Weicker, Merle & Schulz, Petra. 2024. Children and adults privilege linguistic over visual information when creating comparison classes for prenominal gradable adjectives. *Glossa: a journal of general linguistics* 9(1). pp. 1–36. DOI: https://doi.org/10.16995/glossa.9912

Open Library of Humanities

Children and adults privilege linguistic over visual information when creating comparison classes for prenominal gradable adjectives

Merle Weicker, Goethe University Frankfurt, Germany, weicker@em.uni-frankfurt.de

Petra Schulz, Goethe University Frankfurt, Germany, p.schulz@em.uni.frankfurt.de

Mastering the semantics of gradable adjectives is a complex acquisition task, since their interpretation requires the calculation of a threshold relative to a comparison class. Context plays a role in determining the comparison class in multiple ways; empirical research has focused on non-linguistic cues, such as world knowledge and visual context, and less on linguistic cues, such as the modified head noun. The present study is the first to examine how preschoolers and adults use the taxonomic category encoded by the modified head noun as well as visual cues to interpret relative (*big, small*) and absolute (*clean, dirty*) gradable adjectives. Using existing objects from different taxonomic categories (basic-level and superordinate-level), we developed a picture-choice task that allowed us to infer the comparison class and the threshold from participants' choices. Forty-three children (aged 3 to 5) and 26 adults, all monolingual German-speaking, were tested. Confirming previous research, children, like adults, showed a context-sensitive interpretation of relative but not of absolute gradable adjectives. The taxonomic label was preferred to visual cues in creating the comparison class already by age three; this advantage of linguistic over non-linguistic information was also present in the adults. We explain this finding with the privileged role of information provided by the taxonomic label: the modified head noun encodes the cognitive representation of an object category and the corresponding taxonomic hierarchy. Accordingly, this representation is easily accessible while creating the comparison class, which provides further evidence that language is a powerful tool to modulate our mental computations.

1 Introduction

Imagine the following scenario: a kindergarten group has to move to another room. The children are helping their teachers to pack up their toys in boxes, among them all kinds of balls including bouncy balls, basketballs, soccer balls, ping-pong balls, and tennis balls, as well as hopper balls, and many other toys such as marbles. In this context, one of the teachers may say *Look at this one; this one is big,* pointing to one of the balls. Then, pointing to a different ball, she may say *Look at this one; this one is small.* In order to evaluate whether these utterances are appropriate descriptions of the objects the teacher is pointing to, the child has to understand what her teacher uses as a basis of comparison for the adjectives. Put differently, the child needs to know the comparison class. *Big* and *small* are gradable adjectives and their interpretation crucially relies on knowledge of the comparison class. That is, judging an entity as big or small in a given context requires determining the threshold for *big* or *small*, which depends on the class of entities it belongs to and how it compares to the entities in this class (e.g., Cresswell 1976). Utterances such as *This one is small/big* differ from utterances like *This one is small/big for a ball* in that the latter indicate the comparison class overtly, via the *for*-PP. In the former case, the child needs to determine the comparison class herself: is the object big for a ball, big for toys, big for fitting in a box, etc.? Cues to the comparison class may be non-linguistic or linguistic in nature. Non-linguistic cues concern the visual context (e.g., presence of other toys in our scenario), most relevant for the current study, but they may also concern the discourse context (e.g., the current topic of the conversation like packing boxes) and world knowledge (e.g., the typical size of balls). Linguistic cues include the modified head noun, for instance *balls* in *big balls.* The child can employ the category label *ball* to restrict the comparison class to the set of balls. Accordingly, in order to fulfill a request such as *Please hand me the big balls*, the child's task is to decide which of the balls in a specific situation are big and which ones are not. In the scenario sketched above, the child may opt for the soccer balls and basketballs and decide against the tennis balls, bouncy balls, and ping-pong balls. If the teacher in the same scenario says *Please hand me the big toys*, the child may choose the hopper balls and the basketballs and ignore the other balls, along with other small toys like marbles.

Adjectives such as *big* and *small* are referred to as relative gradable (henceforth RGA) and are distinguished from absolute gradable adjectives such as *clean* and *dirty* (henceforth AGA) (Kennedy & McNally 2005). The two classes of gradable adjectives differ regarding the role of the comparison class in that the comparison class seems to affect the threshold much less for AGAs than for RGAs. The requests *Please hand me the clean balls* and *Please hand me the clean toys*, for instance, will most likely result in picking out the same set of balls, whereas the requests *Please hand me the big balls* and *Please hand me the big toys* would not.

Capitalizing on these properties of RGAs and AGAs, acquisition research has asked whether children are sensitive to the distinction between RGAs and AGAs. Most experimental studies have focused on the role of non-linguistic cues for the interpretation of RGAs and have examined which

contextual information children use for calculating the threshold. Here, we extend this line of research and investigate the role of linguistic cues for children's interpretation of RGAs, comparing them to non-linguistic cues and to the role of both types of cues for AGAs. More specifically, we investigate whether the threshold shifts as a function of changes in the taxonomic category encoded by the modified head noun (basic-level vs. superordinate-level, henceforth called 'taxonomic label'), and we compare the effect of this linguistic change with the effect of a non-linguistic change involving the object categories present in the visual context. German-speaking preschoolers and adults participated in a picture-choice task targeting the comprehension of the RGAs *groß* 'big' and *klein* 'small' and the AGAs *sauber* 'clean' and *dreckig* 'dirty' in different linguistic and visual contexts. This experimental design allows us to uncover the role of taxonomic labels relative to the role of visual contextual cues in determining the comparison class and in calculating the threshold. It also allows us to seek further support for the finding that children distinguish RGAs and AGAs. By comparing child and adult data, we explore whether children and adults differ in the cognitive representation of the comparison class (henceforth abbreviated as CC).

The paper is structured as follows: Section 2 summarizes the theoretical analysis of RGAs and AGAs and spells out the acquisition task for the child. Section 3 reviews previous research on children's interpretation of RGAs and AGAs. Our study is presented in Section 4, the findings are discussed in Section 5. Section 6 offers a conclusion.

2 Gradable adjectives

Gradable adjectives are commonly analyzed within degree-based frameworks (e.g., Cresswell 1976; von Stechow 1984; Kennedy 2007).[1] According to this analysis, gradable adjectives denote relations between individuals and degrees on a scale associated with some dimension of measurement such as height (type $<d,<e,t>>$); this is illustrated in (1) for *big*. Degrees are understood as representations of measurement; a scale is a set of degrees that is totally ordered regarding a dimension such as height. In (1), **big** is a measure function that maps its argument onto the height scale (type $<e,d>$).[2]

(1) $[\![\text{big}]\!] = \lambda d \lambda x.\mathbf{big}(x) \geq d$

To evaluate whether the property denoted by a gradable adjective is true of an individual, the individual must be related to a degree that exceeds a certain threshold to a sufficiently salient extent (Kennedy 2007). The unmarked positive form is assumed to combine with a null degree

[1] Alternative semantic approaches to gradability – so-called delineation approaches – analyze gradable adjectives as context-dependent partial functions that cause a partition of a comparison class into a positive extension, a negative extension, and an extension gap consisting of entities that fall in neither of the two (e.g., Klein 1980; Burnett 2016).

[2] In some degree approaches (e.g., Kennedy 2007), gradable adjectives are analyzed as measure functions that assign to individuals a unique degree (type $<e,d>$). This analysis shares the same underlying assumptions about degrees and scales with the relational analysis in (1).

morpheme (*pos*) that binds the degree variable and introduces the threshold. Kennedy's (2007) implementation of *pos* is given in (2):

(2) $[\![\text{pos}]\!] = \lambda g_{<e,d>}\lambda x_e.g(x) \geq \mathbf{s}$, with **s** is a context sensitive function that chooses a standard of comparison in such a way as to ensure that the objects that the positive form is true of 'stand out' in the context of utterance, relative to the kind of measurement that the adjective encodes (Kennedy 2007: 17).

There is agreement that context typically affects the threshold for RGAs but not for AGAs. Both adjective classes are discussed in turn in Section 2.1 and 2.2. In Section 2.3, we sketch the child's acquisition task, given the theoretical analysis of gradable adjectives.

2.1 Relative gradable adjectives

RGAs such as *big* have variable, context-dependent thresholds.[3] The variability of the threshold has been explained by the assumption that different CCs can result in different thresholds. The threshold for *big*, for example, differs across object categories (i.e. CCs) such as elephants and mice. In addition, the threshold for *big* can differ across object categories at different taxonomic levels: it differs for mice and animals just as it does for balls and toys. Importantly, the CC can be created based on linguistic and/or non-linguistic information. As far as linguistic markers are concerned, *for*-PPs express the CC overtly (see (3a), (4a)) and the head nouns modified by the adjective express the CC implicitly (see (3b), (4b)). In both cases, the linguistic markers can restrict the CC to the category encoded by the noun. A change of the linguistic markers affects the truth conditions of the respective sentences. Whereas the sentences in (3a–b) are true, the sentences in (4a–b) are false: compared to other animals, such as dogs or elephants, a mouse does not count as big. Accordingly, a single referent (e.g., Otto) may be evaluated as being big, if it is construed as a member of a basic-level category (mouse), but not if it is construed as a superordinate-level category (animal).

(3) a. Otto is big for a mouse.
b. Otto is a big mouse.

(4) a. Otto is big for an animal.
b. Otto is a big animal.

Examples (3b) and (4b) indicate that the set denoted by the modified head noun is a potential candidate for the CC. However, as illustrated in (5a–b) (adapted from Kamp & Partee 1995), the modified noun does not always provide the CC.

[3] A discussion of the different notions of 'context' proposed in the literature is beyond the scope of the present paper. Here, we follow Burnett (2016) in assuming that variation across CCs is a major source of contextual variation and we limit ourselves to the CC as an important aspect of context. Accordingly, we use the terms context and CC interchangeably.

(5) a. The toddlers built a big snowman.
b. The teachers built a big snowman.

(6) This snowman is big.

The adjective *big* modifies the noun *snowman* in (5a) and (5b), but the thresholds for *big* are surely not the same. This is because different agents, expressed by the subject DP, carried out the snowman-building, and we attribute different snowman-building abilities to these agents. Accordingly, other linguistic material besides the modified noun (here: the subject DP) together with world knowledge (here: regarding toddlers' and teachers' capacity for building snowmen) can contribute to the calculation of the threshold. The CC can also be moderated by non-linguistic information such as the visual context or the discourse context: (5a–b) could be considered true if the statement refers to a specific snowman in a region where snow is uncommon, but false in a region that receives a lot of snow. In case linguistic information about the CC is absent as in (6), non-linguistic information is even more crucial. Listeners have to rely exclusively on non-linguistic cues to determine the appropriate CC and to calculate the threshold. They can do so by evaluating the visual context (e.g., the speaker is looking at other, bigger or smaller snowmen), the discourse situation (e.g., the conversation takes place in a country that receives a lot of snow or little), or their world knowledge (e.g., the typical size of snowmen).

2.2 Absolute gradable adjectives

In contrast to RGAs, the threshold for AGAs is typically not dependent on the CC (Kennedy & McNally 2005; Kennedy 2007; Burnett 2016). So-called minimum standard AGAs such as *dirty*, *spotted*, *wet*, *bumpy*, and *bent* describe the (minimal) existence of a quality such as dirt, spots, or wetness, and so-called maximum standard AGAs such as *clean*, *flat*, *dry*, *straight*, and *full* describe the (maximal) lack of that quality (Rotstein & Winter 2004; Kennedy & McNally 2005). Within degree-based frameworks, minimum standard AGAs require their arguments to possess a non-zero degree of the respective property. In the case of *dirty*, this would be some degree of dirtiness. Maximum standard AGAs require their arguments to possess a maximal degree of the relevant property. In the case of *clean*, this would be a maximal degree of cleanliness, i.e., no amount of dirt. Here, determination of the threshold via a CC does not play a role.

Nevertheless, the threshold for AGAs may be affected by the object category to some extent. Consider for instance a child's dirty T-shirt compared to a groom's tuxedo: in case of the T-Shirt, the threshold for the partial adjective *dirty* may deviate from the minimal existence of dirt, while for the tuxedo it will not (Toledo & Sassoon 2011). Similarly, consider a full wine glass, compared to a full gas tank: in case of the glass, the threshold for the total adjective *full* need not be maximal (McNally 2011). Accordingly, deviations from the minimal or the maximal threshold

can exist with AGAs, and theoretical accounts agree that the CC may play some role for the interpretation of AGAs, similar to RGAs.

Notwithstanding these parallels, AGAs and RGAs crucially differ in that context-sensitive interpretations for AGAs are highly restricted (e.g., Burnett 2016). First, the threshold for RGAs, but not for AGAs, can shift for any CC. Imagine a CC consisting of two containers of different sizes, one half-filled and one empty. For this CC, it will be possible to determine the tall and the empty container. If the CC changes, however, such that a bigger container half-filled replaces the smaller empty container, it is still possible to determine the tall one because the threshold for the RGA *tall* can shift. But it is not possible to determine the empty container, because the threshold for the AGA *empty* cannot shift. Moreover, even if the threshold for AGAs shifts, it is still very close to the maximal or minimal value of the respective property. A shirt with one or two small spots of dirt, for example, may be considered clean, but a shirt with more spots of dirt will not. Furthermore, knowledge of the CC is necessary for the interpretation of RGAs, but not for AGAs. For example, the AGA *clean* could apply to any object without dirt, but in order for the RGA *big* to apply to an object, we first need to know which entities we should consider.

Many different theoretical approaches have attempted to account for the differences and similarities of RGAs and AGAs (e.g., degree-based approaches, Rational Speech Act models, trope-based approaches, delineation approaches, see Lasersohn 1999; Kennedy & McNally 2005; Kennedy 2007; Moltmann 2009; Lassiter & Goodman 2013; Qing & Franke 2014; Burnett 2016). The question of which theory of gradable adjectives may be better suited to account for acquisition data is beyond the scope of the present study. The different approaches are important for informing psycholinguistic research about the precise knowledge needed to interpret gradable adjectives, but, in our view, they do not yet lend themselves to testable predictions for acquisition (see also Hacquard 2020; Phillips et al. 2021, on this point). This is because these approaches differ in many ways, but they do not make clearly distinct empirical predictions (see also Burnett 2016). Accordingly, in the present paper we focus on the open question of which contextual information listeners use to determine the threshold, which can shed light on the cognitive representation of the comparison class.

2.3 The acquisition task for the child

Mastering the semantics of gradable adjectives is a complex acquisition task. The child has to recognize that gradable adjectives can occur in various morpho-syntactic environments: they can occur with *for*-PPs, with a noun or without a noun as in *This is big*. She also has to understand that gradable adjectives express orderings among objects along a dimension. Accordingly, the child learner needs to figure out that in the unmarked positive form, gradable adjectives must be interpreted relative to a threshold and that the thresholds for RGAs and AGAs typically differ.

In order to calculate the threshold for RGAs, we as speakers and listeners determine the class of objects used for comparison, that is the CC. Therefore, the child needs to learn which linguistic and non-linguistic cues are available to determine the CC in principle and must decide which one to choose in a given situation. Moreover, the child needs to learn that the threshold for RGAs can vary relative to the CC. As far as AGAs are concerned, the learner must identify the circumstances under which the threshold may deviate from the minimal or maximal value with respect to a specific CC.

3 Previous research on children's interpretation of gradable adjectives

Empirical data is crucial for understanding the nature of the CC: experimental studies can inform us about which contextual cues are actually used to calculate the threshold and can inform theoretical accounts that to date do not spell out how exactly the context provides the threshold. In Section 3.1 and 3.2, we summarize previous research on gradable adjectives addressing non-linguistic cues and linguistic cues, respectively. RGA studies are discussed regarding the effects of different contextual cues for children's interpretations; AGA studies are discussed regarding the question of whether children and adults distinguish them from RGAs. Note that adults are asked directly about which CC they have in mind in a specific context (e.g., Tessler & Goodman 2022), while children's interpretation of gradable adjectives is inferred from their responses to questions or from their object choices.

3.1 Non-linguistic cues

3.1.1 World knowledge

In a series of experiments, Ebeling & Gelman (1988; 1989) investigated the role of world knowledge for determining the CC. Testing *big* and *little*, Ebeling & Gelman (1988) showed that children as young as 2;6 years[4] use their knowledge about typical sizes of objects to decide whether an object is *big* or *little*. Children were presented with a single object at a time, either an unfamiliar object labeled with a fantasy noun (e.g., *wug*), or a familiar object labeled with an existing noun (e.g., *mitten*). Children were not able to provide consistent judgments when asked about the unfamiliar objects (e.g., *See this wug? Is it big or is it little?*), but when asked about the familiar objects, they judged normatively big objects *big* and normatively little objects *little*. This finding indicates that children have stored mental thresholds or "normative" standards for RGAs with respect to specific object categories. A subsequent study by Ebeling & Gelman (1989) found that children aged three also employed "functional" standards for RGAs. When asked to judge an object according to their function (e.g., a hat for a doll), children were able to apply a standard with regard to the intended use of the object.

[4] Children's age is given in the format years;months.

3.1.2 Visual context

Several studies have examined whether children use visual context to interpret RGAs and AGAs (Syrett et al. 2006; Foppolo & Panzeri 2013; Gotowski & Syrett 2020). Children were presented with several objects from the same object category to find out where they locate the threshold, via different methods. In a Scalar Judgement Task, participants had to evaluate whether an object has a property (e.g., *Is this big?*) for each of several objects displaying a linear increase of the adjectival property. Testing RGAs and AGAs, Syrett et al. (2006) report that three-year-olds exhibited a threshold for the RGAs *big* and *long* around the center of the object series. For the minimum AGA *spotted*, children showed a minimal threshold, and for the maximum AGA *full*, some children did not show a maximal threshold, accepting not maximally full containers (see Foppolo & Panzeri 2013 for a similar finding). Using the stimuli from Syrett et al. (2006) in a sorting task (e.g., *Let's put all the big ones over here. […] Let's put all the other ones over here*), Gotowski & Syrett (2020) found a similar pattern for RGAs and AGAs in three- to five-year-old children. Taken together, the findings for these visual setups indicate that by age three, children have different thresholds for RGAs, minimum and maximum AGAs.

3.1.3 Visual context and world knowledge

Several studies have examined the role of visual context and world knowledge together (Smith et al. 1986; Ebeling & Gelman 1988; 1994; Tribushinina 2013). Ebeling & Gelman (1988; 1994) presented two-year-olds with two prototypically big objects of different sizes. When asked to decide whether the smaller of the two objects was big or little (e.g., *Is this big or little?*) children labeled this object more frequently as *little*. Similarly, Tribushinina (2013) reported that four-year-old children used the visual context, supplied via pictures, to categorize objects as being big and small. Notably, the threshold was not affected by the object category, i.e., prototypical size of the objects (see Smith et al. 1986, for a similar finding regarding *high/low*). Overall, these studies suggest that children between the ages of two and four prefer the information provided by the visual context to world knowledge when calculating the threshold for RGAs.

3.1.4 Changes in the visual context

A further line of research investigated whether the threshold for gradable adjectives shifts when the visual context changes (Smith et al. 1986; Ebeling & Gelman 1994; Barner & Snedeker 2008; Syrett et al. 2010; Booij & Sassoon 2014). In these studies, children saw either pairs of objects (Ebeling & Gelman 1994; Syrett et al. 2010) or several objects (Smith et al. 1986; Barner & Snedeker 2008) that differed regarding a specific property, such as size.

Providing children with a choice between two objects, Ebeling & Gelman (1994) showed two- to four-year-olds the same test object in two visual contexts, together with a smaller or with

a bigger object. When asked *Give me the big one*, children selected the test object above chance in the condition with the smaller, but not with the bigger object. The authors concluded that two- to four-year-olds are able to adjust their threshold for RGAs depending on the visual context in which an object was presented. Using a similar task with RGAs (*big, long*) and AGAs (minimum: *spotted, bumpy;* maximum: *full, straight*), Syrett et al. (2010) found that children as young as age three accepted requests with RGAs in contexts that differed regarding the absolute size or length of the objects. This finding indicates that children are aware that RGAs denote relative properties, i.e., that an object is always big or long relative to another object. In the case of AGAs, rejections were expected when the two objects violated either the uniqueness or the existence presupposition of the definite determiner. That is, either both objects (e.g., two maximally full containers) or none of the objects (e.g., two filled but not maximally full containers) could be described by the adjective. Beginning at age three, children performed target-like for *spotted* and *bumpy*, which suggests that the visual context did not affect the threshold for minimum AGAs. Children accepted the request for maximum AGAs, however, when the existence presupposition was violated, which suggests that children's threshold can deviate from the maximal amount of the respective property. Providing more than two objects in the visual context, Barner & Snedeker (2008) (henceforth B&S) tested children's interpretation of the RGAs *tall* and *short* in two different scenarios. One group of four-year-olds saw nine fantasy objects (*pimwits*) of the same appearance, except for their height (Experiment 1). In Experiment 2, a second group of four-year-olds saw the same setup, but with an additional four shorter/taller *pimwits*. The test prompt was: *Can you look at all of the pimwits and find the short/tall pimwits?*. The rationale was the following: the addition of shorter/taller objects to objects that already differ in height results in a change of the distribution of height such that the mean height increases or decreases, respectively. Their results indicate that four-year-olds exhibit consistently different thresholds for *tall* when the mean height of objects changed, while for *short* there was variation. Using a similar manipulation of the visual context, Smith et al. (1986) tested the RGAs *high* and *low* with real-world objects in a series of five experiments. Different from B&S, they found that four-year-olds did not shift their judgements. In summary, evidence to date is mixed regarding the age at which children are able to calculate a threshold that is sensitive to the visual context.

3.2 Linguistic cues

Our understanding of children's interpretation of gradable adjectives has been informed by a large body of research on different non-linguistic cues for determining the CC. Notably, the test prompts in the studies discussed above typically referred to the objects in question without using a noun label (e.g., *Give me the big one!*). The only study to examine this linguistic cue for adjective interpretation is B&S (see Section 3.1): they manipulated the modified noun in addition to the visual context using fantasy names and fantasy objects. Two groups of four-year-old children saw the

same set of objects: nine unknown objects, which looked identical except for their height, and four shorter unknown objects, which looked different from the nine objects. Group 1 (Experiment 3) learned that the nine objects were *pimwits* and the four objects were *tulvers*; Group 2 (Experiment 4) learned that all 13 objects were *pimwits*; their task was to select *the tall pimwits* (*Can you look at all of the pimwits and find the tall pimwits?*). Children's object choices indicated a higher threshold for Group 1 than for Group 2. When calculating the threshold, Group 1 did not consider objects present in the visual context but not labeled by the noun, whereas Group 2 considered all objects labeled by the noun despite their perceptual difference. B&S concluded that four-year-old children determine the CC based on linguistic information, i.e. the names given to the objects, rather than based on perceptual categories. Their evidence pointing to a preference for linguistic information is based on the specific setup of the task, namely that the set of fantasy objects was "linguistically partitioned" into two in Group 1, but not in Group 2. Due to the use of fantasy names denoting fantasy objects, two issues remain unresolved. First, it is open how children determine the CC for lexical nouns that denote real-world objects. Different from fantasy objects, existing objects are part of given natural taxonomic classes, which exhibit sub- and superset relations that we can refer to with basic- or superordinate-level nouns. Second, for fantasy objects, categorization is more variable; hence, it is unclear how children interpreted the linguistic information they received with respect to categorization and CC construction. More specifically, for Group 1, it is open whether children interpreted the labels *pimwits* and *tulvers* as referring to two taxonomic categories that belong to the same or to different superordinate categories. Children from Group 2 may have understood *pimwits* as a basic-level or as a superordinate-level noun. The present study addresses these issues by extending the B&S design to existing objects and existing nouns. The use of existing objects and their taxonomic labels can help uncover the role of the noun in children's creation of the adjective's comparison class.

4 The present study

4.1 Goals and rationale

Our study aims to contribute to the fundamental question of how listeners determine the comparison class for gradable adjectives, via the modified head noun and/or via the visual context, and whether children use the same cues as adults. Nouns play a crucial role for adjective learning in general. Studies on adjective acquisition showed that children were more successful in learning the novel meaning of fantasy adjectives such as *blikish*, when the noun phrase contained a lexically specified head noun (e.g., *horse*) compared to indefinite pronouns (*one*) or unspecific nouns (*thing*) (e.g., Klibanoff & Waxmann 2000; Mintz & Gleitman 2002). This suggests that knowledge about category membership helps in identifying the property described by adjectives. For learning the meaning of RGAs, access to category membership is especially important. Bigness, for example, cannot be interpreted without a relation to some category, unlike cleanliness, which

may be evaluated in the absence of any category information. In the current study, we examine the role of the nature of the modified noun in determining the comparison class. Any given object can be a member of different taxonomic classes. Take a mouse: it belongs to the classes 'mouse', 'mammal', 'animal', among others. These classes are hierarchically structured and exhibit subset-superset inclusion relations: being a member of the category 'mouse', for example, entails membership in the category 'animal'. The respective taxonomic category is encoded by a noun. The taxonomic category affects the interpretation of the adjective. Imagine you see a mouse of substantial size: if you construe the referent as a mouse, you may call it *big*, but if you construe it as an animal, you would most likely not do so. Notably, relative gradable adjectives, such as *big*, can revoke the inclusion relation, e.g., being a big mouse does not entail being a big animal.

In short, the category labeled by the noun is a salient candidate for the CC. At the same time, visual cues can serve to determine the threshold (see Section 2). The current study investigates which source of information children and adults privilege when both sources of information are available. We use existing nouns denoting basic-level categories (e.g., *water balloons*) and the corresponding superordinate-level category (*toys*) to address the unresolved question of how the taxonomic label affects the creation of the CC, and how non-linguistic, in our case visual, cues affect the determination of the CC. Each adjective was presented in three conditions, varying the taxonomic label (basic-level, superordinate-level head noun) or the visual setup (one or two basic-level categories), or both, the taxonomic label and the visual setup. This design allows us to test children's interpretation of the same linguistic description for different visual arrays as well as children's interpretation of different linguistic descriptions for the same visual array (different from B&S, see Section 4.3 for details).

The use of existing basic-level and superordinate-level head nouns has several consequences: first, different from fantasy objects and their names, we do not need to teach the participants that objects belong to the same category or to different categories; we can immediately observe whether participants make use of the nominal label to establish category membership and a corresponding CC. Second, the fact that existing objects are part of a given natural taxonomic hierarchy allows us to test how this categorization and the respective labels affect the creation of the CC; this question cannot be addressed using fantasy objects. Consider a visual array with two types of objects that are part of a given natural taxonomic hierarchy, such as water balloons and soccer balls of different sizes, which both belong to the same superordinate-level category 'toys'. It may be that soccer balls are more likely to affect judgements for big water balloons, compared to tulvers affecting the judgements for tall pimwits, for which their taxonomic categories are under-determined.

The non-linguistic information of interest for the present study is the visual context, i.e. cards depicting objects from one or two types of toys. Note that using existing objects automatically

introduces the factor world-knowledge, which can be regarded as a further dimension of the non-linguistic context (see Section 2.1). We controlled for the factor world knowledge by keeping the superordinate-level category 'toys' constant throughout the experiment (see Section 4.3).

An adult group was included in the current study to find out which information (taxonomic label and/or visual setup) is privileged in adults and whether adults and children exhibit the same preferences. Adults' CCs have been shown to be sensitive to category knowledge and to the visual context (Schmidt et al. 2009; Solt & Gotzner 2012; Qing & Franke 2014; Tessler & Goodman 2022). Therefore, it is possible that they can use cues to the CC more flexibly than children.

As a starting point, we wish to substantiate the observation that children exhibit different thresholds for AGAs and RGAs, using a picture-choice task and presenting pictures of objects, which to date has not been used to study gradable adjectives. Physical objects have been used in Scalar Judgement tasks by Syrett et al. (2006) and Foppolo & Panzeri (2013), and in the sorting task by Gotowski & Syrett (2020). If our results for Q1 and Q2 below confirm previous findings, we can conclude that children's ability to distinguish between RGAs and AGAs is unaffected by the specific task and by the mode of presentation (pictures vs. objects), that is, by preschool age the distinction between RGAs and AGAs is firmly established. Based on these results, we address Q3 and Q4 below, which are central to the current study.

As for the child's acquisition task (see Section 2.3), our study addresses the following research questions:

Q1: Do children have different thresholds for AGAs and RGAs in a picture-choice task?
Q2: For which type of gradable adjectives (RGAs and/or AGAs) do children shift the threshold when the context changes in a picture-choice task?
Q3: Which contextual changes (taxonomic label, visual setup) trigger a shift of the threshold?
Q4: Are children sensitive to the same contextual changes as adults?

4.2 Participants

Forty-three monolingual German-speaking children and 26 monolingual German-speaking adults participated. Children's age ranged from 3;02 years to 5;09 years (mean = 4;07 years). All children were typically developing as ensured by a standardized language test (SETK 3–5, Grimm 2001). The children were tested at their day-care centers in the Frankfurt am Main metropolitan area after parents were informed about the study and gave their written consent. **Table 1** summarizes the child participants by age group.

The monolingual German-speaking adults were undergraduate students (age range = 19–36 years, mean = 23 years) at Goethe University Frankfurt with little or no linguistic background and received compensation for participation (10 € or course credit); they were tested in a quiet room at the university.

Group	N	Age range (years)	Mean age (years)
3-year-olds	11	3;02–3;11	3;07
4-year-olds	15	4;01–4;11	4;06
5-year-olds	17	5;00–5;09	5;04

Table 1: Child participants by age group.

4.3 Methods

4.3.1 Materials

Four easy-to-depict adjectives were tested: *groß* 'big', *klein* 'small', *sauber* 'clean', and *dreckig* 'dirty'. We included a positive RGA (*big*) and its negative counterpart (*small*) to control for the role of adjective polarity (see B&S; Pagliarini et al. 2022). A maximum AGA (*clean*) and its minimum counterpart (*dirty*) were included to examine whether interpretation patterns differ for minimum and maximum AGAs (see Syrett et al. 2010).

Participants saw sets of picture cards; each picture card had the same size (14 x 14 cm) and depicted one object from the category 'toys'.[5] The object category 'toys' was chosen for all adjectives: as toys often come in different sizes, corresponding to different extents of smallness or bigness, we could create natural transitions between objects. In addition, toys can clearly be depicted as dirty or clean. Importantly, the toys we used as well as their names were likely to be familiar to preschool children.

Each object was of a different color, except for the soccer balls, and showed the property described by the respective adjective to a different extent. All test prompts had the form *Please hand me the* ADJ_{Plural} N_{Plural}, with neutral, non-contrastive intonation.

As shown in **Figure 1** for RGAs and in **Figure 2** for AGAs, each adjective occurred in three conditions, with two items per condition. In all three conditions, participants received linguistic information – via the taxonomic label in the test prompt – and non-linguistic information – via the objects in the visual array – regarding the possible CC. What differed across conditions was the combination of the specific test prompt and the specific visual array. By comparing participants' thresholds across conditions, we could detect which contextual changes were most prominent in inducing a threshold shift: changes in the taxonomic label (keeping the visual array constant), changes in the visual array (keeping the taxonomic label constant), or changes of both, the taxonomic label and the visual array.

[5] The toys we selected are not miniatures of real world entities in the sense in which a toy car resembles a real car. This way, we minimized the influence of a discrepancy between miniature models of objects and the real world objects, which are ordinarily quite large, on children's interpretation of the adjectives (see Huang & Snedeker 2013).

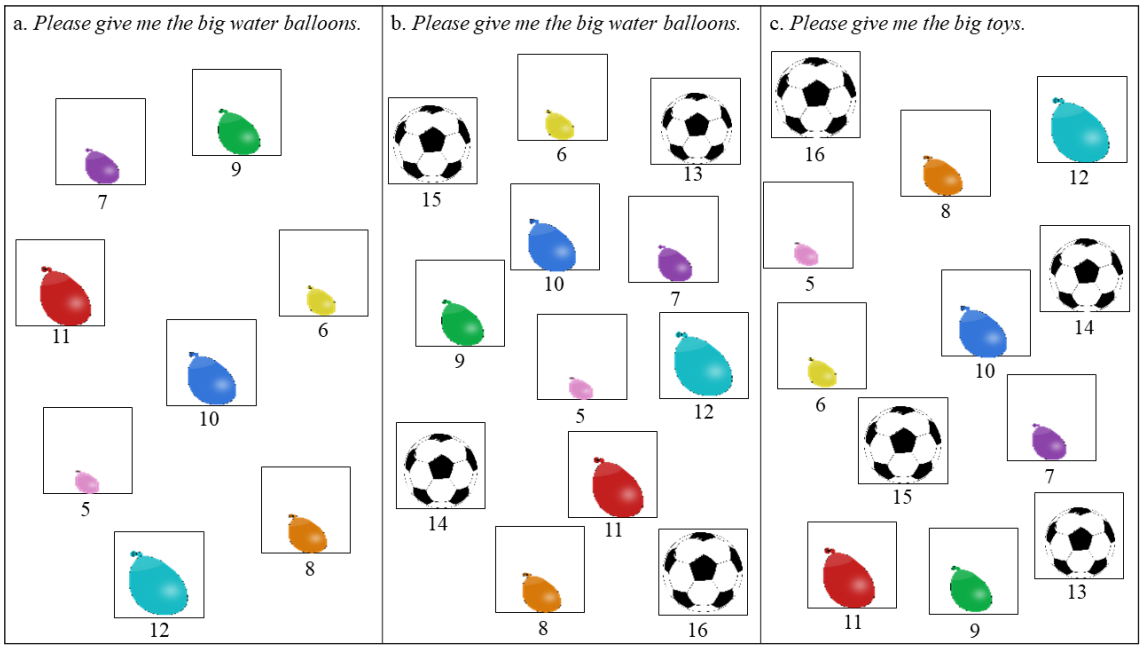

Figure 1: Example items for the RGA *big*. Conditions: a. BASIC_BASELINE, b. BASIC_EXPANSION, c. TOYS_EXPANSION. The numbers were not present in the experiment; they are added for easier reference to the objects in the remainder of the paper.

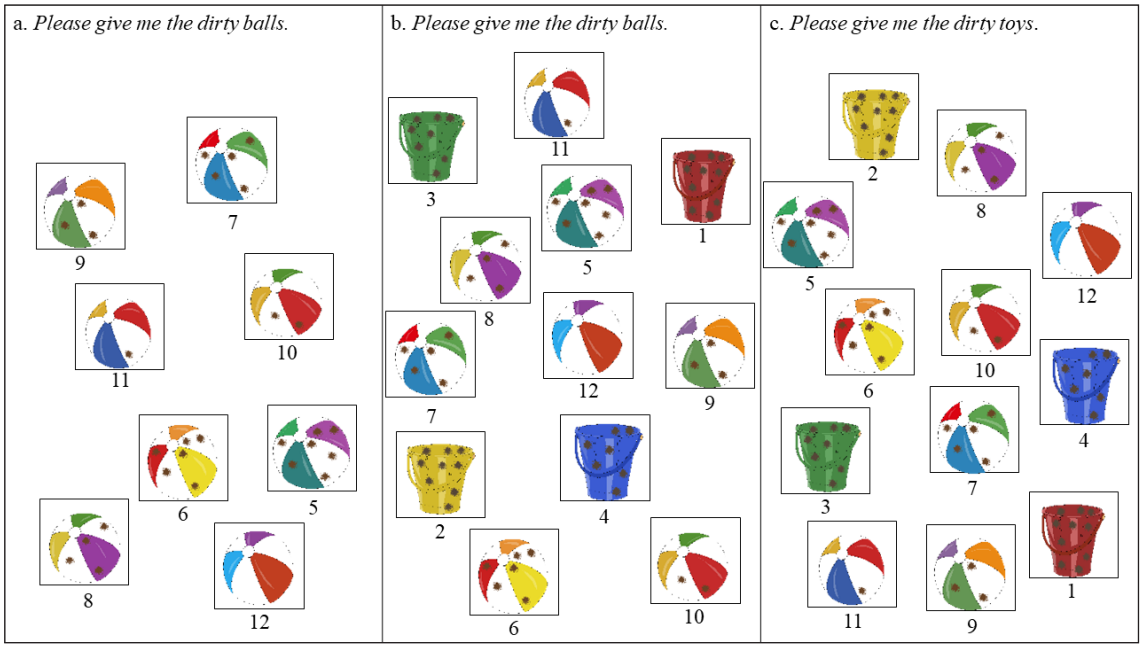

Figure 2: Example items for the AGA *dirty*. Conditions: a. BASIC_BASELINE, b. BASIC_EXPANSION, c. TOYS_EXPANSION. The numbers were not present in the experiment; they are added for easier reference to the objects in the remainder of the paper.

In Condition 1 (BASIC_BASELINE[6], **Figure 1a/2a**), the visual setup consisted of a set of eight toys from one of four basic-level categories: water balloons or hopper balls, all varying in size, and buckets or balls, all varying in dirtiness. We refer to them with numbers 5 to 12, with object 5 being the smallest/dirtiest object, and object 12 being the biggest/cleanest object. The four additional smaller/dirtier objects in BASIC_EXPANSION and TOYS_EXPANSION are labeled 1 to 4; the four additional bigger/clean objects 13 to 16. In the BASIC_BASELINE test prompts, the adjectives modified a basic-level head noun (*Wasserbomben* 'water ballons', *Hüpfbälle* 'hopper balls', *Teddys* 'teddy bears', *Bälle* 'balls').

In Condition 2 (BASIC_EXPANSION, **Figure 1b/2b**), children saw 12 toys from two different basic-level categories: the eight original objects of the BASIC_BASELINE condition and four toys from a different basic-level category (soccer balls, buckets, or dolls). The additional objects expanded the object array either at the upper end of the scale, i.e., by big (**Figure 1**) or clean objects, or at the lower end, i.e., by small or dirty (**Figure 2**) objects. By adding additional objects from a second basic-level category of toys, the distribution of the adjectival property changed compared to the BASIC_BASELINE condition. The noun in the test prompt was a basic-level noun, identical to the BASIC_BASELINE condition.

In Condition 3 (TOYS_EXPANSION, **Figure 1c/2c**), the visual setup was identical to the BASIC_EXPANSION condition, but in the test prompts, the adjectives modified the superordinate-level noun *toys*.

In short, BASIC_BASELINE differed from TOYS_EXPANSION regarding the visual setup and the taxonomic category encoded by the head noun. BASIC_BASELINE differed from BASIC_EXPANSION only regarding the visual setup, and BASIC_EXPANSION differed from TOYS_EXPANSION only regarding the taxonomic category encoded by the head noun.

The visual setups and the nouns used were the same for *big* and *small* and for *clean* and *dirty*, respectively. This design allowed us to compare children's responses for adjectives of different polarity in the same visual context and in the same child. **Table 2** provides the complete list of test items.

Ten filler trials were added in-between test trials to minimize potential influence from the prior test item. Each filler trial consisted of eight picture cards (see **Figure 3**). Similar to the test trials, the picture cards displayed toys. In each trial, objects from two different basic-level categories were shown (buckets, dices, Lego® bricks, soccer balls, books) which differed in shape (round or square) and color (red or blue). The fillers were similar to the test trials, but had different adjectives (*red, blue, round, square*) and were uttered by the child (see Section 4.3.2 for the procedure).

[6] The names of the conditions read as follows: the first part refers to the taxonomic category of the noun and the second part to the type of visual setup.

Condition	Item	Noun in test prompt	Visual setup
BASIC_ BASELINE	big-1/small-1	water balloons	8 water balloons of different size
	big-2/small-2	hopper balls	8 hopper balls of different size
	clean-1/dirty-1	teddies	8 teddies of different dirtiness
	clean-2/dirty-2	balls	8 balls of different dirtiness
BASIC_ EXPANSION	big-1/small-1	water balloons	8 water balloons + 4 bigger soccer balls
	big-2/small-2	hopper balls	8 hopper balls + 4 smaller buckets
	clean-1/dirty-1	teddies	8 teddies + 4 clean buckets
	clean-2/dirty-2	balls	8 balls + 4 dirtier buckets
TOYS_ EXPANSION	big-1/small-1	toys	8 water balloons + 4 bigger soccer balls
	big-2/small-2	toys	8 hopper balls + 4 smaller buckets
	clean-1/dirty-1	toys	8 teddies + 4 clean buckets
	clean-2/dirty-2	toys	8 balls + 4 dirtier buckets

Table 2: Full list of test items.
Note. There are two items per condition for each adjective, referred to by –1 and –2.

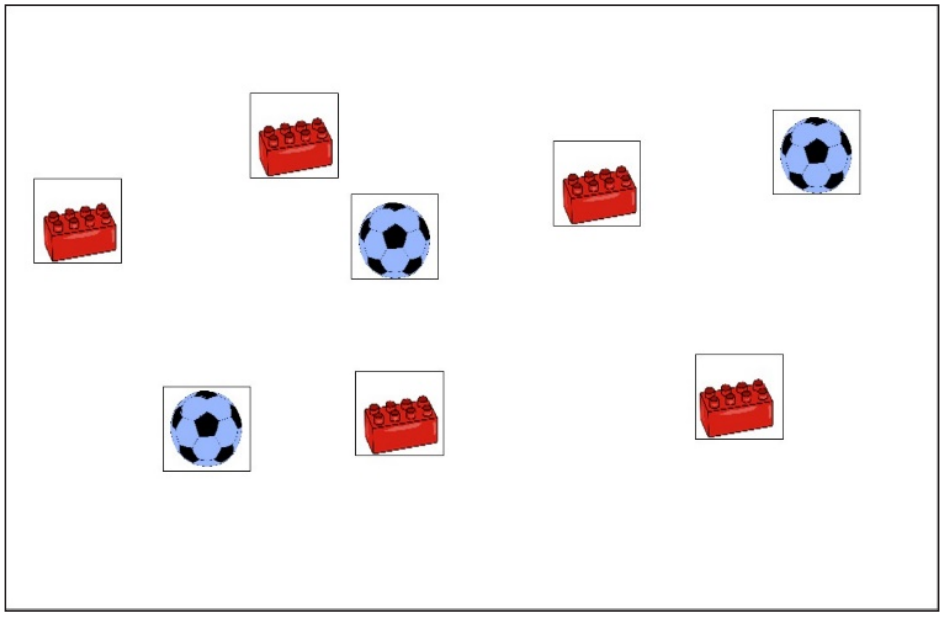

Figure 3: Example filler trial.

The test phase was preceded by a practice phase, where the children were introduced to the objects and their names and to the task. The experimenter first showed pictures of single exemplars of the test objects and asked the child to label them. Then, the child saw all exemplars together and was asked *And what are all these?* to prompt the superordinate-level noun *toys*. Next, the children received three practice trials, where they saw four objects from different basic-level categories and heard a test prompt, which did not contain any adjective (e.g., *Please hand me the dolls*). Following B&S, we opted for plural definite DPs in the practice phase and in the test trials. The definite plural DP presupposes the existence of objects and signals reference to the maximal element of a given set (Link 1983), which can be plural individuals, or in singleton sets, atomic individuals. We assured participants that they could select as many or as few objects as matched the description, and that choosing one object was a licit choice. Participants also received feedback in the practice phase, if they noticed that only one object showed the property: the experimenter explained that the puppet's request was the same independent of the number of objects matching the description. If, in the practice phase, the child forgot to select one or some required objects, the experimenter gave feedback as well.

4.3.2 Procedure

The study was carried out in two sessions (session 1: *big* and *clean;* session 2: *small* and *dirty*). Visual stimuli and order of presentation were the same across sessions to allow for a comparison of the antonymous adjectives. To minimize possible influence from choices in session 1 on performance in session 2, the two sessions were about twelve days apart.

To make the task engaging for the child, we implemented the experiment as a game: a puppet played by the experimenter and the child took turns selecting picture cards. The children sat next to the experimenter on the floor or at a table large enough to show all picture cards at the same time. At the beginning, the experimenter explained that the puppet wanted to play a game. The puppet and the child each received a die, with arbitrary patterns for the puppet and with a square, a circle, a red dot, and a blue dot for the child. In the test trials, the puppet made a request and the child had to select the matching picture cards; in the filler trials, the roles were reversed. At the beginning of each trial, the experimenter distributed the picture cards. Importantly, the picture cards were presented in random order (see **Figure 4**). This mode of presentation does not suggest a specific relation between the objects to the participants. Instead, participants need to establish their own ordering, silently, verbally, or by moving the cards, before they can respond to requests such as *Please hand me the big water balloons*. We believe that this is a close proxy to real-world scenarios, since relationships between objects in reality are rarely as transparent as they appear in an ordered series of objects.

Figure 4: Example presentation of visual stimuli.

On the puppet's turns, the puppet rolled her die without letting the child see the outcome and made her request, which corresponded to the test trials. In the child's turns, the child rolled her die. When the die showed *blue*, for example, the child had to ask the puppet to hand her the blue toys. The puppet sometimes chose the correct and sometimes the incorrect items. The participant's (and the puppet's) task was to select those objects that in their opinion matched the request; the experimenter emphasized that the person selecting the right picture cards more often would win the game.

4.3.3 Data analysis and expected responses

4.3.3.1 Data analysis

Numbers are used to refer to the objects. Recall that the eight objects in BASIC_BASELINE are labeled 5 to 12, with object 5 being the smallest/dirtiest object, and object 12 being the biggest/cleanest object. The four additional smaller/dirtier objects in BASIC_EXPANSION and TOYS_EXPANSION are labeled 1 to 4; the four additional bigger/clean objects 13 to 16.

Participants' responses were analyzed in three steps. First, for each trial we analyzed which objects were selected. In a next step we determined for each trial the participants' individual 'cutoff point' (henceforth: cutoff), which marks the rank of the transition between entities for which an adjectival property is true and for which it is not. For *big* this was the smallest

object considered as being big (together with all bigger objects), for *small* it was the biggest object considered as being small (together with all smaller objects). For *clean* it was the dirtiest object considered as being clean (together with all cleaner objects), and for *dirty* it was the cleanest object considered as being dirty (together with all dirtier objects). To illustrate, if a child selected object 10, 11, and 12 when asked for *big water balloons* in the scenario of **Figure 1**, the cutoff is 10, marking the rank of transition between big and non-big objects. If participants missed an object within a series after the cutoff, this response was classified as unanalyzable.[7] Proportion of choices for each object in an array and the cutoff were taken to express participants' threshold.

In the third step of the analysis, we compared the cutoffs of each condition in a pairwise manner (BASIC_BASELINE – TOYS_EXPANSION, BASIC_BASELINE – BASIC_EXPANSION, BASIC_EXPANSION – TOYS_EXPANSION) for each item (*big*-1, *big*-2 etc.; see **Table 2**) and analyzed whether the cutoffs shifted. Unlike B&S, who compared mean cutoffs, our within-subjects design allowed us to identify individual shifts in the cutoff. This way we could uncover which of the contextual changes (visual setup + taxonomic label, taxonomic label, or visual setup) more often caused participants to shift their thresholds. Consequently, we can answer the question of whether the CC is preferably determined by linguistic or by non-linguistic visual cues.

Based on the third step of the analysis, a dependent variable *cutoff shift* was created with values classified as yes, no, other, or undefined. We illustrate each type with example cutoffs for item *big*-1 and the contextual change 'visual setup + taxonomic label', i.e. a pairwise comparison of the conditions BASIC_BASELINE and TOYS_EXPANSION (see **Figure 1a** and **1c**): a cutoff shift was present, if the cutoff was shifted in the expected direction given the distribution of size in the two conditions, e.g., when the cutoff is 9 in BASIC_BASELINE and 11 in TOYS_EXPANSION (the cutoff in TOYS_EXPANSION should be higher because bigger objects were added).[8] No cutoff shift occurred when the cutoffs were identical across conditions, e.g., 9 in BASIC_BASELINE and in TOYS_EXPANSION. *Other responses* were those that showed a shift, but given the distribution of size in the two conditions, in the unexpected direction, e.g., cutoff 9 in BASIC_BASELINE and cutoff 8 in TOYS_EXPANSION. A shift was coded as *undefined* when the cutoff for one of the conditions in the comparison was unanalyzable (see above).

---

[7] In total, the cutoff point was unanalyzable in 5.9% of responses (17/552 in BASIC_BASELINE, 53/552 in BASIC_EXPANSION, 27/552 in TOYS_EXPANSION).

[8] Few of the cutoff shifts (12%) included object choices specific to TOYS_EXPANSION. Here, participants occasionally opted for two separate cutoffs, one for each basic-level category. For example, the cutoff in BASIC_BASELINE was 9 and the cutoffs in TOYS_EXPANSION were 9 (for water balloons) and 15 (for soccer balls).

4.3.3.2 Expected responses

To address question (Q1), we considered participants' cutoffs in the BASIC_BASELINE trials. Under the assumption that AGAs describe the maximal or minimal value of a property, we expected the threshold for *clean* to be the maximal lack of dirtiness (= object 12) and the threshold for *dirty* the minimal existence of dirtiness (= object 11). For the RGAs *big* and *small,* we expected thresholds around the center of the object series.

To answer research question (Q2), we analyzed to what extent the cutoff shifted according to a contextual change (visual setup, taxonomic label, visual setup + taxonomic label). Following the assumption that the interpretation of AGAs is not context-dependent (*modulo* cases of potential deviance from a minimal or maximal threshold, see Section 2.2), we expected the same cutoff across all three conditions for AGAs. The cutoffs of RGAs should shift between conditions, as their interpretation is context-dependent.

As for our main research question (Q3), we compared the frequency of cutoff shifts for each type of contextual change (visual setup + taxonomic label, taxonomic label, visual setup). For RGAs, we expect a shift of the cutoff when both, visual setup and taxonomic label, change, i.e., we contrast BASIC_BASELINE and TOYS_EXPANSION. That is, participants' object choices should differ in BASIC_BASELINE and in TOYS_EXPANSION: for example, participants may select water balloons 10, 11, 12 in BASIC_BASELINE (**Figure 1a**) and only water balloon 12 together with all the bigger soccer balls in TOYS_EXPANSION (**Figure 1c**). We expect the same number of cutoff shifts when only the taxonomic category of the noun changes (i.e., contrasting BASIC_EXPANSION, **Figure 1b**, and TOYS_EXPANSION, **Figure 1c**), and we predict significantly fewer cutoff shifts when only the visual setup changes (i.e., contrasting BASIC_BASELINE and BASIC_EXPANSION), under the assumption that comparison classes are preferably based on linguistic cues. In the latter case, participants should select the same objects, e.g. water balloons, in BASIC_BASELINE (**Figure 1a**) and BASIC_EXPANSION (**Figure 1b**), if they determine the CC based on the taxonomic label of the noun and ignore the visual presence of the bigger soccer balls. In case the cutoff is shifted, it should be higher when objects are added at the upper end of the object array, and it should be lower when objects are added at the lower end of the object array.

For AGAs, we do not expect participants to shift the threshold: for instance, participants should consider the same balls dirty across conditions (**Figure 2a–c**) and should select the additional dirty buckets in TOYS_EXPANSION (**Figure 2c**).

To answer (Q4), we analyzed the adult data in a parallel way. Given that adults are proficient speakers with rich linguistic and non-linguistic knowledge, it may be that they use the different cues to the CC more flexibly than children do.

4.4 Results

The analysis of the thresholds in the BASIC_BASELINE condition is reported in Section 4.4.1, addressing (Q1). The analyses of the cutoff shifts, for children and for adults, are reported in Section 4.4.2, addressing (Q2), (Q3) and (Q4).

4.4.1 Thresholds in BASIC_BASELINE

We first analyzed children's individual object choices. **Figure 5** displays how often children selected the objects 5 to 12 in the AGA trials (for the numbering see **Figure 2a**). In the two *clean* trials, the object without dirty spots (object 12) was selected in almost all cases (with one exception by a 3-year-old). Choices for objects with some dirty spots are attributable to a few individual children: four children (from all age groups) sometimes also treated object 11, with one spot of dirt, as being clean; and six children (aged 3 and 4) selected all objects in one or both of the *clean* trials. In the two *dirty* trials, the objects with at least one dirty spot (object 5–11) were selected very frequently. Two children (aged 3 and 5) did not select all objects with dirt; one child in each age group selected all objects as dirty, including object 12, which had no dirty spots.

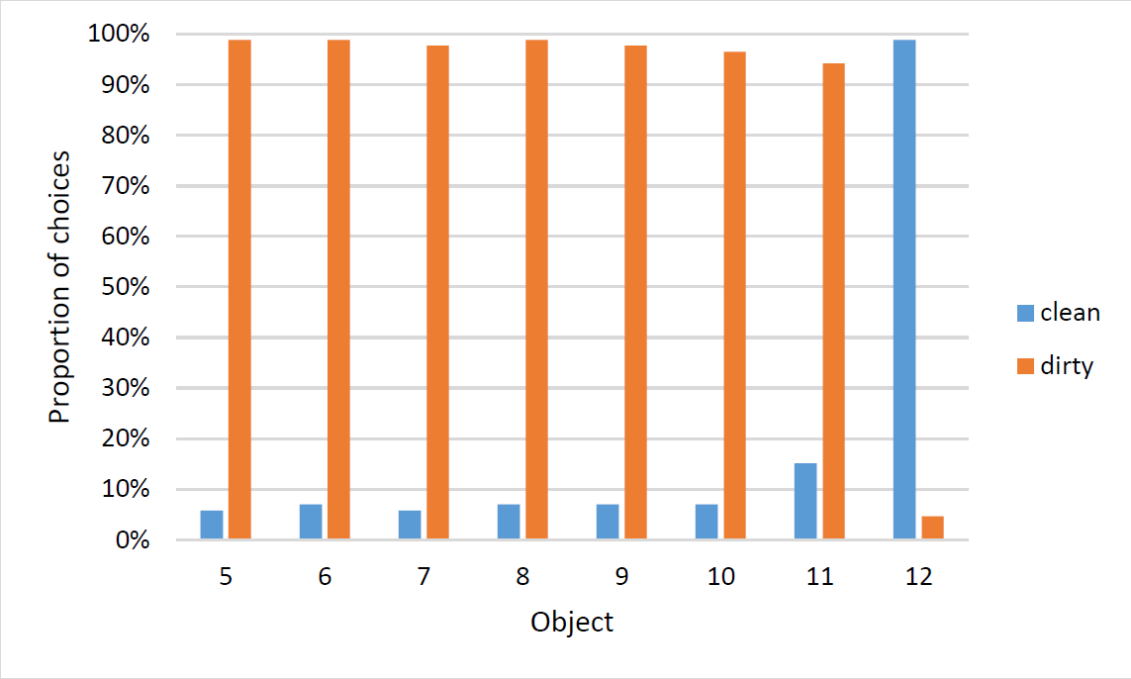

Figure 5: Children's object choices for *clean* and *dirty* in BASIC_BASELINE (5 = dirtiest object, 12 = cleanest object).

Figure 6 displays how often children selected the objects 5 to 12 in the RGA trials (for the numbering see **Figure 1a**). In the two *big* trials, the biggest object was selected in all cases except one (by a 3-year-old). The proportion of choices increased around the center of the scale. In the two *small* trials, the smallest object was selected in all cases except one (by a 3-year-old); the proportion of choices decreased around the center of the scale.

We now turn to the analysis of the cutoffs in BASIC_BASELINE, which is based on 86 observations per adjective (two trials per child). **Table 3** lists for each of the adjectives the median cutoff together with the cutoff that occurred most often (mode) and the range.

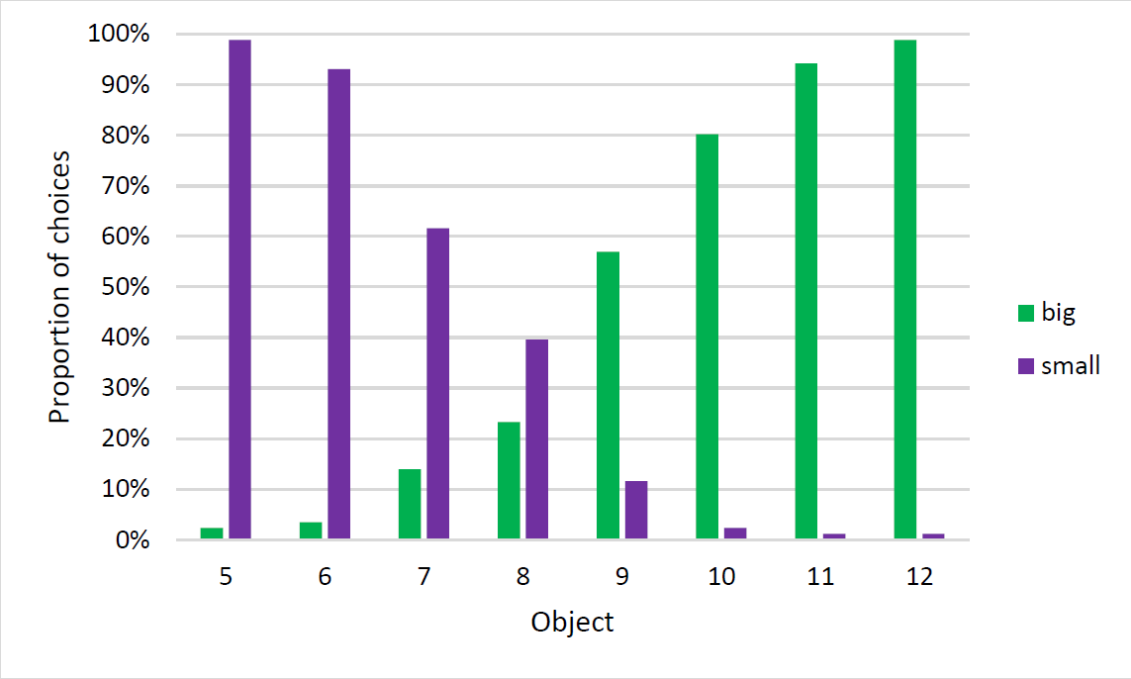

Figure 6: Children's object choices for *big* and *small* in BASIC_BASELINE (5 = smallest object, 12 = biggest object).

Adjective	Median	Mode	Minimum	Maximum
Big	9	9	5	12
Small	7	6	5	10
Clean	12	12	5	12
Dirty	11	11	5	12

Table 3: Children's cutoff: median, mode, minimum, and maximum per adjective.

We analyzed whether children's behavior differed depending on the age group. Notably, the cutoffs did not differ between the three-, four- and five-year-olds (Kruskal-Wallis test; *clean*: $X^2(2) = 4.018$, $p = .134$; *dirty*: $X^2(2) = 2.878$, $p = .237$; *big*: $X^2(2) = 4.709$, $p = .095$, *small*: $X^2(2) = 1.411$, $p = .494$). Within each age group, the cutoffs for *clean* and *dirty* differed significantly (Wilcoxon test; 3-year-olds: $z = -3.638$, $p < .001$, $r = 0.6$, $n = 18$; 4-year-olds: $z = -3.170$, $p = .002$, $r = 0.4$, $n = 29$; 5-year-olds: $z = -5.568$, $p < .001$, $r = 0.7$, $n = 33$), with strong effect sizes for the 3- and 5-year olds and a moderate effect size for the 4-year-olds. The cutoffs for *big* and *small* differed significantly within each age group as well (Wilcoxon test; 3-year-olds: $z = -3.638$, $p < .001$, $r = 0.6$, $n = 18$; 4-year-olds: $z = -3.170$, $p = .002$, $r = 0.4$, $n = 29$; 5-year-olds: $z = -5.568$, $p < .001$, $r = 0.7$, $n = 33$), with strong effect sizes for the 3- and 5-year olds and a moderate effect size for the 4-year-olds.

Overall, the proportion of object choices and the resulting cutoff points in the BASIC_BASELINE condition indicate a threshold around the center of the scale for *big* and *small*, a maximal threshold for *clean*, and a minimal threshold for *dirty*. These results are very similar to the adult data; a detailed description is provided as Supplementary Material.

4.4.2 Shift of cutoff points

For our main analysis, we examined whether children's and adults' cutoffs were sensitive to changes in the context. Comparing the cutoffs across the three conditions (BASIC_BASELINE, BASIC_EXPANSION, TOYS_EXPANSION) in a pairwise manner, we investigated which of the contextual changes (visual setup + taxonomic label, taxonomic label, visual setup) trigger a shift of the cutoff. We first report the results for the children and then the results for the adults.

4.4.2.1 Children

For each item and each adjective, we analyzed three contextual changes (visual setup + taxonomic label, taxonomic label, visual setup). In total, there are 1,032 observations, 516 for AGAs (*clean*, *dirty*) and 516 for RGAs (*big, small*). Recall that shifts of the cutoff are expected for RGAs, but not for AGAs. As far as AGAs are concerned, the cutoff was shifted in only 5.8% (30/516) of the cases. In 79.1% (408/516) of the cases the cutoff was not shifted; 13.3% (69/516) of the responses were classified as undefined, because the cutoff in one of the conditions could not be determined (see footnote 7); finally, 1.7% (9/516) of the responses were other shifts. As for RGAs, the cutoff was shifted in 44.8% (231/516) of the cases. In 23.6% (122/516) of the cases the cutoff was not shifted; 15.1% (78/516) of the responses were shifts in the unexpected direction, and 16.5% (85/516) of the cases were classified as undefined. Undefined responses were excluded from further analyses.

Recall that for RGAs we expected different shifting patterns depending on the type of contextual change. A shift of the threshold should occur when both the taxonomic category encoded by the noun and the properties of the visual setup changed. Under the assumption that the CC is preferably determined based on linguistic cues and that the threshold is calculated accordingly, a shift of the cutoff should occur when only the taxonomic category encoded by the noun changed, but less often when only the visual setup changed. Therefore, we inspected the shifts for RGAs for each type of contextual change. The results are shown in **Figure 7**; the width of the violin plots varies by the density of data points in a specific region. The frequency of shifts varied with type of contextual change. The violins for two of the three conditions (visual setup + taxonomic label, taxonomic label) for both *big* and *small* resemble each other and differ from the violins for the visual setup condition. These latter show wider parts in the middle and in the lower part, indicating that the majority of children rarely shifted the threshold; they did so in about half of the cases or almost never (*big*: mean = 33.3%, SD = 36.8%; *small*: 27.3%, SD = 29.8%). The former violins show wider parts in the middle and in the upper part, indicating that children shifted the threshold more frequently, in about half of the cases or nearly always. Children shifted the threshold for the RGAs *big* and *small* most frequently, when both the taxonomic label

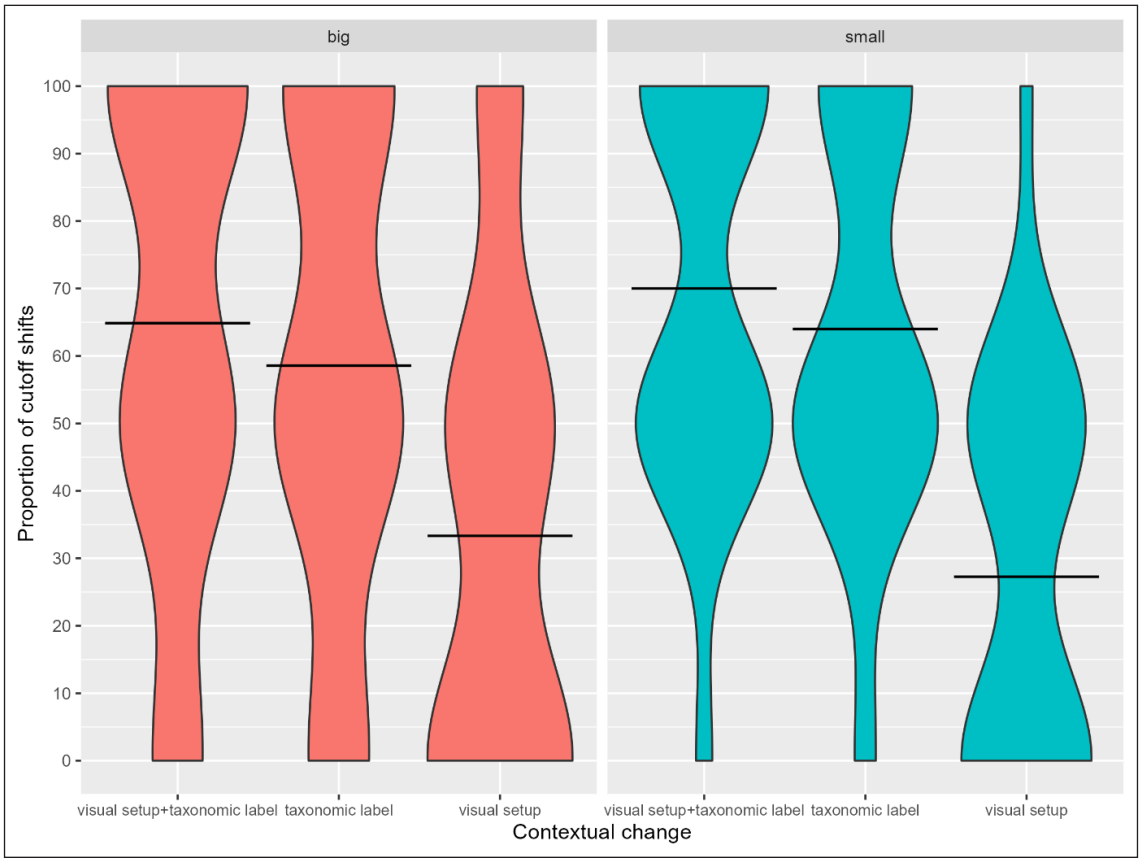

Figure 7: Frequency of cutoff shifts for the relative gradable adjectives by adjective and type of contextual change in the child group. The black line indicates the mean.

and the visual setup changed (*big*: mean = 64.9%, SD = 37.0%; *small*: mean = 70.0%, SD = 30.2%). Descriptively, the threshold shifted slightly less often when only the taxonomic label changed (*big*: mean = 58.6%, SD = 37.3%; *small*: mean = 64.0%, SD = 30.7%).

The shifting data for RGAs were analyzed with generalized mixed effects logistic regression (1 = shift-yes, 0 = shift-no or shift-other); we used the *glmer* function in the *lme4* package (Bates et al. 2015) in R (R Core Team 2022). The shifting data were fitted into a model using maximum-likelihood estimation (Laplace Approximation). Contextual change (visual setup + taxonomic label, taxonomic label, visual setup) was entered as a fixed effect, using treatment coding and *visual setup + taxonomic label* as the reference level, with a random intercept for Participant.[9] The model revealed a highly significant difference between *visual setup + taxonomic label* and *visual setup* ($\beta = -1.4904$, $Z = -5.774$, $p < .001$), but not between *visual setup + taxonomic label* and *taxonomic label* ($\beta = -0.3598$, $Z = -1.456$, $p = .145$). Further inspection via pairwise comparisons with Tukey adjustment (*emmeans* package, Lenth 2022), showed a significant difference between *visual setup + taxonomic label* and *visual setup* ($p < .001$), and between *visual setup* and *taxonomic label* ($p < .001$), but not between *visual setup + taxonomic label* and *taxonomic label* ($p = .312$). Put differently, changes in both visual setup and taxonomic label, and changes in the taxonomic category of the noun trigger a shift of children's threshold for RGAs more often than changes in the visual setup.

A second model was built adding Adjective (*big, small*) as fixed effect. This model did not improve the fit of the data compared to the simpler model (AIC = 563.35 vs. 564.96), and the difference in fit from the model without Adjective as a fixed effect was not significant ($X^2(1) = 0.3962$, $p = .529$). That is, children's shifts of the cutoff were not influenced by the polarity of the RGA.

4.4.2.2 Adults

Parallel to the child group, we analyzed for each item and each adjective the contextual changes (visual setup + taxonomic label, taxonomic label, visual setup). In total, there are 624 observations by 26 adults, 312 for AGAs (*clean*, *dirty*) and 312 for RGAs (*big, small*). In 99.4% (310/312) of the AGA cases, the threshold was not shifted; a shift occurred only once (0.3%), and one additional shift (0.3%) was classified as *other*. The cutoff for *clean* was consistently the object without dirt (object 12). The cutoff of *dirty* was consistently object 11, i.e., adults selected all objects with at least one spot of dirt, and the object without dirt was never selected.

Overall, the cutoff of RGAs was shifted more often than the cutoff of AGAs, namely in 50.6% (158/312) of the cases. In 35.6% of the cases (111/312), the cutoff did not shift; 11.2% of the cases (35/312) were shifts in the unexpected direction, and 2.6% of the cases (8/312) were classified as undefined. Undefined responses were excluded from further analyses.

[9] We also ran a model with random by-participant slopes, but this resulted in overfitting of the model.

In a next step, we examined the adult shifting pattern more closely to address the question of whether adults and children are sensitive to the same contextual changes (Q4), i.e., whether the shift of the threshold in the adult group is triggered by the same contextual changes as in the child group. The results, displayed in **Figure 8**, are largely similar to the data for the child group. The frequency of shifts varied with type of contextual change; the violins for two of the three conditions (visual setup + taxonomic label, taxonomic label) for both *big* and *small* resemble each other, compared to the violins for the visual setup condition. Adults shifted the threshold for the RGAs *big* and *small* most frequently, when both the taxonomic category encoded by the noun and the visual setup changed (*big*: mean = 67.3%, SD = 31.4%; *small*: mean = 60.0%, SD = 28.9%). Descriptively, the threshold shifted marginally less often when only the taxonomic category encoded by the noun changed (*big*: mean = 58.0%, SD = 31.2%; *small*: mean = 58.3%, SD = 31.9%) and much less often when only the visual setup changed (*big*: mean = 44.0%, SD = 33.3%; *small*: 22.9%, SD = 29.4%).

A model similar to the child group was run for the adult group, with Contextual change as a fixed effect, and a random intercept for Participant. The model revealed that the difference

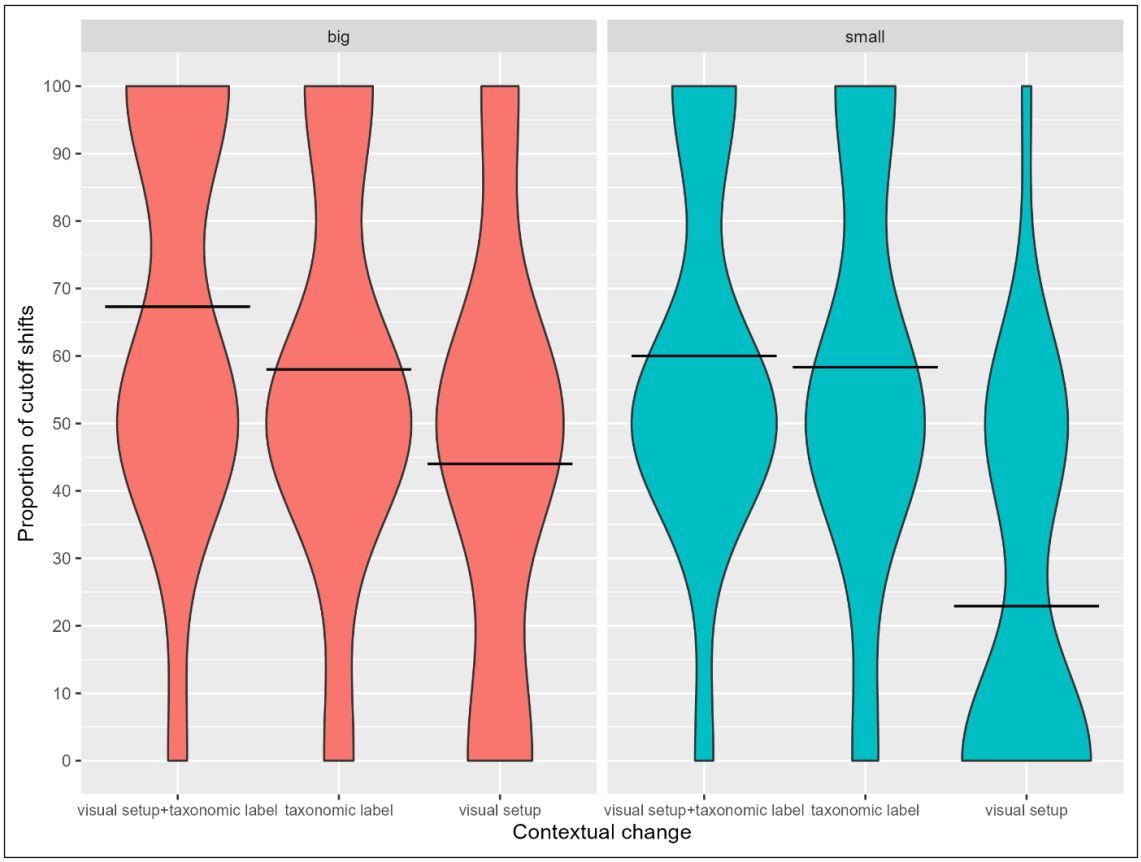

Figure 8: Frequency of cutoff shifts for the relative gradable adjectives by adjective and type of contextual change in the adult group. The black line indicates the mean.

between *visual setup + taxonomic label* and *visual setup* was highly significant ($\beta = -1.261$, $Z = -4.219$, $p < .001$), but the difference between *visual setup + taxonomic label* and *taxonomic label* was not significant ($\beta = -0.2248$, $Z = -0.773$, $p = .439$). Further inspection via pairwise comparisons with Tukey adjustment (*emmeans* package, Lenth 2022), showed a significant difference between *visual setup + taxonomic label* and *visual setup* ($p < .001$), and between *visual setup* and *taxonomic label* ($p = .001$), but not between *visual setup + taxonomic label* and *taxonomic label* ($p = .719$). Similar to the child group, changes in both visual setup and taxonomic label, and changes in the taxonomic category of the noun trigger a shift of adults' threshold for RGAs more often than changes in the visual setup.

A second model was built adding Adjective (*big, small*) as a fixed effect. This model did not improve the fit of the data compared to the simpler model (AIC = 407.37 vs. 406.63), and the difference in fit from the model without Adjective as a fixed effect was not significant ($X^2(1) = 2.7408$, $p = .098$). Adults' cutoff shifts were unaffected by RGA polarity in parallel with children.

5 Discussion

The current study tested children's and adults' interpretation of absolute gradable and relative gradable adjectives (AGA and RGA) to examine how listeners determine the comparison class (CC) and calculate the threshold. We created a picture-choice task that varied the taxonomic category encoded by the modified head noun in the test prompt (basic-level, superordinate-level) and/or the composition of the visual setup (baseline, expansion). Our first finding addresses the question of whether children have different thresholds for AGAs and RGAs (Q1). Our analysis of children's responses in the BASIC_BASELINE condition revealed that already the three-year-olds exhibited different thresholds for AGAs and RGAs. The threshold for RGAs was located around the center of the object series. As for AGAs, children's threshold for the maximum AGA *clean* was oriented towards the maximal lack of dirtiness and the threshold for the minimum AGA *dirty* towards the minimal existence of dirtiness. This pattern is in line with findings from previous studies using Scalar Judgement Tasks and Sorting Tasks with physical objects (see Syrett et al. 2006; Foppolo & Panzeri 2013; Gotowski & Syrett, 2020). Addressing the question of for which type of gradable adjectives (RGAs and/or AGAs) changes in the context result in shifts of the threshold (Q2), we found that children shifted the threshold for RGAs when the context changed, whereas their thresholds for AGAs were not affected by contextual changes. This pattern indicates that by age three children's threshold for RGAs, but not for AGAs, is context-dependent, strengthening the earlier observation that beginning age 3, children are sensitive to the differences between RGAs and AGAs. The main conclusion here is that the ability to distinguish between RGAs and AGAs was unaffected by the specific task (Picture-choice vs. Scalar Judgement/Sorting) and by the mode of presentation (pictures of objects vs. physical objects), that is, by preschool age the distinction between RGAs and AGAs is firmly established.

Our main goal was to investigate which contextual changes trigger a shift of the threshold (Q3). More specifically, we examined whether children use linguistic cues that indicate the taxonomic category (basic-level, superordinate-level), i.e. the modified head noun, for determining the CC and the threshold in the presence of non-linguistic visual cues, i.e. the number of basic-level categories and the distribution of the specific property displayed in the visual setup. Our data provides novel evidence that children preferably use the information provided by the taxonomic category encoded by the modified head noun to calculate the threshold for RGAs. More precisely, children exhibited different thresholds across trials when the taxonomic category encoded by the modified noun changed (e.g, *water balloons* vs. *toys*), and they exhibited the same thresholds when the adjective modified the same basic-level noun. Notably, objects in the visual setup not denoted by the noun affected children's thresholds much less.

Overall, our data reveal that the taxonomic category encoded by the modified head noun plays a more prominent role in determining the CC of RGAs than the composition of the visual setup. By relying on familiar objects that allowed for clear taxonomic categories, this novel finding corroborates Barner & Snedeker's (2008) results, which were based on fantasy objects and fantasy nouns. Regarding question (Q4) of whether children are sensitive to the same contextual changes as adults, we found that the preference for the linguistic cue observed in adults holds for children as well, as shown by their sensitivity to the same contextual changes.

Our findings for RGAs allow us to explore two further questions regarding the role of the modified head noun in the creation of comparison classes, which we discuss in turn below: Why is the linguistic cue via the modified noun preferred to the non-linguistic cue via the visual setup in determining the CC? Why is there a certain amount of variability in terms of the contextual changes that caused the threshold to shift or not to shift?

The first issue relates to the overarching question of the interaction between language and conceptual organization. We propose that linguistic cues may be privileged over non-linguistic cues for three reasons. First, language can influence cognitive processing during mental computations, because it guides attention to aspects of the perceptual environment that are relevant for communication; other not verbalized aspects are deemphasized (see the review in Ünal & Papafragou 2016). Accordingly, the modified head noun may provide a more salient cue than the visual setup, because it articulates the features of the context relevant for determining the CC. Second, it has been found that linguistic labels serve to emphasize categories, especially at the more abstract superordinate level (Waxman & Markow 1995). Assuming that linguistic labels direct listeners' attention to similarities between objects, linguistic information can facilitate the establishment of object categories in a way that non-linguistic information cannot. After all, perceptual similarity is less obvious at the superordinate than at the basic level. Toys, for example, come in very different shapes, materials, composition, colors, and sizes, whereas teddy bears do not. Accordingly, nouns like *toy*, just like *teddy bear*, provide a shared label that initiates the search for commonalities among sets of objects that are otherwise very different. Linguistic labels

have a further important advantage over non-linguistic cues: by labeling a conceptual category, children understand that objects that share a label also share their underlying unobservable properties, e.g., toys share the property of being artifacts that are used to play with (Waxman & Markow, 1995; Fairchild et al. 2018). Put differently, words, in particular nouns, invite the formation of categories. We propose that this property of nouns helps us to create comparison classes. These comparison classes are a prerequisite for further internal categorization, for example when evaluating properties expressed by adjectives.

Let us now turn to the open issue of why there was some variability regarding whether child and adult participants did or did not shift the threshold. On the one hand, the threshold shifted very frequently when the taxonomic category encoded by the noun or both taxonomic label and visual setup changed, but in 43% of the cases it did not. On the other hand, the threshold rarely shifted when only the visual setup changed, but in some cases (28%) it did shift.[10] We propose that both of these findings are related to a specific semantic property of RGAs: vagueness. Vagueness can be understood as uncertainty regarding the threshold, even if the CC is established. Typically, vagueness is described by three characteristics: fuzzy boundaries, the existence of borderline cases, and susceptibility to the Sorites Paradox (e.g., Solt 2015; Burnett 2016, for an overview). The notion of fuzzy boundaries captures the observation that there seem to be no sharp boundaries between entities satisfying a vague adjective (e.g., being big) and entities that do not satisfy it (e.g., not being big). Think of a big car: if we make this car smaller by shrinking its size millimeter by millimeter, it will be impossible to determine the exact size at which this big car turned into a car that no longer counts as being big. Borderline cases are entities for which it is difficult or impossible to judge whether they satisfy the property denoted by the gradable adjective. Consider different types of cars: whereas a truck can be clearly identified as a big car and a Smart car as not big, the judgement for a Honda Civic, for instance, is more difficult. For this borderline case, speakers are inclined to judge it as being neither big nor not big or as both big and not big (Solt & Gotzner 2010; Alxatib & Pelletier 2011; Égré & Zehr 2018). Finally, vague adjectives give rise to the Sorites Paradox,[11] which follows from the lack of sharp boundaries and insensitivity or indifference to small changes in the relevant property (Morzycki 2015; Solt 2015).

We suggest that the relaxed dependency of the threshold on the CC and the resulting flexibility of the threshold can be accounted for in terms of vagueness as follows. Flexible thresholds are a reflex of the uncertainty regarding the threshold due to fuzzy boundaries and the existence of borderline cases. In our experiment the size differences between the single objects were

[10] According to one reviewer, the fact that children's cutoff sometimes shifted when only the visual context changed could also reflect their difficulty with inhibition of the not-mentioned but still visible objects. We have to leave this question for future research, since we did not test children's inhibitory skills.

[11] The Sorites Paradox is illustrated in (i) (see Kennedy 2007: 2):

(i) P1. A $5 cup of coffee is expensive (for a cup of coffee).
P2. Any cup of coffee that costs 1 cent less than an expensive one is expensive (for a cup of coffee).
C. Therefore, any free cup of coffee is expensive.

rather small (8–10 mm) so that the objects were very similar regarding the property encoded by the adjectives (*big, small*). It is possible that participants were sometimes indifferent to small differences regarding size and assumed no precise boundary between big and not-big and between small and not-small objects. Consequently, this boundary may but not need shift from trial to trial, i.e. between CCs. Likewise, variance may exist regarding the objects in the borderline area: a borderline case may or may not be selected across trials. Note that this explanation may also account for the infrequent shifts in the unexpected direction.

How children discover in which situations the threshold for AGAs is susceptible to contextual changes, we leave for future research. We acknowledge that in our experimental task there was little need for tolerating imprecision, i.e., to deviate from the minimal or maximal threshold. At this point, we leave open whether children may opt for non-minimal or non-maximal thresholds for AGAs when reasoning about imprecision is called for (e.g., with specific CCs).

In summary, the current study provides novel empirical evidence that the taxonomic category encoded by the modified head noun is a privileged cue for determining the CC and for calculating the threshold. The creation of CCs accesses cognitive representations of object categories and hierarchies that can be articulated via the modified head noun. If this linguistic information is not available, non-linguistic visual contextual information is considered.

The present study examined the taxonomic category encoded by the noun as a linguistic cue and the visual context as a non-linguistic cue, by systematically varying both factors. More research is needed to better understand to what extent our findings are related to the objects' organization into natural taxonomic hierarchies (i.e. subset-superset relations). As pointed out by one reviewer, this can be tested by extending our design to taxonomically unrelated objects (e.g., water balloons and watermelons). To develop a comprehensive picture of the role of non-linguistic information in the creation of comparison classes, future studies should manipulate other sources of non-linguistic information, such as world knowledge and discourse context. Our study used non-propositional requests, such as *Please hand me the big water balloons*, as did Barner & Snedeker (2008). More research needs to be done to see if what we found extends to propositions, such as *This water balloon is big*, which make an assertion about one entity at a time. Finally, findings by Tessler et al. (2020) for adults suggest that adjective position mediates how the noun contributes to inferring the comparison class: when the adjective was in predicative position, adult listeners were less likely to use the noun as a cue to the comparison class than when the adjective was in attributive position (*This is a big water balloon*). Whether the same applies to children remains to be seen.[12]

[12] In a corpus study, Weicker (2019) found that two-year-old German-learning children use relative gradable adjectives, but not absolute gradable adjectives, more often in attributive than in predicative position, which suggests that children may be sensitive to these syntactic differences also regarding the creation of comparison classes.

6 Conclusion

Mastering the semantics of gradable adjectives is a complex acquisition task. Our study contributes to the question of how children and adults interpret relative and absolute gradable adjectives. We assessed to what extent German-speaking three- to five-year olds and adults use the taxonomic category encoded by the modified head noun (basic-level, superordinate-level) and visual contextual information to determine the comparison class and to calculate the threshold for two relative gradable (*big, small*) and two absolute gradable (*clean, dirty*) adjectives. Using the method of picture-choice, we presented children with an unordered display of existing objects aiming at a close proxy for real-life scenarios. We found that, in this experimental setup, children and adults interpret relative, but not absolute gradable adjectives as context-dependent, thus providing further evidence that central aspects of adjective semantics are mastered by age three. Furthermore, we showed that children are able to establish an ordering among multiple objects also in unordered arrays, organized in basic-level and superordinate-level categories, and to evaluate for which objects a gradable adjective is true. Most importantly, our study is the first to demonstrate for existing objects that, from age three, linguistic information provided by the taxonomic label is privileged over non-linguistic visual information in creating the comparison class of relative gradable adjectives. This finding adds to the notion that language is a powerful tool to modulate our mental computations.

Data availability

Data files, R script, additional experiment materials and results: https://osf.io/hc2p3/?view_only=b42dd5c131d84f38ab93c2c4456021cb.

Ethics and consent

This study was approved by the Research Ethics Board of DIPF Leibniz Institute for Research and Information in Education (#DIPF_EK_2021_30). Informed consent was obtained from each adult participant and the legal guardian of each child participant. All research data has been anonymized.

Funding information

This research was supported by a grant from the German Research Foundation (DFG, GRK 2016/1 "Nominal Modification", PI: Esther Rinke). This study was also part of the project *Vari*, supported by the IDeA – Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany.

Acknowledgements

Thanks to Alex Lowles for creating the pictures. We are grateful to Louise McNally, Thomas Ede Zimmermann, Galit W. Sassoon, Tom Roeper, and three anonymous reviewers as well as Lyn Tieu for their valuable comments and suggestions.

Competing interests

The authors have no competing interest to declare.

References

Alxatib, Sam & Pelletier, Francis J. 2011. The psychology of vagueness: borderline cases and contradictions. *Mind & Language* 26. 287–326. DOI: https://doi.org/10.1111/j.1468-0017.2011.01419.x

Barner, David & Snedeker, Jesse. 2008. Compositionality and statistics in adjective acquisition: 4-year-olds interpret 'tall' and 'short' based on the size distributions of novel noun referents. *Child Development* 79. 594–608. DOI: https://doi.org/10.1111/j.1467-8624.2008.01145.x

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Booij, Elbert J. & Sassoon, Galit W. (2014). Big differences: The standard for 'big' as used by adults and children. In Melnik, Nurit (ed.), *Proceedings of IATL 29. MIT Working Papers in Linguistics* 72. 1–14.

Burnett, Heather. 2016. *Gradability in Natural Language: Logical and Grammatical Foundations*. Oxford Studies in Semantics and Pragmatics 7. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198724797.001.0001

Cresswell, Max J. 1976. The semantics of degree. In Partee, Barbara (ed.), *Montague Grammar*, 61–292. New York: Academic Press. DOI: https://doi.org/10.1016/B978-0-12-545850-4.50015-7

Ebeling, Karen S. & Gelman, Susan A. 1988. Coordination of size standards by young children. *Child Development* 59. 888–869. DOI: https://doi.org/10.2307/1130256

Ebeling, Karen S. & Gelman, Susan A. 1989. Children's use of nonegocentric standards in judgments of functional size. *Child Development* 60. 920–932. DOI: https://doi.org/10.1111/j.1467-8624.1989.tb03524.x

Ebeling, Karen S. & Gelman, Susan A. 1994. Children's use of context in interpreting 'big' and 'little'. *Child Development* 65. 1178–1192. DOI: https://doi.org/10.1111/j.1467-8624.1994.tb00811.x

Égré, Paul & Zehr, Jeremy. 2018. Are gaps preferred to gluts? A closer look at borderline contradictions. In Castroviejo, Elena & McNally, Louise & Sassoon, Galit W. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*, 25–58. Berlin: Springer. DOI: https://doi.org/10.1007/978-3-319-77791-7_2

Fairchild, Sarah & Mathis, Ariel & Papafragou, Anna. 2018. Linguistic cues are privileged over non-linguistic cues in young children's categorization. *Cognitive Development* 48. 167–175. DOI: https://doi.org/10.1016/j.cogdev.2018.08.007

Foppolo, Francesca & Panzeri, Francesca. 2013. Do children know when their room counts as clean? In Kan, Seda & Cantwell, Claire M. & Staubs, Robert (eds.), *Proceedings of NELS 40: Volume 1*, 205–218. Amherst, MA: GLSA.

Gotowski, Megan & Syrett, Kristen. 2020. Investigating the Hypothesis Space for Children's Interpretations of Comparatives. In Brown, Megan M. & Kohut, Alexandra (eds.), *Proceedings of the 44th Boston University Conference on Language Development*, 154–167. Somerville, MA: Cascadilla Press.

Grimm, Hannelore. 2001. *SETK 3–5. Sprachentwicklungstest für 3- bis 5-jährige Kinder*. Göttingen: Hogrefe.

Hacquard, Valentine. 2020. The child in semantics. In Bhatt, Rajesh & Frana, Ilaria & Menéndez-Benito, Paula (eds.), *Making Worlds Accessible. Essays in Honor of Angelika Kratzer*, 32–46.

Huang, Yi Ting & Snedeker, Jesse. 2013. The Use of Lexical and Referential Cues in Children's Online Interpretation of Adjectives. *Developmental Psychology* 49. 1090–1102. DOI: https://doi.org/10.1037/a0029477

Kamp, Hans & Partee, Barbara. 1995. Prototype theory and compositionality. *Cognition* 57. 129–191. DOI: https://doi.org/10.1016/0010-0277(94)00659-9

Kennedy, Christopher. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30. 1–45. DOI: https://doi.org/10.1007/s10988-006-9008-0

Kennedy, Christopher & McNally, Louise. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81. 345–381. DOI: https://doi.org/10.1353/lan.2005.0071

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4. 1–45. DOI: https://doi.org/10.1007/BF00351812

Klibanoff, Raquel S. & Waxmann, Sandra R. 2000. Basic level object categories support the acquisition of novel adjectives: Evidence from preschool-aged children. *Child Development* 71. 649–659. DOI: https://doi.org/10.1111/1467-8624.00173

Lasersohn, Peter. 1999. Pragmatic Halos. *Language* 75. 522–551. DOI: https://doi.org/10.2307/417059

Lassiter, Daniel & Goodman, Noah D. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Snider, Todd (ed.), *Proceedings of SALT 23*, 587–610. Linguistic Society of America. DOI: https://doi.org/10.3765/salt.v23i0.2658

Lenth, Russel V. 2022. *emmeans: Estimated Marginal Means, aka Least-Square means*.

Link, Godehard. 1983. The logical analysis of plurals and mass terms: A Lattice-Theoretic approach. In Portner, Paul & Partee, Barbara H. (eds.), *Formal Semantics – The Essential Readings,* 127–147. New York: Wiley. DOI: https://doi.org/10.1515/9783110852820.302

McNally, Louise. 2011. The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In Nouwen, Rick & van Rooij, Robert & Sauerland, Uli & Schmitz, Hans-Christian (eds.), *Vagueness in Communication*, 151–168. Berlin. Springer. DOI: https://doi.org/10.1007/978-3-642-18446-8_9

Mintz, Toben H. & Gleitman, Lila R. 2002. Adjectives really do modify nouns: the incremental and restricted nature of early adjective acquisition. *Cognition* 84. 267–293. DOI: https://doi.org/10.1016/S0010-0277(02)00047-1

Moltmann, Friederike. (2009). Degree structure as trope structure: a trope-based analysis of positive and comparative adjectives. *Linguistics and Philosophy* 32. 51–94. DOI: https://doi.org/10.1007/s10988-009-9054-5

Morzycki, Marcin. 2015. *Modification*. Cambridge. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511842184

Qing, Ciyang & Franke, Michael. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Snider, Todd & D'Antonio, Sarah & Weigand, Mia (eds.), *Proceedings of SALT 24*, 23–41. Linguistic Society of America. DOI: https://doi.org/10.3765/salt.v24i0.2412

Pagliarini, Elena & Barlassina, Alice & Sanfelici, Emanuela. 2022. The Acquisition of Antonymous Dimensional Adjectives by Italian Preschoolers. In Gong, Ying & Kpogo, Felix (eds.), *Proceedings of the 46th annual Boston University Conference on Language Development*, 615–628. Somerville, MA: Cascadilla Press.

Phillips, Colin & Gaston, Phoebe & Huang, Nick & Muller, Hanna. 2021. Theories all the way down: remarks on "theoretical" and "experimental" linguistics. In Goodall, Grant (ed.), *The Cambridge Handbook of Experimental Syntax*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108569620.023

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rotstein, Carmen & Winter, Yoad. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12. 259–288. DOI: https://doi.org/10.1023/B:NALS.0000034517.56898.9a

Schmidt, Laura A. & Goodman, Noah D. & Barner, David & Tenenbaum, Joshua B. 2009. How tall is 'tall'? Compositionality, statistics, and gradable adjectives. In Taatgen, Niels A. & van Rijn, Hedderik (eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 3151–3156. Austin, TX: Cognitive Science Society.

Smith, Linda B. & Cooney, Nancy J. & McCord, Carol. 1986. What is high? The development of reference points for high and low. *Child Development* 57. 583–602. DOI: https://doi.org/10.1111/j.1467-8624.1986.tb00229.x

Solt, Stephanie. 2015. Vagueness and imprecision: Empirical foundations. *Annual Review of Linguistics* 1. 107–127. DOI: https://doi.org/10.1146/annurev-linguist-030514-125150

Solt, Stephanie & Gotzner, Nicole. 2010. *Expensive, not expensive or cheap? An experimental investigation of vague predicates*. Paper presented at the 11th Szklarska Poreba Workshop on Experimental Pragmasemantics, Szklarska Poreba.

Solt, Stephanie & Gotzner, Nicole. 2012. Experimenting with degree. In Chereches, Anca (ed.), *Proceedings of SALT 22*, 166–187. Linguistic Society of America. DOI: https://doi.org/10.3765/salt.v22i0.2636

Syrett, Kristen & Bradley, Evan & Kennedy, Christopher & Lidz, Jeffrey. 2006. Shifting standards: Children's understanding of gradable adjectives. In Deen, Kamil U. & Nomura, Jun & Schulz, Barbara & Schwartz, Bonnie D. (eds.), *Proceedings of GALANA*, Vol 2, 353–364. Cambridge, MA: UConn Occasional Papers in Linguistic 4.

Syrett, Kristen & Kennedy, Christopher & Lidz, Jeffrey. 2010. Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics* 27. 1–35. DOI: https://doi.org/10.1093/jos/ffp011

Tessler, Michael H. & Goodman, Noah. 2022. Warm (for Winter): Inferring Comparison Classes in Communication. *Cognitive Science* 46. 1–27. DOI: https://doi.org/10.1111/cogs.13095

Tessler, Michael H. & Tsvilodub, Polina & Snedeker, Jesse & Roger P. Levy. 2020. Informational goals, sentence structure, and comparison class inference. In Stephanie Denison., Michael Mack, Yang Xu, Blair C. Armstrong (eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 537–543. Austin, TX. Cognitive Science Society. DOI: https://doi.org/10.31234/osf.io/n8eyj

Toledo, Assaf & Sassoon, Galit W. 2011. Absolute vs. relative adjectives – variance within vs.between individuals. In Ashton, Neil & Chereches, Anca & Lutz, David (eds.), *Proceedings of SALT 21*, 135–154. Linguistic Society of America. DOI: https://doi.org/10.3765/salt.v21i0.2587

Tribushinina, Elena. 2013. Adjective semantics, world knowledge and visual Context: Comprehension of size terms by 2- to 7-year-old Dutch-speaking children. *Journal of Psycholinguistic Research* 42. 205–225. DOI: https://doi.org/10.1007/s10936-012-9217-3

Ünal, Ercenur & Papafragou, Anna. 2016. Interactions between language and mental representations. *Language Learning* 66. 554–580. DOI: https://doi.org/10.1111/lang.12188

von Stechow, Arnim. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3. 1–77. DOI: https://doi.org/10.1093/jos/3.1-2.1

Waxman, Sandra R. & Markow, Dana B. 1995. Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology* 29. 257–302. DOI: https://doi.org/10.1006/cogp.1995.1016

Weicker, Merle. 2019. *The role of semantic complexity for the acquisition of adjectives.* Frankfurt am Main, Germany: Goethe University Frankfurt dissertation.