Are generics defaults? A study on the interpretation of generics and universals in 3 age-groups of Spanish-speaking individuals

Elena Castroviejo, José V. Hernández-Conde, Dimitra Lazaridou-Chatzigoga, Marta Ponciano & Agustín Vicente

Abstract

This paper reports an experiment that investigates interpretive distinctions between two different expressions of generalization in Spanish. In particular, our aim was to find out when the distinction between generic statements (GS) such as Tigers have stripes *and universally quantified statements (UQS)* such as All tigers have stripes *was acquired in Spanish-speaking children of two different age groups (4/5-year-olds and 8/9-year-olds), and then compare these results with those of adults. The starting point of this research was the semantic distinction between GS and UQS in that the former admits exceptions, unlike the latter. On the other hand, several authors have observed a Generic overgeneralization effect (GOG) consisting in allowing for UQS to be felicitous in the face of exceptions, thus proposing that this "error" stems from GS being defaults (simpler, more easily learned and processed). In the current paper we aimed to test the "Generics as Default" (GaD) hypothesis by comparing GS and UQS in three different age ranges. Our data show that, overall, the accuracy of GS is greater than the accuracy of UQS. Moreover, we also confirm a hypothesized interaction between age and NP type (GS vs UQS). Further, we present several data points that are not predicted by the GaD, including an observed decline in the accuracy of GS in the older group of children as well as in adults, and that children fail at rejecting statements that are not considered to be true generalizations.*

1. Introduction

The expression of generalization is pervasive in everyday language. Across languages and within the same language, different mechanisms are used to this effect. The goal of the present paper is to compare the acquisition and interpretation of two such strategies in Spanish. Specifically, generic statements reporting non-accidental regularities (henceforth GS), (1)-(2), and universally quantified statements (henceforth UQS), (3)-(4).

(1) Cats have whiskers. ENGLISH
(2) Los gatos tienen bigotes.[1] SPANISH
DET.PL cats have.3PL whiskers
'Cats have whiskers.'
(3) All cats have whiskers. ENGLISH
(4) Todos los gatos tienen bigotes. SPANISH
all DET.PL cats have.3PL whiskers
'All cats have whiskers / every cat has whiskers.'

As will be developed in the following sections, one key distinction between GS and UQS is that the former, but not the latter, are tolerant to exceptions. Following the lead of Lazaridou-Chatzigoga et al. (2013, 2019), in this paper, we want to stress the importance of exceptions as a means to identify differences in the interpretation of GS and UQS both in children and adults. Moreover, in the same line, we also make a case for the need of experimental studies in this domain that direct our attention to languages other than English and which pay close attention to linguistic differences in the expression of generalizations. In particular, here we will analyze the results of an experiment in Spanish featuring a comparison between three age groups: a group of 4/5 year-old children, a group of 8/9 year-olds and a group of adults. The main insight from Spanish is that, unlike in English, GS are typically expressed by means of definite plurals, as in (2),[2] and UQS include the definite determiner, as in (4).[3] In a nutshell, we show that, while GS may be considered to be easier than UQS at first sight, a more thorough look at the results of our experiment proves otherwise. While young children seem to be more tolerant to exceptions for generics than older children, they do not have an adult-like behavior in the interpretation of GS. As to UQS, we suggest that there is a difference in the acquisition of universal quantifiers depending on whether their quantificational domain is explicitly established, or it is implicitly treated as the entire domain.

This article is structured as follows: in the next section we lay the ground for the study of generalizations, both from a semantic and a psychological perspective, identifying their weaknesses and the predictions made for language acquisition and processing. In

[1] List of abbreviations: 1,2,3 = first, second, third person, COND = conditional, DET = determiner, DP = determiner phrase, PL = plural, SG = singular.
[2] But not only. See e.g. Ionin et al. (2011) for an experimental study on the cross-linguistic expression of GS, which takes into consideration the different DP structures that admit generic interpretations in Spanish. As for theoretical accounts in Romance languages, see e.g. Zamparelli (2002), Farkas & de Zwart (2007), Mari et al. (2013), and references therein.
[3] We leave out of the scope of this research determinerless universally-quantified sentences of the shape *Todo hombre es mortal* '(lit.) All man is mortal', which typically appear in categorical statements, and especially in the premises of syllogisms. For the purposes of comparison with English, it is important to note that Spanish UQS need to occur with the definite determiner when the noun is plural, (i).
(i) *Todos *(los) hombres son mortales.*
all the.PL men are mortal.PL

section 3 we report our experiment, and discuss results of the three groups. Section 4 concludes.

2. Theoretical and experimental background

As said before, generalization, i.e., the expression of regularities, is conveyed through different linguistic structures, both within the same language and across different languages. Focusing on GS and UQS, and their differences, one of the main properties of GS is that they tolerate exceptions. Nevertheless, as pointed out by Pelletier (2010), it is by no means clear how many exceptions a can GS admit and still be true. In the literature on generics, several types of generics have been proposed. For instance, Leslie et al. (2011) consider the following: quasi-definitional, majority characteristic, minority characteristic, majority and striking, (5) (see Lazaridou-Chatzigoga et al. 2015 for an overview).

(5) a. Quasi-definitional: *Triangles have three sides.*
b. Majority characteristic: *Tigers have stripes.*
c. Minority characteristic: *Lions have manes*.
d. Majority: *Cars have radios.*
e. Striking: *Sharks attack people.*

Quasi-definitional generics (5)a are statements speakers take to be analytical truths. Majority characteristic generics (5)b are statements that are true of a majority of individuals of a kind in virtue of some law-like connection between the essence of the kind and the property referred to by the predicate. Majority characteristic generics are usually compared with accidental regularities, such as *Canadians are right-handed*, which, according to Leslie et al. do not qualify as true generics. Minority characteristic generics (5)c are statements that are true of a qualified minority of individuals of a kind also in virtue of some law-like connection between the essence of the kind and the predicated property. Majority generics (5)d are statements that hold for a majority of individuals of a kind. Finally, striking generics (5)e are statements that are typically true of only a small subset of individuals of a kind, but the predicated property, which again is not an accidental property of such individuals, is striking or particularly interesting for us humans.

In a nutshell, from the viewpoint of their semantics (cf. Krifka et al. 1995, Dahl 1995, Zamparelli 2002, a.m.o.), GS, compared to UQS, are characterized by two main properties:

(6) a. They tolerate exceptions.
b. They are not associated with an overt dedicated quantifier.
That is, sentences such as (1) and (2) are true even in the face of a cat that does not have whiskers — which would make (3) and (4) false; moreover, cross-linguistically, there does not seem to be a quantifier which expresses a generalization that admits exceptions. In fact, one way of expressing GS in English is a bare plural, (1), but in other languages, such as Spanish or Greek, a plural definite determiner is recycled for this purpose.

Generalizations have recently caught the attention of cognitive and experimental psychologists, especially in view of the so-called 'Generic Overgeneralization' (GOG) effect, an error consisting in interpreting or recalling UQS as GS. For instance, participants in an experiment would hear *All ducks lay eggs* and would recall the sentence as *Ducks lay eggs* (Leslie & Gelman, 2012). In fact, several studies on adults have observed a tendency to treat UQS as true even in the face of exceptions. In view

of these results, Leslie (2007, 2008) and Gelman (2010) have endorsed the Generics as Default (GaD) hypothesis, according to which GS are simpler and hence more easily acquired and processed than UQS, which would explain an asymmetry that is not otherwise observed or predicted by semantic theories.

More than this, this generic bias is founded on a dual view of cognition as proposed by Kahneman & Frederick (2002) among others, which posits the existence of two cognitive systems which have different properties. System 1 is fast, automatic and effortless, while System 2 is slow, effortful, higher-level, and rule-governed. Leslie (2007) builds on this dual mechanism to claim that GS are part of System 1, while UQS are part of System 2. According to her, issuing and verifying generic generalizations (i.e., GS) is a matter of checking whether a certain category exhibits a certain feature by accessing a conceptual representation of such a category, whereas issuing and verifying universal generalizations (i.e., UQS) involves working memory and checking statements against possible exceptions. Note also that System 1 cognitive processes are taken to be evolutionary ancient and to arise early in development, and that, accordingly, GS are mastered earlier than UQS (Leslie & Gelman, 2012). In Leslie's view, then, the overgeneralization of UQS as GS is an example of the "lazy" overuse of System 1. Relatedly, she interprets the lack of a dedicated overt operator to express GS as following from the fact that System 1 is simpler. Hence, lack of "markedness" is associated with less complexity.

As summarized in Lazaridou-Chatzigoga et al. (2015), among the 20 articles on the acquisition of generics published so far, only two studies are identified as clear arguments in favor of the GaD view. First, Hollander et al. (2002) asked 3- and 4-year-olds questions such as *Are {fires, all fires, some fires} hot?* They found that 3-year-olds were adult-like only in GS, while 4-year-olds were adult-like in GS and UQS; these two things together seemed compatible with the GaD hypothesis. The second study is the recall study mentioned above (Leslie and Gelman, 2012), where children and adults were asked to recall novel facts about familiar animal kinds, which were introduced either as GS or as UQS. Both adults and children recalled many UQS as generic. Again, this suggested that the GaD could be on the right track.

As far as the acquisition of universal quantification is concerned, the major cross-linguistic work has been Katsos et al. (2016). In a report of 31 languages representing 11 language types, the authors raise the question concerning the order of acquisition of different quantifiers in different languages, and the linguistic constraints that may have an effect in potential differences across languages. The materials used consist of sentences of the form 'all of the N are in the boxes', where N is a placeholder for balls, sandwiches, dinosaurs, pens and shoes. In this design, children (mean age 5.5 years old) are instructed to help a cavegirl who wants to learn the language and, so, they have to say whether sentences are true or false when watching different images representing quantificational relations. In this work, the proposed criteria for the acquisition of quantifiers are monotonicity (quantifiers that validate inferences from subsets to supersets are acquired before), totality (quantifiers that attribute a property to all or none of the members of the set are easier to acquire), complexity (more calculation is necessary to process *most* than *some*) and informativeness (children will be stricter in violations of truth than pragmatic felicity). Important for our purposes is that monotonicity and totality support that *all* implies higher performance than other quantifiers such as *some* and *none*. After data collection and comparison, these seem valid generalizations across languages.

Interestingly, Katsos et al. (2016) ran the cavegirl task in Spanish, including sentences such as (7). With typically developing children whose mean age was 6.4, the proportion of correct responses in the *all* condition was above 95%.

(7) Todas las pelotas están dentro de las cajas.
all DET.PL balls are.3pl inside of DET.PL boxes
'All the balls are inside of the box.'

Barberán-Recalde (2019) obtains similar results with L1 speakers of Spanish in her comparison to Spanish-Basque bilinguals using the same task, with results at ceiling percentages of success already at age 4. Note, however, that the cavegirl task measures the command of the universal quantifier in partitive, non-generic statements, what we have called the *restricted* interpretation of UQS (as in the studies of Barner et al. 2009). Here, the total amount of relevant members (balls, shoes, sandwiches, etc.) and the boxes are displayed on the screen, so the children only pay attention to the set relation, but do not need to resort to world knowledge to verify the relation denoted by the quantifier. Hence, we can assume that 4-year-old Spanish speakers have a good command of *todos* 'all' in this specific condition.

In the case of generic statements, the study of languages other than English from an experimental perspective is unquestionably motivated. For one, in languages other than English the expression of genericity is different. Spanish (as well as Greek) recycles a determiner that is used in other contexts and thus we could even predict that GS will be harder to learn (or to have a good command of), since the child has to figure out the ambiguity of the determiner. As described by Pease Gorrissen (1980), sentence (8) is ambiguous between a generic and a specific reading in Spanish. In the non-generic reading, both the subject and the object refer to a specific set of shepherds and sheep, while in the generic one, the sentence concerns shepherds and sheep in general.

(8) Los pastores llevan (a) los borregos a pastar.
DET.PL shepherds take (to) DET.PL sheep to graze
'(The) Shepherds take (the) sheep to graze.'

Hence, Greek and Spanish could pattern together and differently from English in view of the different morphosyntax of GS.[4]

The literature on GS in Spanish is scarce. We would like to report a second experiment, due to Gelman et al. (2016), which tackles the comparison we are interested in, albeit in an indirect way. In this study, 48 Spanish-speaking children (mean age 5) and 48 Spanish-speaking adults were instructed to recall sentences in two conditions: DET.PL (GS) and quantified NPs, which was in turn, either *muchos* ('many') or *todos* DET.PL ('all the', UQS). Gelman et al. motivated the need for a study in another language to rule out the possibility that quantified statements were recalled as GS because, in English, the latter are simpler ("[…] the generic has one fewer word […]", p.1233). Since, as said, Spanish GS are not bare nouns, this language makes a good testing ground for investigating whether previous results are replicated (and hence, more evidence in favor of the GaD hypothesis is obtained), or else, the lack of differences in morpho-syntactic complexity is key in determining whether or not quantified sentences are recalled as GS (in line with Leslie 2007). In fact, the comparison between *todos* DET.PL and *muchos* is

[4] In fact, it cannot also be taken for granted that results about *all* should extend to other universal quantifiers in the same language or in other languages. This is tackled in Lazaridou-Chatzigoga et al. (2013, 2019).

included so it can be tested whether there is a difference between UQS and GS (where the UQS condition has one word more), but not between GS and *muchos* sentences (where both NPs have the same number of words). The three conditions are illustrated in (9), where the GS is realized as a plural definite determiner (DET.PL), the *muchos* sentence does not include a DET.PL, and the UQS sentence does.

(9) a. Los osos trepan árboles.
DET.PL bears climb.3PL trees
'Bears climb trees.'
b. Muchos osos trepan árboles.
many bears climb.3PL trees
'Many bears climb trees.'
c. Todos los osos trepan árboles.
all DET.PL bears climb.3PL trees
'All bears climb trees.'

Alongside the Spanish-speaking participants, 48 English-speaking adult controls carried out a version of the GS-*muchos* condition (bare N-*many* in English).

Building on Leslie & Gelman (2012), in the design, participants were shown a series of photographs of animals and learned a novel fact about them (half realized as GS and half realized as a quantified statement), then they had a 4-minute distractor task, and finally they were shown the pictures again and were asked to recall what the experimenter had told them about each picture.

The results can be summarized as follows:

The English-speaking group exhibits a generic bias, whereby correct recall was higher for the generic condition. There is a higher tendency to recall *many* as a GS than vice versa.

In the case of Spanish-speaking children and adults:

- There is a main effect of age group indicating that adults recalled sentences more accurately than children; and a main effect of sentence type, such that generics are recalled more accurately than quantified sentences (both in the UQS and the *muchos* condition).
- In the examination of errors, participants tend to interpret the quantified statements as GS more often than the other way around across quantifier types. And while there is no age effect, there is an age x sentence type interaction, such that there is a larger difference between GS and quantified sentences in children than adults.

Gelman et al. interpret these results as evidence in favor of the GaD hypothesis. That is, irrespective of the length of the NP in a particular language, GS are recalled correctly more than quantified statements, and even quantified statements are often recalled as GS.

Despite these results, we believe that it is essential to run further experiments in languages other than English, because we would like to test in a more direct way — i.e., not through a recall experiment — whether differences can be observed in the interpretation of GS vs. UQS. Recall designs can give information about cognitive processes related to memory, but they only *indirectly* inform us about the acquisition of the adult interpretation of linguistic expressions. Certainly, it may be that GS are easier

to retain in the hearers' memory than quantified statements. However, this may not necessarily have to do with generic generalizations being defaults (as the GaD contends). It could also be a difference in the syntactic-semantic properties of the two structures. To make sure we circumvent this problem, we propose a new study, inspired by Lazaridou-Chatzigoga et al. (2013, 2019), that compares the interpretation of GS and UQS in Spanish-speaking children (and the corresponding adult controls).

3. Our study

3.1. Research questions and hypotheses

According to the Generics as Default (GaD) view, GS are defaults (see Gelman et al. 2016 for Spanish), and so they should be easier to process and easier to learn. Hence, we would expect the comprehension of GS to be adult-like from the beginning (the young group), while we may expect the comprehension of UQS to depart from the adult behavior in the child groups. Moreover, unlike what could follow from Leslie (2007), given the results in Gelman et a. (2016), differences in morpho-syntactic markedness should not mirror processing complexity. Hence, while GS in Spanish are more complex than GS in English, cross-linguistic differences should not arise. Let us remember that GS in Spanish are still less complex than UQS in terms of number of words, but they can be said to be less complex structurally and, in any case, they are superficially ambiguous between an exemplar and a generic reading (see Pease Gorrissen 1980, 1981, who calls the exemplar reading "specific"), which could be argued in favor of adding processing complexity.

Concerning the background on the acquisition of UQS (Katsos et al. 2011, 2016, Barberán-Recalde 2019), we should expect 4-year-olds to master universal quantification. In fact, *all* is easier to acquire than other quantifiers, including *some* and *most*. However, we can only be sure about UQS whose domain restriction is made explicit by the utterance context, i.e., when the total number of objects being quantified over are visibly present, and *all* is followed by a partitive phrase as in *all of the.*

Our research questions are the following, paired up with our research hypotheses:

RQ1) Are children sensitive to the reported differences between GS and UQS? We hypothesize that children are sensitive to the differences between GS and UQS: The interpretation of GS is more akin to the adult interpretation, whereas the interpretation of UQS is less adult-like.

RQ2) Is there an interaction between NP type and age? Is the joint effect of NP type and age on accuracy predictable from the effect of the two factors individually? We hypothesize that there is one.

3.2. Method

3.2.1. Participants

A total of 55 Spanish-speaking children (30 male; 25 female) divided into two age groups, a 4/5-year-old group (N = 31, M = 68.16 months, SD =6.8, henceforth "young") and an 8/9-year-old group (N = 24, M = 108.75 months, SD = 6.3, henceforth "old"), were recruited from a local (primary) school in Vitoria-Gasteiz (Spain). The participants in the young group belonged to different school years: 2nd and 3rd years of "Infantil" (child-care). We decided to collapse their results given that we did not have enough participants for each group (N=15 for each). On the other hand, in a preliminary study we did not see

significant differences in their performance.[5] The participants in the old group all belonged to the 3rd year of primary school.

The criteria for including these two groups were the following: given the information obtained from Hanlon (1988) and Barberán-Recalde (2019), among others, we could assume that the young group would be competent in the interpretation of *todos*, at least in the restricted sense, where the set of members quantified over are made available. The rationale for recruiting a group of 8/9 years is that we plan to run the study with a group of autistic people whose minimal IQ is equivalent to that of a 9-year-old child. The aim of that study is to test whether in the autistic population there may be a marked intolerance to exceptions (which is the case in the social domain: Strang et al., 2017), and so a higher rejection rate of GS. We want to test whether such intolerance to exceptions only concerns social rules or it is rather a cognitive trait. By recruiting this group of children as controls for a forthcoming study we also collected data that prove interesting for other reasons.

Alongside, we included a third group of adult controls (N = 26, 12 male; 14 female, henceforth "adults"). Following the same strategy as in Gelman et al. 2016), we did not test college undergraduates, but rather adults of various ages (M = 38, SD = 13.92) and in various schooling degrees (ranging from elementary education to a BA university degree). 25 of them had Spanish as their L1, and 1 of them had Basque as her L1, although she claimed to be a fully Basque-Spanish bilingual and to have perfect command of Spanish. 13 Spanish speakers also declared some knowledge of other languages such as Basque, English, Galician or Hassānīya.

All participants were Spanish speakers and residents in Vitoria-Gasteiz. All volunteered to take part in the experiment. This study was carried out in accordance with the recommendations of the Human Beings Research Ethics Committee ("CEISH: Comité de Ética de Investigación con Seres Humanos de la UPV/EHU") with written informed consent from all subjects. Parents or caretakers gave written informed consent for their children to participate in the investigation prior to the inclusion in the study. Children also assented to participating in the experiment.

3.2.2. Design

Building on Lazaridou-Chatzigoga (2013, 2019), our experimental design places the role of context and, more specifically, of exceptions, at the center of the discussion. To keep the experiment simple, we also focused only on majority characteristic statements. We manipulated the NP create a generic condition (DET.PL) and a universal condition (*todos/as* DET.PL 'all'). By contrast, we did not manipulate the context variable between a supportive and a contradictory condition; all our critical items were uttered on the basis of a contradictory context, i.e., in the face of exceptions. Below is an example.

(10) [Faced with the picture of a rabbit with one ear and hearing the utterance "a rabbit with one ear"]
¿Dirías que todos los / los conejos tienen dos orejas?
say.COND.2SG that all DET.PL DET.PL rabbits have.3pl two ears
'Would you say that all / rabbits have two ears?'

[5] We could not collect reliable information on the linguistic background of the children. We should assume some competence in Basque in some of them, as well as proficiency in languages other than Spanish in children whose families are originally from non-Spanish-speaking countries.

However, to prevent participants from developing a response strategy, we included supportive contexts in the fillers.

Each participant in the study saw 16 critical items and 32 distractors. In the critical items, NP type (GS vs. UQS) was a within-individual independent variable. We also had a between-individual variable, namely the three age conditions: young, old and adult. The dependent variable was a yes/no response to the question prompted as part of the experiment's procedure (see example in (11)).

Given the manifest existence of exceptions and following the semantic-logical properties of the different expressions, GS and UQS were true in different scenarios. For instance, given the context described in (10), the truth values are the ones in (11).

(11) a. All rabbits have two ears. → False
b. Rabbits have two ears. → True

Coming back to the distractor items, we created two conditions. First, we had similar sentences with proper name subjects (henceforth "Name"), which was a mix of true and false statements, to control for the participants' *yes*-bias and for their understanding of the procedure, (12). Second, we had generics which should be treated as false[6] and which were presented with a supportive context (henceforth "Controlgen"), (13).

(12) [Faced with a picture of Plaza de la Virgen Blanca and hearing a voice uttering "La Plaza de la Virgen Blanca".] [Proper name]
¿Dirías que la Plaza de la Virgen Blanca está en Vitoria?
say.COND.2SG that det.sg Plaza de la Virgen Blanca is in Vitoria
'Would you say Plaza de la Virgen Blanca is in Vitoria?

(13) [Faced with a picture of a square-shaped pizza hearing a voice uttering "Una pizza cuadrada"]
¿Dirías que las pizzas son cuadradas? [Controlgen]
say.COND.2SG that det.pl pizzas are square-shaped.pl
'Would you say pizzas are square-shaped?'

3.2.3. Materials

As previously mentioned, participants saw 16 critical items (8 GS and 8 UQS) and 32 distractors (16 Names and 16 Controlgen). They were distributed in two lists that were randomly assigned to the participants. So, each critical item came in two different conditions (GS or UQS), and participants in the different lists did not see the same item in the same condition. All participants saw the same set of fillers. Apart from the set of 16 critical items and 32 distractors, there were 4 training items, also common to all participants (at the very beginning of the experiment to ensure that participants followed the dynamics of the experiment). The materials were counterbalanced across participants. Also, experimental items were randomized every time a participant started a new experimental session.

The 16 critical items consisted of majority characteristic statements like *Cats have whiskers* and *Horses have four legs*. Special care was taken to select properties about which young children would be knowledgeable. Also, to avoid the possible ambiguity in

[6] Remember from §2, example (5), true generics can be of 5 classes (quasi-definitional, majority characteristic, minority characteristic, majority and striking: Leslie et al., 2011). None of them would encompass cases such as "Pizzas are square-shaped". We therefore assume that according to current proposals, such generic-like statements should be rejected and treated as false. See more on this issue in §3.4.3.

the GS conditions (remember from §2 that DET.PL can be ambiguous between a generic and an exemplar reading), the picture that was shown always depicted a single individual exhibiting an exceptional property (the same strategy of showing only one individual in the image to avoid the exemplar reading of the definite plural was followed by Gelman et al. 2016).

Also, bear in mind that, for each critical generalization, a picture describing an exception was presented. To create these materials, we used two strategies: we either selected a subkind of the species (a sphinx cat, which does not have whiskers, or a black pig, which is gray) or an animal that, for an accidental reason, may have suffered a mutation (a three-legged horse or a hen with four wings).

As indicated in §3.2.2, fillers came in 2 conditions (Name and Controlgen). Concerning the former, we mixed true and false sentences. Accompanying pictures here were more or less irrelevant, usually representing the referent of the subject. For instance, a picture of Donald Duck for the sentence *Would you say that Donald Duck has a beak?*. Regarding Controlgen, unlike critical GS, they described minority characteristics, so the sentences could not be considered to belong to any of the known types of generics (see (5) above), not majority, not striking, etc. For instance, square-shaped pizzas, black roofs or blue butterflies. In these cases, accompanying pictures were supportive, so the participant would see a picture of a square-shaped pizza while she was asked whether pizzas are square-shaped.

As a final remark, the pictures that were used for the experiment were all selected from google image searches. We chose those that seemed more realistic and similar to one another in terms of the proportions of the target animal/object. Of course, images of mutant animals may be the result of Photoshop manipulation, but we chose either actual pictures or pictures whose manipulations were less obvious/had a realistic appearance.

3.2.4. Procedure

Prior to the experiment, we conducted a pilot study with 4 participants to test the materials we had created. No further modification had to be made before the items were finalized for the final experiment. We first collected data from the young group and then from the old group (the procedure was kept the same). This was done in that particular order due to participants' availability during school classes. Children were tested individually in a quiet room in their school. So, we conducted the experiment with individuals separately and in succession. Children had been previously told that they would play a game on the computer. In the case of adults, they were administered the study either at a quiet place of their convenience (for instance their home) or at Micaela Portilla Research Center (UPV/EHU) in Vitoria-Gasteiz.

Participants sat in front of a computer screen with the investigator beside them. In the case of children, the experimenter was in charge of clicking the "yes" or "no" key on the computer's mouse; while adults handled the mouse themselves.

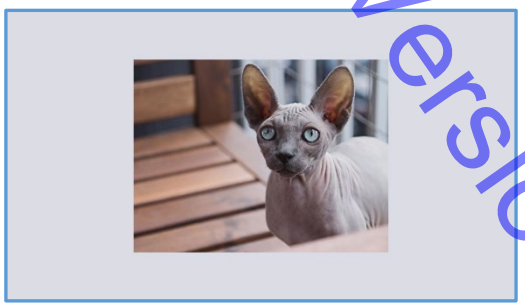
The software used was E-prime 3 ("Psychology Software Tools") on a PC running windows. At the beginning of the test, and for the sake of convenience, the investigator read the instructions that appeared on the screen. They were asked to listen carefully. The instructions highlighted the fact that the participants would see a series of images, some of which would be surprising, thus trying to bring the participant's attention to the content of the image. During training, feedback was given to make sure that participants understood how they should reply to the question asked (by answering either "yes" or "no"). No feedback was given during the main task. Supportive feedback was given at

the end, when the task was completed. The testing process took approximately 15 minutes to complete.

In the study, each target item was composed of two parts. In the first part, an image occupying the center of the screen presented one individual contradicting or supporting the generalization to be judged. For instance, a cat without whiskers. At the same time, a pre-recorded audio of a female voice said: “un gato sin bigotes” (‘a cat without whiskers’). By pressing a key, we moved to the second part, where an image of a girl, a cartoon character, appeared on the right-hand-side of the screen and asked “¿Dirías que *los* gatos tienen bigotes?” (‘Would you say that cats have whiskers?’) or alternatively “¿Dirías que *todos los* gatos tienen bigotes?” (‘Would you say that *all* cats have whiskers?’). The pre-recorded audio with the question was played twice to enhance comprehension of the question and to possibly draw the participant’s attention to the building blocks of the sentence. There was no going back to the first screen. In the meantime, participants could see the previous picture in the left-corner of the screen as a reminder of the context at the moment of listening to the target question. Participants were asked to judge whether they agreed (or not) with the statement they had just heard. It was a two-alternative forced choice task, whereby participants were instructed to choose “yes” or “no” in light of the picture that had been presented to them. Responses were recorded by button press.[7]

Below is an example of a critical trial in the study.

(14) SCREEN 1: “un gato sin bigotes” (‘a cat without whiskers’)



(15) SCREEN 2: “¿Dirías que {los gatos, todos los gatos} tienen bigotes?” (x 2) (‘Would you say {cats, all cats} have whiskers?’)

3.2.5. Data analysis

[7] Alongside the participant’s yes/no response, E-Prime also registered reaction times. Nevertheless, children did not press their own key response, and hence this data point was not reliable, so we decided to disregard reaction times altogether.

To analyze data, we used R (R Core Team, 2020), together with the lme4 package (Bates et al., 2015) to carry out generalized mixed-effects linear analyses of the effects between determiner and context on the yes/no response, specifying a binomial family. We carried out mixed-effects logistic regressions on response type with age group and NP type as categorical predictors, and random intercepts by participant and item. Responses were treated as a dummy coded categorical variable and were modeled with glmer. We fitted a linear mixed-effects regression model with age group and NP type as categorical fixed effect predictors (with an interaction term), and random effects by participant and item. In addition, we used ANOVA to perform likelihood ratio tests to compare the statistical significance both of the interaction term and of the considered random slopes (i.e., random by-participant slopes for NP type, and random by-item slopes for age group), and also to test the individual effects of age group and of NP type. Finally, two models will be shown and discussed for the critical items —and other two models for the filler items —, which will capture simple effects for age in each condition (i.e., GS condition vs UQS condition, for the case of critical items; and GS condition vs Controlgen condition, for the case of filler items).

3.3. Results

3.3.1. Critical items[8]

Table 1 summarizes the mean percentages of yes-responses and their standard deviation by NP type (GS vs. UQS) and age group (young children, old children, adult).

Table 1 Mean percentage of yes-responses by age group and critical condition (GS vs. UQS)

TABLE 1 GOES HERE

As graphically represented in Figure 1, the acceptance of the GS condition was higher than the acceptance of the UQS condition for all age groups. Young children accepted GS 92% of the time, and UQS only 70% of the time; older children accepted GS only 72% of the time, and UQS 36% of the time; and adults accepted GS 79% of the time and UQS only 28% of the time. Figure 1 also shows that the older the age group, the lower the percentage of acceptance of UQS.

In addition, and even though the percentage of acceptance of GS is not monotonically decreasing, the difference between the percentages of acceptance of GS and UQS (22% in young children, 36% in older children, and 50% in adults) is an almost perfect increasing linear function with respect to age (as shown by the red curve in Figure 1). These numbers confirm an evident developmental effect, which could be the result of a combination of: (i) a greater intolerance for exceptions in UQS with age, and/or (ii) a greater tolerance for exceptions in GS with age.[9]

FIGURE 1 GOES HERE

Figure 1 Mean percentage of yes-responses by age group and critical condition (GS vs. UQS)

In the box-plots below (Figure 2) we can see the distribution of the mean percentage of yes-responses by each individual participant, across the different age groups and the

[8] To verify that the participants understood the task, the reader is invited to check section 3.3.2 on filler items.

[9] We thank an anonymous reviewer for drawing our attention to the developmental effect in the differences between the “Yes” responses to GS and UQS across ages.

GS-UQS condition. It shows that the mean percentages of yes-responses are well distributed at the level of participant, as proven by the fact that medians are close to the means given in Table 3, and that interquartile ranges are well distributed around them. In addition, it shows a slightly greater variability in the case of the UQS condition.

FIGURE 2 GOES HERE

Figure 2 Box-plots of the percentage of yes-responses aggregated over participants and showing the means by age group and critical condition (GS vs. UQS)[10]

Then we conducted mixed-effects logistic regressions on responses with age group and NP type as fixed effect predictors, and random effects by participant and item. For all the analyses described below, we used a dummy coding (also known as treatment coding) scheme for predictors, so that the reference level is coded as 0, and the other condition is coded as 1. Firstly, we fitted a model including all the relevant fixed effects (i.e., age group and NP type), an interaction term between them, and random intercepts by participant and item. Then we carried out a likelihood ratio test of the model with an interaction term against a model without the interaction term and the comparison proved that the model with the interaction term was significantly better (χ^2=9.97(2), p=0.0068). Given that the interaction term significantly improved the goodness of fit of the model, the model with that term was used for all the subsequent analyses and comparisons.

We then fitted versions of this base model, in which a single effect was removed and we compared the reduced model to the base model. To test the main effect of age, we removed age group. A likelihood ratio test of the base model against the model without age proved significant (χ^2=39.92(4), p<0.001). Thus, we concluded that there was a main effect of age. To test the main effect of NP type, we removed NP type. A likelihood ratio test of the base model against the model without NP type proved significant (χ^2=272.89(3), p<0.001). Thus, we concluded that there was a main effect of NP type.

Then, we included random by-participant slopes for NP type, and random by-item slopes for age group in the model. We started including the by-item random slopes for age group and tested their effect. The new model was significantly better than the model that only had random intercepts (χ^2=11.4(5), p=0.044). Next, we compared this new model with a full model with both random by-participant and by-item slopes. The comparison proved significant (χ^2=6.69(2), p=0.035). Thus, adding random slopes significantly improved model fit, so we used the model with maximal random-effects structure justified by design for all subsequent analyses and discussion.

Table 2 shows the full model parameters of two mixed effects logistic regressions carried out for the set of critical items. Both models are regressions on responses with age group and NP type (GS or UQS) as fixed effect predictors, and random effects by participant and item. We used adult and GS as the reference levels of the first model, and adult and UQS as the reference levels of the second model. The aim of including these two models is to show how differently response depends on age for the GS condition versus the UQS condition.

[10] A small amount of random variation (i.e., 10%) has been included in the representation of each individual data point in Figures 2 and 4, in order to handle the overplotting due to discreteness of small datasets.

As we show in Table 2, in the case of adults the difference between GS and UQS was statistically significant ($p<0.001$). This was as anticipated, as generics were expected to be accepted even in the face of exceptions, whereas universals were expected to be rejected. (The mention of an exception in the case of UQS just served as a reminder.) Adults accepted universals in similar percentages as in previous experiments (~30%). In the case of UQS, adults differed statistically from young children ($p<0.001$), but were not significantly different from the old children group ($p=0.1766$); while in the case of GS there were not significant differences between adults and the groups of young and old children.

Table 2 Mixed-effects logistic regressions (full model with interaction term) [Critical Items]

TABLE 2 GOES HERE

Then we conducted comparisons for the different age groups and GS-UQS conditions. We fitted the two models shown in Table 2 changing the reference level of age group from adult to young and old. Thus, we obtained four additional models with the following reference levels: old-GS, old-UQS, young-GS, and young-UQS; in addition to the two models already shown in Table 2, with reference levels: adult-GS, and adult-UQS. In all these models, the difference between GS and UQS was statistically significant ($p<0.001$), independently of the reference level chosen for age group. This had been anticipated for the case of old children, because they were expected to have adult-like behavior; and it is notable that the difference between GS and UQS is also significant for young children.

In the GS condition, old children did not significantly differ from adults ($p=0.2683$), but they were significantly different from young children ($p=0.0048$); lastly there was not a significant difference between young children and adults ($p=0.1204$). In the UQS condition, adults significantly differed from young children ($p<0.001$), but were not significantly different from old children ($p=0.1766$); while old children differed significantly from young children ($p<0.001$). All in all, both in the GS condition and in the UQS condition the behavior of young children were significantly different from that of old children, and the differences between old children and adults were not statistically significant. Finally, the difference between young children and adults is statistically significant in the UQS condition, but not in the GS condition. In the light of these results, it seems that the adult-like way of answering UQS is acquired sometime between 4/5 years-old and 8/9 years-old.

3.3.2. Filler items

Moving to the filler items, Table 3 presents the mean percentages of yes-responses, and their standard deviations, of the three age groups depending on the filler type. This is graphically represented in the bar graph shown in Figure 3. The results of the proper name condition (namely, mean percentage of yes-responses close to 0.5, and standard deviations almost equal to 0.5) confirm that all the participants understood the task, were paying attention and had some amount of world knowledge — which increased with age. (As shown below, these results were confirmed at the level of participant by the box-plots presented in Figure 4.) More revealing is the Controlgen condition, which shows relatively high values (51% and 56%) in the child groups (young and old, respectively), with respect to the low value for the case of adults (16%).

Table 3 Mean percentage of yes-responses by age group and filler condition (Controlgen vs. Proper name)

TABLE 3 GOES HERE

FIGURE 3 GOES HERE

Figure 3 Mean percentage of yes-responses by age group and filler condition (Controlgen vs. Proper name)

Figure 4 represents the distribution of the mean percentage of yes-responses by each individual participant, across the different age groups and the Controlgen-ProperName condition. Again, it confirms that the yes-responses are well distributed at the level of participant, since medians are close to the means given in Table 5. In addition, it corroborates — at the participant level — that participants understood the task and were paying attention, as proven by the very narrow interquantile ranges in the proper name condition.

FIGURE 4 GOES HERE

Figure 4 Box-plots of the percentage of yes-responses aggregated over participants and showing the means by age group and filler condition (Controlgen vs. Proper name)

Next, in order to compare the GS condition with the Controlgen condition, mixed-effects logistic regressions were carried out on responses with age group and item type as fixed effect predictors, and random effects by participant and item. First, we fitted a model with all the relevant fixed items (age group and NP type) an interaction term between them, and random intercepts by participant and item. We performed a likelihood ratio test of that model against a model without the interaction term and — as also happened in the case of critical items — the comparison proved that the model with the interaction term was significantly better (χ^2=78.8(2), p<0.001), so the model with interaction was used for all the later analyses and comparisons.

We then fitted multiple variations of this base model, where a single effect was removed and compared the reduced model with the base model. In order to test the effect of age, age group was removed, and a likelihood ratio test of the base model against the model without age proved that age group was statistically significant (χ^2=113(4), p<0.001), so there was a main effect of age. In order to test the effect of item type, we removed it and a likelihood ratio test of the base model against the model without item type proved that the GS-Controlgen condition was statistically significant (χ^2=125(3), p<0.001). Hence, we concluded that there was a main effect of item type as predictor.

Next, we included in the model random slopes by-participant for NP type, and random slopes by-item for age group. First, we included the by-item random slopes for age group and tested their effect. The new model significantly improved the model only with random intercepts (χ^2=20.8(5), p<0.001). Then, we compared the new model with a full model with both random by-participant and by-item slopes, and the comparison was significant (χ^2=110(2), p<0.001). Since adding random slopes significantly improved model fit, we used the model with maximal random-effects structure for all subsequent analyses and discussion.

Table 4 Mixed-effects logistic regressions (full model with interaction term) [Generic Statements]

TABLE 4 GOES HERE

Table 4 shows the full model parameters of two mixed effects logistic regressions carried out for the set of Generic Statements. Both models are regressions on responses with age group and Item type (critical or filler) as fixed effect predictors, and random effects by participant and item. We used adult and critical as the reference levels of the first model, and adult and filler as the reference levels of the second model. The aim of including these two models is to show how differently response depends on age for the GS condition (critical item) versus the Controlgen condition (filler item).

As it is shown in Table 4, for the case of adults the difference between GS (critical) and Controlgen (filler) was statistically significant ($p<0.001$). This was as anticipated, as GS were expected to be accepted more often than Controlgen. In the case of GS, adults statistically differed from young children ($p=0.019$), but were not significantly different from the old children group ($p=0.346$). By contrast, in the case of Controlgen, adults were statistically different from both young children ($p<0.001$) and old children ($p<0.001$).

Next, we carried out comparisons for the different age groups and the GS-Controlgen conditions. We modified the two models shown in Table 4 changing the reference level of age group from adult to young and old, so we obtained four additional models with reference levels old-GS, old-Controlgen, young-GS, and young-Controlgen, in addition to the two models in Table 4 (with reference levels adult-GS and adult-Controlgen). In all these models, the difference between GS and Controlgen was statistically significant, independently of the reference level chosen for age group ($p<0.001$ for adult; $p=0.0343$ for old children; $p<0.001$ for young children). This was expected for the case of old children, since they could have adult-like behavior, and it is worth saying that the difference between GS and Controlgen was also significant for the case young children.

In the GS condition, old children did not significantly differ from adults ($p=0.346$), but they were significantly different from young children ($p=0.0011$); while young children differed significantly from adults ($p=0.019$). By contrast, in the Controlgen condition, adults significantly differed from young and old children ($p<0.001$); while old children did not significantly differ from young children ($p=0.64$). Therefore, in the GS condition the behavior of young children was significantly different from those of old children and adults, and the differences between old children and adults were not statistically significant; in the Controlgen condition the behavior of adults differed significantly from those of young and old children, and the differences between the behavior of young children and old children were not statistically significant. In the light of these results, we may conclude that, while the adult-like way of answering GS is acquired sometime between 4/5 years-old and 8/9 years-old, the acquisition of the adult-like way of assenting to Controlgen statements occurs after the age of 8/9.

3.4. Discussion

In this investigation we collected data from children in two different age groups regarding their acquisition and interpretation of generic statements (GS) and universally quantified statements (UQS). The view from semantics is that the former are different from the latter in that GS admit exceptions, unlike UQS. However, cognitive psychology has observed a tendency of children and adults to retrieve UQS as GS (the Generic Overgeneralization [GOG] effect). Authors like Leslie and Gelman take a step further and claim that GS are easier to process and more easily learned than UQS because they are associated with two different cognitive systems. The so-called “Generics as Default” (GaD) hypothesis makes an interesting prediction that can be experimentally tested against new data. Our goal in this study was to bring data from Spanish to the debate. Here we provide the data from three Spanish-speaking age groups of individuals.

Prior to this study, we hypothesized that children would be sensitive to the differences between GS and UQS observed in the literature. Indeed, compared to the adult behavior, we have found a difference in the "Yes" responses to GS and UQS. Specifically, the interpretation of UQS is less adult-like. Moreover, we expected to find an interaction between NP type and age, which is precisely what we have observed.

In what follows, we present an overview of the main features of the three different age groups (§3.4.1), we then discuss whether the collected data can be interpreted as supporting the Generics as Default (GaD) view (§3.4.2), and finally we conclude by entertaining alternative explanations for the facts and discussing ideas that could verify these views.

For the ease of exposition in the upcoming discussion, from now on we employ the term "accuracy", defined as in Table 5 and 6 both for critical and filler items, respectively.

Table 5 Accuracy (Critical)

TABLE 5 GOES HERE

Table 6 Accuracy (Filler)

TABLE 6 GOES HERE

Taking this stance allows us to establish a standard for what would have been the correct answer for each question depending on NP type and, thus, assess the performance of the three age groups with respect to this baseline.

3.4.1. Developmental overview

Remember that we had three age groups. The one we called "young" (4/5-year-olds), the "old" group (8/9-year-olds) and the adults. The striking thing about the young group is that their accuracy rate of universals is very low (i.e. they accept UQS in the face of exceptions). It is striking because children of that age can interpret restricted universal quantifiers (see Hanlon 1988, Barner et al. 2009, Katsos et al. 2011, 2016, Barberán-Recalde 2019 and §2 above). Given that we were giving them the exception and thus facilitating the verification process, we had expected that they would reject more universals than they did. They excelled at our critical generics, apparently confirming that generics are easy, and that they tolerate exceptions. If we delve into individual items, we observe notable variation, from *People have teeth*, which got 100% accuracy, to *Elephants have tusks*, which obtained 75%. The variation in rejection rates of UQS was even more noteworthy, not to mention the distribution of acceptance rates of generics vs rejection rates of universals item by item. As an illustration, we have pairs like: acceptance of 93% for *Sheep are white* (GS), and acceptance of 50% for *All sheep are white* (UQS), compared to: acceptance of 94% for *Horses have four legs* (GS) and acceptance of 93% for *All horses have four legs* (UQS). These latter rates are striking again if we compare them to the results obtained with the similar pair *People have two arms* (94% acceptance), and *All people have two arms* (73% acceptance).[11] Importantly, young children did not perform well at Controlgen, which we had as control items. Remember, these were minority features generics like *Pizzas are square shaped*, which were accompanied by a picture of a square-shaped pizza. However, they excelled at

[11] We did not hypothesize any relation between accepting a GS and rejecting its corresponding UQS or vice versa, so the reported difference in the accuracy rates of GS vs UQS pairs is pending further analysis. However, it is striking to see that apparently there is *no* such relation between GS and UQS (say, between rejecting a UQS and accepting its corresponding GS).

fillers asking about properties of well-known individual objects, e.g. *The Eiffel Tower is tall*.

Concerning the old group, the surprising result was that they performed worse on generics than the young group, and that their performance on GS *vis à vis* UQS was very similar (to the point of not being statistically significant). The pattern of responses to GS in this group (as well as in the adults) is similar to the pattern found by Lazaridou-Chatzigoga et al. (2019). In that paper, they suggested that they could not discard that participants could have been misled by an ambiguity in the Greek statements (remember, the DET.PL ambiguity also present in Spanish). As the stimuli they had concerned several exemplars (e.g. three white tigers), the question "Do tigers have stripes?" could be interpreted as referring to those three white tigers. In our experiment, we controlled for this issue by presenting only one exception, and we got similar rejection rates. However, the most interesting data is not that the old children (and adults, as we will see) rejected a 28% of the GS, but that both they and the adults performed worse than the young group. As in the young group, the distribution of rejecting rates for UQS was broad, and broader than the distribution of acceptance of GS. Also, there is a striking distribution in acceptance rates in GS-UQS pairs. For instance, *Cats have whiskers* gets an acceptance rate of 75%, while the UQS *All cats have whiskers* gets an acceptance rate of 33%. Then, while *Bees have wings* gets a similar acceptance rate as *Cats have whiskers*, the UQS *All bees have wings* gets an acceptance rate of 67%. *People have two arms*, in turn, gets an acceptance rate of only 58%, with the UQS *All people have two arms* getting an acceptance rate of 25%. Additionally, in the same line, the old group did not do well on Controlgen, which should be rejected according to the current views about generics and how adults seem to understand them. The performance of the group on Controlgen was at chance levels.

Finally, turning to adults, they had an accuracy rate of 72% in UQS and of 79% in GS, consolidating the trend found in the old group. Compared to this latter group, the distribution of rejections of UQS was as broad, with an interesting qualification: 100% of rejections was reserved to those items that represented sub-kinds (we will have more to say about this below). The distribution of acceptances of GS was even broader, with 23% for *Pigs are pink* to 100% for *Giraffes have long necks*. In a similar vein as in the UQS condition, the lowest acceptance rates for GS was for the sub-kind cases. In contrast to the two child groups, adults performed as current views predicts on Controlgen, treating them as false generics, with 84% of accuracy.

While we have not been able to extract any meaningful patterns from the detailed look at the individual items for the young and old groups, as just mentioned, we have done so for adults. The most interesting emerging pattern in adults was their sensitivity to a class of predications: pink vs black pigs, white vs black sheep, spotted vs non-spotted cows. These are sub-kind predications: a sub-kind of pigs are black, a sub-kind of cows are spotted, etc. — whereas there is no sub-kind of rabbits that have only one ear. When presented with an exception to a sub-kind which at the same time is an exemplar of the alternative sub-kind, adults rejected the UQS 100% of the time and accepted the GS between 23% (*Pigs are pink*) and 54% (*Cows have spots* and *Sheep are white*) of the time. If we subtract the items *black pig*, *white cow* and *black sheep*, the performance on GS increases to 87%, while the performance on UQS drops to 65%. It seems that exception type may have had an effect on the behavior of adults, at least. This is a topic that should be further researched from an experimental perspective (there is some discussion on the generics literature about this sort of sub-kind or sub-species

generalizations, see e.g. Lazaridou-Chatzigoga et al., 2015; Leslie, Khemlani & Glucksberg, 2011).

3.4.2. Do these data support the GaD hypothesis?

After providing a panoramic view of the interpretation of GS and UQS across ages, let us turn to the Generics as Default (GaD) hypothesis, about which our data have something to say. Remember that it is supported by two kinds of experiments: experiments that ask participants whether a certain (universal or generic) statement is true or false, and experiments that ask people to recall what they have been told concerning some (universal or generic) feature of some kind. In both cases, it is observed that even adults have some strong tendency to interpret universals as generics, the explanation being that interpreting a generic is easier than interpreting a universal. Interpreting a universal requires an explicit formulation of it that goes into working memory and knowing whether the universal is true or false requires checking it against exceptions that have to be brought back to working memory from long term memory. Moreover, even when exceptions are accessible, the process of verifying a universal is costly: the exception(s) have to be held in working memory together with the universal statement while the verification process takes place. If the exception(s) is new to the participant, they also have to update their beliefs. In contrast, interpreting a generic is easy: one only has to access the conceptual structure associated with the kind and check whether a certain feature is one of the features of that category. This is summarized in (16).

(16) Unfolding the GaD hypothesis
- a. Generics are verified by accessing a conceptual structure.
- b. Verifying a generic is easy and does not involve working memory (System 1).
- c. Generics are held to be true even if we are aware that there are exceptions.
- d. Generics are true of features that are characteristic, highly prevalent, or striking.
- e. Universals are difficult to verify properly (System 2).
- f. Universals are many times misinterpreted as generics.

We have not collected the kind of data that can have a say in a. and b. However, concerning c., we have observed that generics are not always verified as they should be, so this is not true across development. We cannot say whether d. holds or not: we only tried with majority characteristic generics. The conclusion should be that majority characteristic generics are not universally taken to be true. Finally, concerning e. and f., we think we should resist the idea that since universals are costly, there is a tendency to interpret them as generics out of “laziness”, which is a key component of the GaD. For one thing, the hypothesis does not explain why only some UQS are rejected (are they more difficult to verify? If so, why?). However, apart from these main claims of the GaD hypothesis, and how our results relate to them, the hypothesis seems unable to predict some of the results of our study.

Let us consider a summary of the data that we have collected and judge whether they support the GaD. The fact that the accuracy in GS is higher than in UQS collapsing all age groups, together with the better performance of the young group in GS than in UQS, the worse performance of the young group in UQS than in the other two groups, and the relatively modest results for UQS in the old group and adults, could be interpreted as a natural consequence of the GaD hypothesis. It looks as though young children find GS easier than UQS (they accept more GSs in the face of exceptions and fail to reject most of the UQSs) and that all the age groups fail at correctly rejecting UQS in many situations.

However, other data points seem incompatible with the GaD. Especially, the high success of the young group in the GS condition (in fact higher than the old group and adults) shows there is a significant decline in the understanding of GS between the young and old groups and even extending to adults. If the greater accuracy of GS in young children, both with respect to the UQS condition and other ages, were to be interpreted as evidence in favor of GS being easy, we would be forced to entertain the idea that GS become more difficult across development. Moreover, the low rates in the Controlgen condition in the child groups clearly show that they behave unlike adults, so there seems to be a development trajectory in the correct understanding of GS, which is not predicted under the hypothesis that verifying a GS involves System 1 and, as such, the correct interpretation of GS should be preserved across ages. In fact, given the poor performance of young children in the UQS condition, the old group is adult-like in their interpretation of UQS, while the output in the generic types taken together (GS and Controlgen) suggest the two child groups are non-adult-like in the interpretation of generics. Finally, the mere fact that the old group and adults do not have at ceiling results for GS can also be viewed as data that go against the GaD (i.e., if GS are easy, we would expect higher accuracy than the one we observe). On the other hand, two factors should be pointed out, that is, the centrality of exceptions in our experiment and the fact that the participants had to reason in the face of a counterexample to the generalization, which may be responsible for these below ceiling effects.

Remember that GaD is an answer to the Generic Overgeneralization (GOG) effect that was observed in various experiments. Now, do we also find a GOG effect in our data? The low values in UQS in the young group could be evidence in favor of this, since it seems that they are interpreting them as tolerating exceptions. The GOG effect would be attenuated in the older group and in adults, but the low performance in UQS in these two groups may also suggest that they tend to interpret some UQS as GS (but see below).

3.4.3. Generics could be difficult after all: novel insights

Our main challenge is to explain whether our data constitute instances of the GOG effect. Since we are not bound by the GaD hypothesis, we should consider whether different features of the GS/UQS distinction prove decisive for the different age groups. Regarding the young group, two possible explanations might be driving their behavior: (a) they either do not pay much attention to exceptions (or they do not update their background knowledge), or (b) they do not master unrestricted UQS, thus treating GS and UQS as sharing the same truth conditions. In future research, we plan to explore this issue further by looking more closely into the distinction between restricted and unrestricted UQS.

On the other hand, the results for UQS in the old group and adults, not being at ceiling, could be explained as a tendency to disregard certain (amount of) exceptions as potential falsifiers for the generalization at hand. Even if the universal quantifier has been treated as a slack regulator in sentences such as *All the townspeople are asleep* (Lasersohn 1999), meaning that the UQS reduces imprecision, it could be that, even in these cases, a certain amount of imprecision or loose talk is tolerated. In fact, there is a tendency (not a strict difference) to be more permissive in GS than in UQS (i.e., to say “yes” in the case of a GS than to say “yes” to a UQS), so this is again compatible with an explanation in terms of loose talk. Note, in this respect, that it is not unusual to have to add modifiers such as *absolutely* to UQS to be sure that hearers get the unrestricted reading of the UQS right (e.g. *Absolutely all horses have four legs*). However, neither the GOG effect nor the loose talk hypotheses explain why only some UQS are taken to be compatible with exceptions. Apparently, some UQS are more exception-tolerant than others (e.g.

compare the tolerant *All horses have four legs* vs the intolerant *All people have two hands*). Again, this is an issue that calls for further research. Just as there are said to be different kinds of GS (majority characteristic, minority characteristic, etc. See (5) in §2), it is possible that generalizations expressed by means of a universal quantifier cluster together and form groups that need to be further explored.

All in all, the data from UQS have proven to not follow from previous theories. But so is the case of generic statements, which seem to be quite difficult for children, at least the ones exemplified in our Controlgen condition. According to the current accounts of GS, true generics are of five kinds (See (5) in §2). GS such as *Pizzas are square-shaped* or *Beds are round* are false because they do not fall under any of such types. According to Leslie and colleagues, a competent speaker should reject these generic-like statements after accessing the conceptual representation of the category and seeing that the predicated feature does not form part of such a representation. Little empirical research has been done on these generalizations. A possible account from the GaD hypothesis could be that the lazy System 1 is good at verifying generics, but not so much at falsifying them, as falsifying generics is not as fast and easy as verifying them. However, apparently, falsifying *Canadians are right-handed* is not particularly effortful. Alternatively, it could be that participants access a context-dependent *ad hoc* conceptual representation (Casasanto & Lupyan, 2015), where the predicated feature appears prominently (e.g. the *ad hoc* prototype of a butterfly is a blue butterfly, or the *ad hoc* prototype of a chair is a plastic chair). The fact is that, as of today, we lack any interesting explanation as to why minority generics like the ones in our Controlgen condition should be particularly difficult: the GaD account can concoct different apparently plausible hypotheses, but so can alternative approaches. For instance, we could take note that many participants in our experiments would complain when faced with a Controlgen by saying “Well, some do…”. To these participants, a direct rejection of the statement is too hard. As pointed out to us by B.H. Partee and A. Kratzer (p.c.), this is not surprising in view of the well-known homogeneity effects observed in plural definites (homogeneity goes back to Fodor, 1970, but see Križ, 2019 for a recent overview), which also seem to apply to generics (Cohen, 2004; Löbner, 2000; Magri, 2012; Križ, 2015). That is, sentences with definite plurals and generic statements and their negations do not have complementary truth conditions. Thus, upon hearing or reading a sentence such as “The guests are here” when some of them have not arrived, the sentence cannot be considered true or false (and likewise for its negation, when some have arrived). It is a presupposition failure that stems from the fact that homogeneity is needed — i.e., “pretty much” the totality of the guests has to be present in order for the sentence to be considered true. It seems plausible to think that, in our study, participants were faced with the challenge of evaluating statements whose falsity conditions are unclear. While adults seem to have taken the rejection route, children may not have had the same reaction in the face of a presupposition failure. While striking generics seem to be a problem for this view, this is by itself an interesting topic for future research. The important point, in any case, is that, whatever the cognitive process that is responsible for the non-adult-like performance of children at least as old as 9, such a process is absent in adults. Given that adults’ performance coincides with what is standard in the literature, the conclusion is that, looking at our three groups, only adults master all aspects of generic meaning. Pending some explanation from the GaD approach about children’s performance, the provisional conclusion is that generics are not as easy as the GaD claims.

As mentioned at the outset, we have proposed a design that made exceptions salient. This decision was intended as a means to establish a clear divide between GS and UQS,

and to alleviate the children's processing cost coming up with potential exceptions themselves. This must have played a role in the non-optimal results both in the GS and UQS conditions in the old group and adults, though. It should be further explored whether seeing the image of an exception has helped participants reject UQS. In the case of GS, exceptions should not affect their acceptance. But what if they have? It would compromise the idea that majority characteristic generics are taken to be true even when we know they have exceptions (see Lazaridou-Chatzigoga et al., 2019, for a similar point). In some cases, making the exception salient seems to reverse the putative GOG effect, making subjects interpret GS as UQS, possibly by bringing into the fore (i.e., into conscience) exceptions that would otherwise be in the background and considered irrelevant.

Let us conclude by assessing the value of collecting experimental data from a language other than English (here as an example of the majority language used in experiments on generics). We have observed in our experiment that the accuracy of GS is higher than UQS, although generics are not unmarked in Spanish. As pointed out in §2, though, UQS contain DET.PL in their structure, so they could be said to be more complex anyway. However, remember that GS are ambiguous between an exemplar and a generic reading, and this could in principle be a source of confusion in language learners. We have not encountered any evidence in favor or against this. In any case, difficulties in the interpretation of generics in our design do not seem to arise as a result of a morpho-syntactic ambiguity, but rather of the complexity of the meaning of genericity, i.e., determining the truth conditions of a generic statement. While in Gelman et al. (2016) the absence of cross-linguistic differences was analyzed on the basis of a recall experiment, here we have provided Spanish data directly on the interpretation of GS and UQS. While we have not carried out any comparisons, we have proposed a design that can be replicated in other languages. If the difficulties have to do with the truth conditions of GS and UQS (given explicit or implicit domain restrictions and different kinds of properties), in view from works such as Lazaridou-Chatzigoga et al. (2019), we may expect differences with respect to various quantifiers present in each language inventory, as well as some commonalities in the overall results in the comparison between GS and UQS.

4. Conclusions

In the present paper we have carried out an investigation about generalizations in Spanish-speaking children of two age groups (and a corresponding group of adult controls). Building on previous work by Lazaridou-Chatzigoga et al (2013, 2019), which addressed Leslie and Gelman's "Generics ad Default" (GaD) hypothesis, we have proposed a design that could test differences between generic statements (realized as definite plurals in Spanish) and unrestricted universally quantified statements, when the participant was faced with a photograph of an individual failing to support the generalization.

The data that we have collected does not support the GaD view. In fact, it describes an interesting picture yet to be fully understood. We would like to emphasize that, while it had been established that 4-year-olds were able to comprehend (restricted) UQS in an adult-like manner, we have found out that they are not adult-like in the interpretation of unrestricted UQS. In fact, by comparing three age groups, we have been able to spot an age group, namely 8/9-year-olds, as having certain adult-like behaviors (for instance in the interpretation of UQS) and non-adult behaviors (for instance in the interpretation of generic-like statements about properties exhibited by few instances of the kind).On the

other hand, it is quite interesting that the two child groups do not reject these latter statements, thus suggesting that generics are not as easy as one may think after all.

Finally, in this study we have analyzed behavioral results from a forced-choice task. However, in view of the apparent mismatches observed in the literature between the information collected from behavioral tasks and e.g. reaction times, we believe that processing data should be key in further informing us on whether the GaD hypothesis can be rejected. Hence, a natural follow-up study on this research question should take the relevance of processing data into consideration.

Acknowledgments

We would like to express our gratitude to Prof. Toben H. Mintz and three LLD reviewers for their suggestions, corrections and requests for clarification. Thanks also go to the audience of Experiments in Linguistic Meaning 1, where a previous version of this work was presented. This research has been partially supported by projects VASTRUD (PGC2018-096870-B-I00) and PROLE (PGC2018-093464-B-I00), funded by the Ministry of Science and Innovation (MCI) / Spanish Research Agency (AEI) and the European Regional Development Fund (FEDER, EU).

5. References

Barberán-Recalde, T. (2019). *The Acquisition of Basque and Spanish Quantifiers: An Empirical Study*. PhD dissertation, UPV/EHU.

Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive psychology*, 58(2), 195-219.

Bates, D., Mächler, M., Bolker, B. M., & Walker S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Casasanto, D. & G. Lupyan. (2015). All concepts are ad hoc concepts. Margolis, E., & Laurence, S. (Eds.). *The conceptual mind: New directions in the study of concepts*. Cambridge, MA: MIT Press, 543-566.

Cohen, A. (2004). Generics and mental representations. *Linguistics and Philosophy*, 27, 529–556.

Dahl, Ö. (1995). The marking of the episodic/generic distinction in tense-aspect systems. In G. Carlson & F. J. Pelletier (Eds.) *The Generic Book*. Chicago: Chicago University Press. 412-425.

Farkas, D. F., & De Swart, H. (2007). Article choice in plural generics. *Lingua*, 117(9), 1657-1676.

Fodor, J. D. (1970). *The linguistic description of opaque contexts*. (PhD Thesis), MIT, Cambridge, Mass.

Gelman, S. A. (2010). Generics as a window onto young children's concepts. In F.J. Pelletier (Ed.). *Kinds, things, and stuff. Mass Terms and Generics*. Oxford/New York: Oxford University Press. 100-123.

Gelman, S. A., Sánchez Tapia, I. & Leslie, S. J. (2016). Memory for generic and quantified sentences in Spanish-speaking children and adults. *Journal of Child Language*, 43(6), 1231-1244.

Hanlon, C. (1988). The emergence of set-relational quantifiers in early childhood. In F.S. Kessel (Ed.), *The development of language and language researchers: Essays in honor of Roger Brown*. New York: Psychology Press, 65-78.

Hollander, M. A., Gelman, S. A. and Star, J. (2002). Children's Interpretation of Generic Noun Phrases. *Developmental Psychology* 36 (6), 883–894.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.). *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press. 49-81.

Katsos, N., Andrés-Roqueta, C., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment Competent with the Pragmatics and Logic of Quantification? *Cognition*, 119: 43-57.

Katsos, N., Cummins, C., Ezeizabarrena, M., Gavarró, A., Kraljević, J. K., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences of the United States of America*, 113(33): 9244–9249.

Križ, M. (2019). Homogeneity effects in natural language semantics. *Language and Linguistics Compass*, 13(11), e12350.

Križ, M. (2015). Aspects of homogeneity in the semantics of natural language. (PhD thesis). University of Vienna.

Lasersohn, P. (1999). Pragmatic halos. *Linguistics and Philosophy* 75: 522–571.

Lazaridou-Chatzigoga, D., L. Stockall & N. Katsos. (2019). Contextualising Generic and Universal Generalisations: Quantifier Domain Restriction and the Generic Overgeneralisation Effect. *Journal of Semantics*, 36: 617–664.

Lazaridou-Chatzigoga, D., N. Katsos & L. Stockall. (2015). Genericity is easy? Formal and experimental perspectives. In Hansen, N. and Borg, E. (Eds.), *RATIO 28(4), Special Issue: Investigating Meaning: Experimental Approaches*, 470-494.

Lazaridou-Chatzigoga, D & L. Stockall. (2013). Genericity, exceptions and domain restriction: experimental evidence from comparison with universals. In Chemla, E., Homer, V. and Winterstein, G. (Eds.), *Proceedings of Sinn und Bedeutung 17*, École Normale Supérieure, Paris, 325-343.

Leslie, S. J. (2007). Generics and the structure of the mind. *Philosophical Perspectives*, 21: 375-405.

Leslie, S. J. (2008). Generics: Cognition and acquisition. *The Philosophical Review*, 117(1): 1-49.

Leslie, S.J., S. Khemlani & S. Glucksberg. (2011). All Ducks Lay Eggs: The Generic Overgeneralization Effect. *Journal of Memory and Language*, 65(1), 15-31.

Leslie, S. J. & S. Gelman. (2012). Quantified statements are recalled as generics: evidence from preschool children and adults. *Cognitive Psychology* 64(3), 186-214.

Löbner, S. (2000). Polarity in natural language: Predication, quantification and negation in particular and characterizing sentences. *Linguistics and Philosophy*, 23, 213–308.

Magri, G. (2012). No need for a dedicated theory of the distribution of readings of English bare plurals, *Proceedings of SALT 22*. Ithaca, NY: CLC, pp. 383–402.

Mari, A., Beyssade, C., & Del Prete, F. (Eds.). (2013). *Genericity* (Vol. 43). Oxford: Oxford University Press.

Pease-Gorrissen, M. (1980). The use of the article in Spanish habitual and generic sentences. *Lingua*, 51(4), 311-336.

Pelletier, F.J. (2010). Are All Generics Created Equal?. In F.J. Pelletier (Ed.). *Kinds, Things and Stuff. Mass Terms and Generics*. Oxford/New York: Oxford University Press. 60-79.

Strang, J. F., Anthony, L. G., Yerys, B. E., Hardy, K. K., Wallace, G. L., Armour, A. C., Dudley, K., & Kenworthy, L. (2017). The flexibility scale: Development and preliminary validation of a cognitive flexibility measure in children with autism spectrum disorders. *Journal of autism and developmental disorders*, 47(8), 2502-2518.

R Core Team (2020). R: A language and environment for statistical computing. Vienna, R Foundation for Statistical Computing. https://www.R-project.org/.

Zamparelli, R. (2002). Definite and bare kind-denoting noun phrases. *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4, 305-344.

Appendix

This appendix provides the materials (practice items, critical items and fillers) used in the experiment. For the critical items, each participant saw only one form of each critical majority characteristic statement (either generic or universal) and a non-supportive context (exceptional picture) for each given statement. All critical statements are given here in both forms: the GS and UQS form. Most of the statements are a subset of the items used in Lazaridou-Chatzigoga et al. (2019). We also included controls (Controlgen) and proper names (Names) as fillers.

Practice-items

1. Background: a picture of a brown donkey.
¿Dirías que esto es un burro marrón?
[English translation: Would you say that this is a brown donkey?]

2. Background: a picture of lorry.
¿Dirías que esto es un camión de bomberos?
[English translation: Would you say that this is a fire engine?]

3. Background: a picture of a white duck.
¿Dirías que esto es un pato blanco?
[English translation: Would you say that this is a white duck?]

4. Background: a picture of a metal spoon.
¿Dirías que esto es una cuchara de madera?
[English translation: Would you say that this is a wooden spoon?]

Critical items

1. Background: a picture of a cat without whiskers.
¿Dirías que los gatos tienen bigotes? Or ¿Dirías que todos los gatos tienen bigotes?
[English translation: Would you say that cats have whiskers? Or Would you say that all cats have whiskers?]

2. Background: a picture of a dog without tail.
¿Dirías que los perros tienen cola? Or ¿Dirías que todos los perros tienen cola?
[English translation: Would you say that dogs have tails? Or Would you say that all dogs have tails?]

3. Background: a picture of a three-legged horse.
¿Dirías que los caballos tienen cuatro patas? Or ¿Dirías que todos los caballos tienen cuatro patas?
[English translation: Would you say that horses have four legs? Or Would you say that all horses have four legs?]

4. Background: a picture of a bee without wings.
¿Dirías que las abejas tienen alas? Or ¿Dirías que todas las abejas tienen alas?
[English translation: Would you say that bees have wings? Or Would you say that all bees have wings?]

5. Background: a picture of a deer without antlers.
¿Dirías que los ciervos tienen cuernos? Or ¿Dirías que todos los ciervos tienen cuernos?
[English translation: Would you say that deer have antlers? Or Would you say that all deer have antlers?]

6. Background: a picture of a toothless man.
¿Dirías que las personas tienen dientes? Or ¿Dirías que todas las personas tienen dientes?
[English translation: Would you say that people have teeth? Or Would you have that all people have teeth?]

7. Background: a picture of a tuskless elephant.
¿Dirías que los elefantes tienen colmillos? Or ¿Dirías que todos los elefantes tienen colmillos?
[English translation: Would you say that elephants have tusks? Or Would you say that all elephants have tusks?]

8. Background: a picture of a disabled man (with a single arm).
¿Dirías que las personas tienen dos brazos? Or ¿Dirías que todas las personas tienen dos brazos?
[English translation: Would you say that people have two arms? Or Would you say that all people have two arms?]

9. Background: a picture of one-ear-rabbit.
¿Dirías que los conejos tienen dos orejas? Or ¿Dirías que todos los conejos tienen dos orejas?
[English translation: Would you say that rabbits have two ears? Or Would you say that all rabbits have two ears?]

10. Background: a picture of a black pig.
¿Dirías que los cerdos son rosa? Or ¿Dirías que todos los cerdos son rosa?
[English translation: Would you say that pigs are pink? Or Would you say that all pigs are pink?]

11. Background: a picture of a bald person.
¿Dirías que las personas tienen pelo? Or ¿Dirías que todas las personas tienen pelo?
[English translation: Would you say that people have hair? Or Would you say that all people have hair?]

12. Background: a picture of a black sheep.
¿Dirías que las ovejas son blancas? Or ¿Dirías que todas las ovejas son blancas?
[English translation: Would you say that sheep are white? Would you say that all sheep are white?]

13. Background: a picture of a short-necked giraffe.
¿Dirías que las jirafas tienen el cuello largo? Or ¿Dirías que todas las jirafas tienen el cuello largo?
[English translation: Would you say that giraffes have long neck? Or Would you say that all giraffes have long neck?]

14. Background: a picture of a chicken with four wings.
¿Dirías que las gallinas tienen dos alas? Or ¿Dirías que todas las gallinas tienen dos alas?
[English translation: Would you say that chickens have two wings? Or Would you say that all chickens have two wings?]

15. Background: a picture of a three-eyed-frog.
¿Dirías que las ranas tienen dos ojos? Or ¿Dirías que todas las ranas tienen dos ojos?
[English translation: Would you say that frog have two eyes? Or Would you say that all frogs have two eyes?]

16. Background: a picture of a white cow.
¿Dirías que las vacas tienen manchas? Or ¿Dirías que todas las vacas tienen machas?
[English translation: Would you say that cows have spots? Or Would you say that all cows have spots]

Controls (Controlgen)

1. Background: a picture of a white gorilla.
¿Dirías que los gorilas son blancos?
[English translation: Would you say that gorillas are white?]

2. Background: a picture of a yellow bird.
¿Dirías que los pájaros son amarillos?
[English translation: Would you say that birds are yellow?]

3. Background: a picture of a boy wearing glasses.
¿Dirías que los niños tienen gafas?
[English translation: Would you say that children (boys) wear glasses?]

4. Background: a picture of a girl with curly hair.
¿Dirías que las niñas tienen el pelo rizado?
[English translation: Would you say that girls have curly hair?]

5. Background: a picture of a blue butterfly.
¿Dirías que las mariposas son azules?
[English translation: Would you say that butterflies are blue?]

6. Background: a picture of a red leaf.
¿Dirías que las hojas son rojas?
[English translation: Would you say that leaves are red?]

7. Background: a picture of a green table.
¿Dirías que las mesas son verdes?
[English translation: Would you say that tables are green?]

8. Background: a picture of a pair of short pants.
¿Dirías que los pantalones son cortos?
[English translation: Would you say that pants are short?]

9. Background: a picture of a purple lettuce.
¿Dirías que las lechugas son moradas?
[English translation: Would you say that lettuces are purple?]

10. Background: a picture of a house with black roof.
¿Dirías que las casas tienen tejados negros?
[English translation: Would you say that houses have black roofs?]

11. Background: a picture of an Italian restaurant.
¿Dirías que los restaurantes son italianos?
[English translation: Would you say that restaurants are Italian?]

12. Background: a picture of a square-shaped pizza.
¿Dirías que las pizzas son cuadradas?
[English translation: Would you say that pizzas are square-shaped?]

13. Background: a picture of a Spanish omelette “pintxo”.
¿Dirías que los pintxos son de tortilla?

[English translation: Would you say that “pintxos” are of Spanish omelette?]

14. Background: a picture of a car with two doors.
¿Dirías que los coches tienen dos puertas?
[English translation: Would you say that cars have two doors?]

15. Background: a picture of a round bed.
¿Dirías que las camas son redondas?
[English translation: Would you say that beds are round?]

16. Background: a picture of a plastic chair.
¿Dirías que las sillas son de plástico?
[English translation: Would you say that chairs are made of plastic?]

Fillers (Proper names)

1. Background: a picture of the Eiffel Tower.
¿Dirías que la Torre Eiffel es alta?
[English translation: Would you say that the Eiffel Tower is high?]

2. Background: a picture of the Virgen Blanca Square (in Vitoria).
¿Dirías que la plaza de la virgen blanca está en vitoria?
[English translation: Would you say that Virgen Blanca Square is in Vitoria?]

3. Background: a picture of Celedón, a character from local festivities in Vitoria.
¿Dirías que Celedón tiene boina?
[English translation: Would you say that Celedon has a beret?]

4. Background: a picture of Peter Pan with friends (from the film adaptation).
¿Dirías que Peter Pan tiene amigos?
[English translation: Would you say that Peter Pan has friends?]

5. Background: a picture of Ibaiondo, a wellness-centre in Vitoria (with swimming-pools).
¿Dirías que Ibaiondo tiene piscina?
[English translation: Would you say that Ibaiondo has a swimming pool?]

6. Background: a picture of PortAventura,a theme park in Spain.
¿Dirías que PortAventura es un parque de atracciones?
[English translation:Would you day that PortAventura is a theme park?]

7. Background: a picture of Donald Duck.
¿Dirías que el Pato Donald tiene pico?
[English translation: Would you say that Donald Duck has a beak?]

8. Background: a picture of a Shakira, a famous singer.
¿Dirías que Shakira es cantante?
[English translation: Would you say that Shakira is a singer?]

9. Background: a picture of Alavés, a local football team from Vitoria.
¿Dirías que el alavés es un equipo de baloncesto?
[English translation: Would you say that Alavés is a basketball team?]

10. Background: a picture of Oscar, a Spanish man.
¿Dirías que óscar es chino?
[English translation: Would you say that Oscar is Chinese?]

11. Background: a picture of a Christmas, with trees covered in snow.

¿Dirías que las navidades son en primavera?
[English translation: Would you say that Christmas is in spring?]

12. Background: a picture of Olentzero, a coal merchant that brings presents at Christmas (similar to Santa Claus).
¿Dirías que el Olentzero lleva corbata?
[English translation:Would you say that Olentzero wears a tie?]

13. Background: a picture of Donald Trump.
¿Dirías que Trump es moreno?
[English translation: Would you say that Trump is dark-haired?]

14. Background: a picture of Mickey Mouse.
¿Dirías que Mickey mouse es una ratoncita?
[English translation: Would you say that Mickey Mouse is a (female) mouse?]

15. Background: a picture of Captain James Hook, an evil character from the film Peter Pan.
¿Dirías que el capitán garfio es bueno?
[English translation: Would you say that Captain James Hook is good?]

16. Background: a picture of the Three Wise Men.
¿Dirías que los reyes magos son dos?
[English translation: Would you say that the Three Wise Men are two?]