Figure 2 shows, structural explanations were more likely to be produced under the structural framing than the non-structural framing in all age groups (Fisher's exact tests on proportion of structural explanations, $p_{younger}$=.048; p_{older}<.001; p_{adults}<.001). There was also an overall trend of producing more internalist explanations under the non-structural framing than the structural framing, reflecting the efficacy of the non-structural framing condition in inducing an internalist construal; the difference was significant for adults (p<.001),

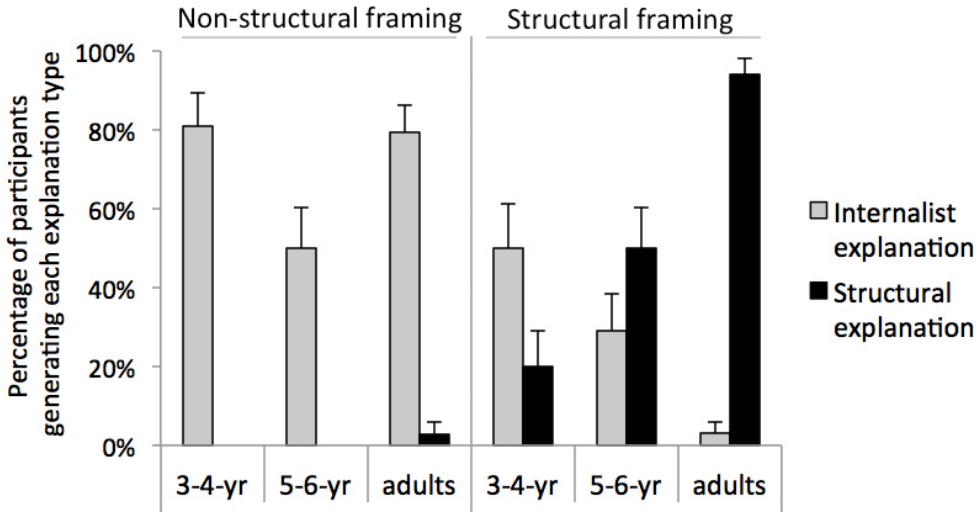

Figure 2. Distribution of internalist and structural explanations generated in response to question about why girls play Yellow-Ball, as a function of framing condition and age group. Error bars represent 1 SEM.

marginal for the younger children (p=.052), and not significant for the older children (p=.238), although the difference was in the predicted direction.

Critically, in the structural framing condition some proportion of participants in each age group produced structural explanations (Figure 2, right panel, black bars). There was also evidence of developmental change in children's response to structural framing, age group (younger, older) x generated explanation (internalist, structural) $\chi^2(1, N=33)=3.86$, p=.049). Specifically, the two age groups showed opposite response trends: whereas younger children were more likely to generate internalist explanations than structural explanations, older children were more likely to generate structural explanations than internalist explanations.

The causal explanation evaluation task (see Figure 3) similarly revealed an effect of framing, but only for older children. Specifically, a mixed ANOVA on children's evaluations as a function of explanation type (internalist, structural, incidental), framing (non-structural, structural), and age group (3-4, 5-6) revealed an interaction between explanation type and condition, $F(2,170)=6.00$, p=.003, η_p^2=.066, qualified by a three-way interaction including age, $F(2,170)=3.73$, p=.026, η_p^2=.042 (also significant if

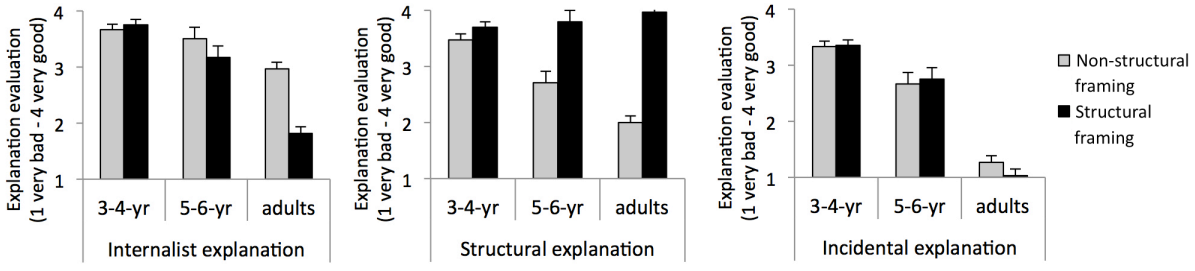

Figure 3. Explanation evaluation as a function of explanation type, framing condition, and age group. Error bars represent 1 SEM.

restricting the analysis to internalist and structural explanations, p=.012). The interaction was driven by the selective effect of framing on 5-6-year-olds' evaluations of the structural explanation: while the youngest group was not sensitive to the framing manipulation, the 5-6-year-olds rated structural explanations higher in the structural condition than in the non-structural condition ($p_{younger}$=.390, p_{older}<.001). There was also a main effect of explanation type, $F(2,170)=9.87$, η_p^2=.104, with lower ratings for the incidental explanations than the internalist (p<.001) and structural (p=.002) explanations, which did not differ from each other (p=.452).

For adults' ratings, an explanation type (essentialist, structural, incidental) by framing (non-structural, structural) mixed ANOVA revealed the expected interaction, $F(2,126)=117.83$, p<.001, η_p^2=.652: structural explanations were rated higher under the structural than non-structural framing, and the reverse held for the internalist explanations (planned pairwise comparisons p's<.001). This interaction also drove a marginal effect of framing, $F(1,63)=3.74$, p=.058, η_p^2=.056), with a trend for higher ratings in the structural condition. Finally, there was a main effect of explanation type, $F(2,126)=171.15$, p<.001, η_p^2=.731: ratings decreased significantly from structural to internalist to incidental explanations (all pairwise p's<.001).

Having found evidence of structural thinking in our open- and close-ended causal explanation tasks, we turn to our additional measures. For the *mutability judgment* task (Figure 4a), we predicted that properties construed as structural (under the structural framing) would be more mutable than properties construed as internalist (under the non-structural framing). Consistent with this prediction, an ANOVA with framing condition and age group as between-subjects factors revealed the predicted main effect of framing,

$F(1,85)=8.95$, $p=.004$, $\eta_p^2=.095$, with no main effect of age group, $F(1,85)=1.05$, $p=.309$, nor interaction, $F(1,85)<.01$, $p=.984$. Similarly, adults rated the target property as more mutable under the structural than non-structural framing, $t(65)=8.04$, $p<.001$, $d=2.00$.

For the *partial definition* task (Figure 4b), we predicted that properties construed as internalist would support category definitions better than properties construed as structural. However, an ANOVA on children's ratings with framing condition and age group as between-subjects factors did not reveal a significant effect of framing, $F(1,85)=.18$, $p=.675$. Neither the age effect, $F(1,85)=.36$, $p=.360$, nor the interaction, $F(1,85)=.02$, $p=.887$, was significant. In contrast, adults displayed the predicted pattern, $t(65)=2.11$, $p=.039$, $d=.52$.

Finally, as predicted, *formal explanation* ratings did not significantly differ across the non-structural and structural conditions for either group of children or for adults, all p's$\geq.916$ (Figure 4c), suggesting that these explanations support both internalist and structural construals.

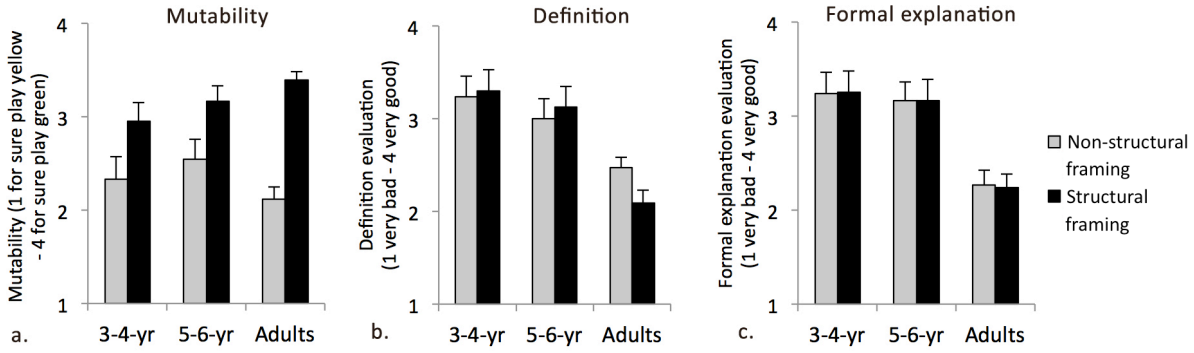

Figure 4. Mutability (a), partial definition (b), and formal explanation ratings (c) as a function of framing condition and age group. Error bars represent 1 SEM.

These results reveal that even young children are capable of structural thinking, as reflected in their open-ended explanations and their judgments concerning the mutability of properties under structural changes. They also provide the first demonstration that across all age groups, formal explanations support two interpretations: internalist and structural. Beyond these age-general effects, we find developmental changes in structural thinking, with older children and adults more readily engaged in structural thinking. Notably, the observed pattern of developmental change is not due to younger children simply not understanding the task or the explanations: in the explanation evaluation task, the youngest children discriminated meaningful (internalist or structural) explanations from merely true statements (incidental explanations), and in the explanation generation task they produced meaningful explanations sensitive to the framing of the property-category association. Finally, our results show that the mutability measure can effectively differentiate internalist from structural thinking across development, and the partial definition task offers an additional measure of differentiation for adults.

General Discussion

Using novel tasks designed to assess structural thinking, we find evidence that even young children are able to reason about social categories in structural terms, as manifested in 3-4-year-olds' self-generated explanations and judgments of property mutability. By 5-6 years, children preferentially generated and accepted structural explanations for a category-property association when a structural constraint was presented. Not until adulthood, however, did participants show sensitivity to structural factors in evaluating partial definitions.

Recognizing structural reasoning as a distinct cognitive phenomenon invites us to

rethink findings in the literature on essentialism. For example, many discussions of essentialism emphasize its capacity to support generalizations across category members (e.g., Gelman, 2003). In fact, generalization tasks are often used to *measure* the extent to which a category representation is essentialized. However, structural representations can also support generalizations when stable constraints act on a category occupying a node. Structural explanations identify broad patterns that hold robustly across "inessential perturbations" within stable structures (Haslanger, 2015). It follows that the stability and generalizability of category properties need not imply internalist (essentialist) representations (see Rhodes & Mandalaywala, 2017, for a related point). Our findings thus lay the groundwork for refining internalist claims and the evidence taken to support them.

We also find that formal explanations support both structural and internalist interpretations. In the structural condition, we suggest that participants were able to construe the category label as a pointer to the node, and that this in turn rendered formal explanations acceptable because the explanations identified a causal or lawful regularity relating the node and the property in question. In the non-structural condition, participants observed a correlation between category membership and game choice that could not be attributed to structural factors. We thus expected participants to infer that girls and boys differed in their internal preferences (Kushnir, Xu, & Wellman, 2010), and the prevalence of internalist explanations confirms that they did. For these participants, we suggest that formal explanations were acceptable because they identified a principled or causal relationship between the category and the property (Prasada and Dillingham, 2006, 2009; Prasada, Khemlani, Leslie, & Glucksberg, 2013). However, it remains an

open question just what kind of relationship participants inferred. In the structural condition, it is unclear whether participants interpreted the node-property connection in specifically causal terms. In the non-structural condition, participants were not offered direct evidence that the relationship between the category and the property was *principled* in Prasada and Dillingham's sense; it remains possible that it was instead taken to support formal explanations because the statistical association was so strong (Haward, Wagner, Carey, & Prasada, in press), and that a truly principled connection would support even stronger endorsements of formal explanations. Identifying the conditions under which children and adults infer different kinds of relationships, and the differential implications of those relationships, is an important step for future research.

Our findings concerning formal explanations raise the intriguing possibility that generics (e.g., "Girls prefer pink") could similarly support structural interpretations. On most accounts, generics are interpreted as expressing something about the underlying nature of the category, reinforcing essentialist beliefs and potentially perpetuating harmful stereotypes (Cimpian & Markman, 2011; Leslie, 2014; Prasada & Dillingham, 2009). For example, Leslie argues that generics are by default interpreted as expressing "generalizations that hold because of common, inherent features of the members of the kind" (p. 217). But if people can interpret generics structurally, this potentially offers a new way to mitigate harmful side-effects of generic language without purging it from everyday speech (or, equally implausibly, convincing people that many associations between properties and social categories are merely "accidental").

More generally, it is valuable to consider whether structural thinking about social categories might mitigate prejudice against category members, and if so, which

element(s) of structural thinking could generate this effect. Rhodes and Mandalaywala (2017) cite evidence suggesting that essentialism promotes the view that the relative social status of different groups reflects objective, underlying reality, thereby supporting prejudice and endorsement of the status quo. A structural construal could have an opposing effect by dissociating social status from objective and intrinsic properties of group members, and/or by promoting belief in the mutability of social properties. Importantly, work on Belief in Social Determinism (BSD) suggests that merely citing external factors is insufficient to mitigate prejudice: Rangel and Keller (2011) find that BSD, a lay theory that "a person's essential features…are shaped *permanently* and *profoundly* by social factors" (p. 1, our emphasis), is associated with the same outcomes as other essentialist beliefs: dispositional thinking, expectations of stability across situations and time, negative stereotyping, prejudice, discrimination, and hierarchy-enhancing ideologies (e.g., nationalism). This suggests that if structural explanations have a mitigating effect on prejudice, it is not merely in virtue of citing external factors, since BSD does so as well. An important difference may be that while BSD invokes external factors, it utilizes them to tell an "inside story" (Plaut & Markus, 2005): a story of how an African American man from a poor neighborhood came to develop criminal tendencies, or how a woman came to be submissive, where both individuals were "permanently and profoundly" shaped by social factors. In contrast, a structural story would emphasize how stable external factors can give rise to similar outcomes, but without necessarily implicating one's character and/or implying permanence. While further research is clearly needed, we speculate that structural explanations that highlight *mutability*, in particular, might be an effective way to mitigate prejudice.

Our studies succeed in differentiating structural explanations from internalist explanations, but the contrast with BSD highlights the need for further distinctions *within* the class of externalist explanations. How does structural thinking differ from reasoning in terms of more transient, situational factors? Do all and only structural explanations give rise to the reported effects, or can other types of externalist (situational) explanations produce similar outcomes? Although the current study does not target this question directly, the fact that our participants generated explanations specifically citing structural factors (stable environmental constraints acting on categories, rather than a broader class of external circumstances; see Appendix), especially in the structural condition, suggests that our participants engaged in genuinely structural reasoning, rather than externalist thinking construed more broadly. In current work, we are contrasting structural with non-structural situational explanations more directly (Vasilyeva, Gopnik, & Lombrozo, in prep).

Future research should also examine reasoning about complex causal systems involving both internalist and structural factors. For example, if one believes that inherent properties determine the assignment of social categories to their "natural niches" (e.g., women's social position reflects their inherent properties), then merely acknowledging the causal influence of nodes on categories may be insufficient to increase the perceived mutability of relevant properties, and thereby mitigate prejudice. An effective structural explanation may need to explicitly reject the idea that the node-category confound is due to the category's inherent properties.

Other important questions about structural reasoning concern the mechanisms underlying developmental change. Although we find evidence of structural thinking by 3-

4, internalist thinking was arguably more robust: even under the structural framing, younger (and, to a lesser degree, older) children frequently generated internalist explanations, and older children rated internalist explanations as relatively good (consistent with prior research on the prevalence of internalist interpretations; Cimpian & Markman, 2009, 2011; Cimpian & Salomon, 2014). If internalist thinking reflects an early-emerging preference or default, it will be important to understand the maturational and learning processes shaping its scope, including the contributions of cultural input and relevant cognitive capacities, such as counterfactual reasoning and executive function. Another important question concerns how internalist (essentialist) and structural thinking co-exist. For instance, if internalist or essentialist thinking serves as a kind of default, it could be that structural thinking requires greater evidence and/or effort.

By introducing structural thinking as a special case of externalist reasoning that exhibits some of the signatures of internalist/essentialist thinking, we have unmasked a gap in our understanding of categorical reasoning, and opened up new directions of study that could help account for some of the mixed evidence in research on the development of relational reasoning, essentialist beliefs about social categories, and reasoning about moral and conventional norms.

Acknowledgements

We thank Grace Goodwin, Sarah Tencher, and Marissa Reid for their help with data collection, and three anonymous referees for their useful suggestions.

References

Allen, E. A. (2008). *Constructing and judging answers to why-questions: What can children tell us?* Doctoral dissertation, New York University. Available from ProQuest Dissertations and Theses Database (Order No. 3320766). Retrieved from https://search.proquest.com/docview/53356078?accountid=14496

Amsterlaw, J. (2006). Children's beliefs about everyday reasoning. *Child Development, 77*(2), 443-464.

Asmuth, J., & Gentner, D. (2016). Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology, 70*(10), 2007-2025. doi:10.1080/17470218.2016.1219752

Astuti, R., Solomon, G. E., Carey, S., Ingold, T., & Miller, P. H. (2004). Constraints on conceptual development: A case study of the acquisition of folkbiological and folksociological knowledge in Madagascar. *Monographs of the society for Research in Child Development*, i-161.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development, 77*(2), 413- 426. doi:10.1111/j.1467-8624.2006.00879.x

Chernyak, N., & Kushnir, T. (2014). The self as a moral agent: Preschoolers behave morally but believe in the freedom to do otherwise. *Journal of Cognition and Development*, *15*(3), 453-464. doi:10.1080/15248372.2013.777843

Cimpian, A. (2015). The inherence heuristic: Generating everyday explanations. In R. Scott & S. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (pp. 1–15). Hoboken, NJ: John Wiley and Sons.

doi:10.1002/9781118900772.etrds0341

Cimpian, A., & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition, 113*(1), 14–25. doi:10.1016/j.cognition.2009.0

Cimpian, A., & Markman, E.M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development, 82*(2), 471–492. doi:10.1111/j.1467-8624.2010.01525.x

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences, 37*(5), 461–480. doi:10.1017/s0140525x13002197

Cimpian, A., & Steinberg, O. D. (2014). The inherence heuristic across development: Systematic differences between children's and adults' explanations for everyday facts. *Cognitive Psychology, 75*, 130–154. doi:10.1016/j.cogpsych.2014.09.001

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Fausto-Sterling, A. (2012). *Sex/Gender: Biology in a social world*. Routledge.

Gelman, S.A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Gelman, S.A., & Wellman, H.M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 23*, 183-209. doi:10.1016/0010-0277(91)90007-q

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155-170. doi:10.1207/s15516709cog0702_3

Gentner, D. (1988). Metaphor as structure-mapping: The relational shift. *Child*

Development, 59, 47-59. doi:10.2307/1130388

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe & D. Rakison (Eds.), *Building object categories in developmental time* (pp. 245–275). Mahwah, NJ: Erlbaum.

Goldwater, M. B., Bainbridge, R., & Murphy, G. L. (2016). Learning of role-governed and thematic categories. *Acta Psychologica*, *164*, 112-126. doi:10.1016/j.actpsy.2015.10.011

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition, 118*(3), 359-376. doi:10.1016/j.cognition.2010.10.009

Gonzales, C. M., Zosuls, K. M., & Ruble, D. N. (2010). Developmental change in the understanding of trait terms. Traints as dimensions or categories? *Developmental Psychology, 46*, 1078-1088. doi:10.1037/a0020207

Harris, P. L. (1989). *Children and emotion: The development of psychological understanding*. Cambridge, MA: Basil Blackwell.

Haslanger, S. (2015). What is a (social) structural explanation? *Philosophical Studies*, 1–18. doi:10.1007/s11098-014-0434-5

Haward, P., Wagner, L., Carey, S., & Prasada, S. The development of principled connections and kind representations. Revision under review in *Cognition*.

Hussak, L. J., & Cimpian, A. (2015). An early-emerging explanatory heuristic promotes support for the status quo. *Journal of Personality and Socisal Psychology, 109*(5), 739-752. doi:10.1037/pspa0000033

Kalish, C. W. (2002). Children's predictions of consistency in people's actions. *Cognition,*

84, 237-265. doi:10.1016/S0010-0277(02)00052-5

Kalish, C. W. (2011). Generalizing norms and preferences within social categories and individuals. *Developmental Psychology*, *48*(4), 1133-1143. doi:10.1037/a0026344

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, *19*(3), 401–416. doi:10.1016/j.cogdev.2004.05.004

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologists*, *28*(2), 107-128. doi:10.1037/h0034225

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134-1140. doi:10.1177/0956797610376652

Lagattuta, K. H., & Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Development, 72*(1), 82-102. doi:10.1111/1467-8624.00267

Lagattuta, K. H., Wellman, H. M., & Flavell, J. H. (1997). Preschoolers' understanding of the link between thinking and feeling: Cognitive cuing and emotional change. *Child Development*, *68*(6), 1081-1104. doi:10.1111/j.1467-8624.1997.tb01986.x

Leslie, S.-J. (2015). The original sin of cognition: Fear, prejudice and generalization. *The Journal of Philosophy, 3*.

Leslie, S.-J. (2014). Carving up the social world with generics. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford Studies in Exp. Philosophy, (pp*. 208-232). doi:10.1093/acprof:oso/9780198718765.003.0009

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better

(or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. doi:10.1016/j.cognition.2013.12.010

Malle, B. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: The MIT Press.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329-358. doi:10.1080/09528130110100252

Miller, J. G. (1984). Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology, 46,* 961-978. doi:10.1037/0022-3514.46.5.961

Nisbett, R.E. (2003). *The geography of thought: How Asians and Westerners think differently… And why*. New York, NY: The Free Press.

Plaut, V. C., & Markus, H. R. (2005). The "inside" story: A cultural-historical analysis of being smart and motivated, Americal style. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 457-488). New York, NY: Guilford Press.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*, 73–112. doi:10.1016/j.cognition.2005.01.003

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: a window onto the formal aspect of common sense conception. *Cogitive Science*, *33*(3), 401–48. doi:10.1111/j.1551-6709.2009.01018.x

Prasada, S., Khemlani, S., Leslie S.J., & Glucksberg, S. (2013). Conceptual distinctions

among generics. *Cognition, 126*(3), 405-422.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child development*, *81*(1), 376-389. doi:10.1111/j.1467-8624.2009.01401.x

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881. doi:10.1037/0012-1649.44.3.875

Rangel, U., & Keller, J. (2011). Essentialism goes social: Belief in social determinism as a component of psychological essentialism. *Journal of Personality and Social Psychology, 100*(6), 1056-1079. doi:10.1037/a0022401

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*, 249–273. doi:10.1016/j.jecp.2006.02.002

Rhodes, M., & Gelman, S. (2009). A developmental examination of the conceptual structure. *Cognitive Psychology*, *59*(3), 244–274.

Rhodes, M., & Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science*. doi:10.1002/wcs.1437

Rholes, W. S., & Ruble, D. N. (1984). Children's understanding of dispositional characteristics of others. *Child Development, 55,* 550-560. doi:10.2307/1129966.

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Sayfan, L., & Lagattuta, K. H. (2008). Grownups are not afraid of scary stuff, but kids are:

Young children's and adults' reasoning about children's, infants', and adults' fears. *Child Development, 79*(4), 821-835. doi:10.1111/j.1467-8624.2008.01161.x

Sayfan, L., & Lagattuta, K. H. (2009). Scaring the monster away: What children know about managing fears of real and imaginary creatures. *Child Development, 80*(6), 1756-1774. doi: 10.1111/j.1467-8624.2009.01366.x

Salomon, E., & Cimpian, A. (2014). The inherence heuristic as a source of essentialist thought. *Personality and Social Psychology Bulletin, 40*(10), 1297–1315. doi:10.1177/0146167214541659

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child Development, 84,* 443–454. doi:10.1111/j.1467-8624.2012.01865.x

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52,* 1333–1336. doi:10.2307/1129527

Taylor, M. G. (1996). The development of children's beliefs about social and biological aspects of gender differences. *Child Development, 67*, 1555-1571. doi:10.2307/1131718

Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development*, 80(2), 461-481. doi:10.1111/j.1467-8624.2009.01272.x

Thibaut, J. P., French, R., & Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive function. *Journal of Experimental Child Psychology, 106*(1), 1-19. doi:10.1016/j.jecp.2010.01.001

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, UK: Cambridge University Press.

Vasilyeva, N., & Lombrozo, T. (in prep.) Structural thinking.

Vasilyeva, N., Gopnik, A., & Lombrozo, T. (in prep.) Structural thinking contrasted with non-structural externalist reasoning.

Waite, S., & Denier, N. (2015). Gay pay for straight work: Mechanisms generating disadvantage. *Gender and Society, 29*(4), 561-588. doi:10.1177/0891243215584761

Wellman, H. M., & Lagattuta, K. H. (2000). Developing understandings of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Appendix

Open-ended explanation coding procedure

The explanations were coded by two independent coders, both blind to age, gender, and condition (one of the coders was hypothesis-blind; the second coder (the first author) was not; both coded 100% of the responses, Cohen's kappa=.87, $p<.001$; disagreements were resolved via discussion). Each explanation could be coded into multiple categories, but in practice all received only one code. In the single case where a response belonged to multiple categories (an adult in the essentialist condition produced internalist, structural and miscellaneous explanations), we retained the code corresponding to the first mentioned reason.

The miscellaneous category included explanations that could be construed as situational but not structural (i.e., they did not cite stable structural constraints), e.g. proximal cause explanations citing prior causal events from the game-selection procedure, e.g., "because they threw a pebble and that's where it landed" (non-structural condition: three 5-6-yr-olds, four adults; structural condition: one 4-yr-old, one 5-yr-old) and explanations citing properties external to the social groups, albeit intrinsic to the games, e.g., "'cause it's fun" (non-structural condition: one 4-yr-old, two 5-yr-olds; structural condition: one 4-yr-old, one 5-yr-old). Given that such externalist-but-not-structural explanations were so infrequent, we combined them with other miscellaneous explanations. Importantly, such explanations were extremely rare in the structural condition, suggesting that participants attended specifically to the structural properties of the environment rather than indiscriminately sampling from a broadly construed set of external circumstances.