Published in final edited form as:

J Speech Lang Hear Res. 2013 October ; 56(5): 1650–1661. doi:10.1044/1092-4388(2013/11-0317).

Using ratings to gain insight into conceptual development

Mary Alt¹, Christina Meyers², and Paul M. Alt³

^{1,2}University of Arizona

³Lower Mill Pond Works

Abstract

Purpose—We explored a novel technique with potential for assessing conceptual development. Participants rated how ‘normal’ to ‘really weird’ an image was in order to determine if: a) participants would rate images by amount of variation (slight/significant) from the standard image b) participants would treat variation related to different concepts equally c) there would be developmental differences in these ratings. Then we asked whether children with Specific Language Impairment (SLI) would demonstrate weaker conceptual skills based on their ratings.

Methods—Adults and school-aged children (with and without SLI) used a 9-point equal appearing interval scale to rate photographic images of animals. These included standard images and images that altered the animals’ shape, pattern, color, and facial morphometry.

Results—Significant differences in ratings were obtained for adults compared to typically-developing children, and children with SLI compared to their age-peers. This is in line with the expectation that adults have stronger representations than children, as do typical children compared to those with SLI. Participants differentially rated images that varied from the standard image (slight/significant) for all parameters except shape.

Conclusions—Probing conceptual representations without the need for verbal response has the potential for exploring conceptual deficits in SLI.

Keywords

Semantic; Conceptual; Typicality; Specific Language Impairment; Ratings

Introduction

Conceptual development is closely related to language development (Smith, 2003; Yoshida & Smith, 2005), yet can be challenging to measure in populations that have limited expressive language abilities (e.g. children with language impairment, people with motor speech disorders). Conceptual knowledge is what a person knows about an idea or object. We often need to measure the conceptual development of children with limited language skills in order to develop appropriate treatment plans. However, a person with limited expressive language abilities may not be able to describe his or her conceptual

¹Contact Author’s Information: University of Arizona, 1131 E. 2nd St., Tucson, AZ 85721, 520-626-6180, Fax: 520-621-9901, malt@email.arizona.edu.

representation. It would be inefficient and frustrating to design and implement an intervention plan to teach a child about a concept that she already has – but cannot tell you she has.

Children with Specific Language Impairment (SLI) are known to have sparse semantic representations compared to their peers. Semantic and conceptual knowledge are closely linked in that conceptual knowledge underlies semantic knowledge. One characteristic of children with SLI is weaker vocabulary skills than peers (e.g., Gray, Plante, Vance, & Henrichsen, 1999). Importantly, this is not only reflected in a smaller lexicon, but also in decreased semantic skills. For example, Alt and colleagues have found that children with SLI were not able to recognize as many semantic features (e.g. color, pattern, shape, presence of eyes) as typically-developing peers given the same exposure to novel words and their referents (Alt, Plante, & Creusere, 2004; Alt & Plante, 2006). The difficulties appear to extend beyond the initial stage of word learning. Sheng and McGregor (2010) found a range of abilities in lexical-semantic organization for 7-year-old children with SLI, and identified a subgroup with significant impairment in this skill set that was associated with word-finding deficits. McGregor, Rost, Guo and Sheng (2010) found semantic deficits in school-aged children with SLI in terms of their ability to use appropriate, semantically-dictated word order and explain the meaning of novel compound words. McGregor, Newman, Reilly and Capone (2002) have also found that naming errors were associated with less richly encoded semantic information. When they asked children to draw pictures of words, the drawings of the words with naming errors were far less detailed than those of that were named correctly. McGregor et al. (2002) found that children with SLI made more naming errors than unimpaired peers, and thus likely had more words with weak semantic representations. Even for typically-developing children, word retrieval appears to be linked to depth of semantic encoding, with greater depth leading to more efficient retrieval (e.g. Capone & McGregor, 2005). The difficulties with encoding and depth of semantic knowledge noted for children with SLI likely reflect a less-developed conceptual knowledge.

One method of measuring conceptual knowledge has come from rating the typicality of category members. However, this method often involves language (e.g. rating words), thus limiting its utility for children with limited language skills because the results would confound conceptual knowledge with linguistic knowledge. This paper presents a novel approach in two studies that explore the use of a procedure that elicits ‘weirdness’ rating of visual information contained in photographic images, a technique that has potential for assessing the conceptual knowledge of people with limited expressive language.

How Might ‘Weirdness’ Ratings Work?

Insight into conceptual knowledge might be gained by asking a person to rate an image’s weirdness. Let us use an example of rabbits to illustrate how this might work. In nature one finds both brown and white rabbits, as well as rabbits with long floppy ears and rabbits with ears that point up. The perceptual differences (and similarities) between them are quite obvious. Their relatively similar shapes, component parts (e.g. four legs, tails), and textures (e.g. short or long fur) are easy to visually perceive, as are their most striking differences (i.e. fur color and ear type). If we were to show people pictures of either a brown rabbit with

long floppy ears or a white bunny with pointy ears, the expectation would be that they would rate either image as a 'normal' or standard exemplars of a rabbit. That either type of rabbit would be deemed acceptable assumes that the rater has a conceptual representation of rabbits that includes more than a single exemplar. If a child had only had experience with brown rabbits and therefore had a less-enriched conceptual representation of 'rabbitness', that child might rate the white rabbit as atypical or 'weird', because he or she did not realize it was perfectly normal for rabbits to be white. Whether that level of knowledge about rabbits resulted from lack of experience or poor learning is not relevant. Any rating other than 'normal' would provide evidence that the rater did not have a conceptual understanding of rabbits that was broad enough to encompass what is a valid characteristic of rabbits – specific variations in color.

Developmental Differences in Ratings

Developmentally, we might expect some differences in ratings as described in the example above. The differences may be simply due to the fact that children's conceptual knowledge is less fully-developed than it will eventually become. We must remember that some of children's information about animals comes through exposure to cartoons, stuffed animals, and other exemplars that are far less detailed, or frankly different, than real-life versions of those items, leading to differences between a child's representations and those of an adult. For example, a child may have a pink stuffed bunny. A pink bunny is clearly not found in nature. If children are building their conceptual knowledge based on exemplars like this, we might expect them to demonstrate their conceptual weakness in two opposing ways. First, they might inappropriately rate a pink bunny as normal, despite the fact that 'pinkness' is not a typical characteristic of a real bunny. However, the lack of well-developed conceptual knowledge in children might also be seen in the other direction. Children might be more likely to rate exemplars with typical characteristics of animals as less 'normal' than adults. Presumably, the child's toy bunny does not have long sharp nails like real rabbits do. Therefore, if they see a picture of a real rabbit, which is not as cute as their toy rabbit, and has some scary-looking nails, they might perceive the picture of the real rabbit as weird because their lack of experience with real animals does not allow them to recognize long nails as a typical characteristics of a rabbit. Adults, with well-developed conceptual representations, would perceive the same picture as 'normal'.

There is evidence for developmental trends in children's rating of prototypicality, in which younger children are less likely to include less-typical exemplars as members of a category than older children or adults (Bjorkland, Thompson, & Orland, 1983; Meints, Plunkett, & Harris, 1999). Rhodes, Brickman, and Gelman (2008) have also reported on developmental trends in which younger children are less likely to use within-category variability for making inductions than adults, and may "focus on certain individual exemplars as standing in for the category as a whole" (p.554). Using evidence from a fast-mapping study, Xu and Tenenbaum (2007) suggest that children may have a different "hypothesis space" than adults in terms of mapping perceptual features and words. We also see this phenomenon mirrored linguistically in the patterns of over- and under-generalizations that young children exhibit when they call all animals "doggie" (overgeneralization) or only their family pet "doggie" (undergeneralization) (e.g. Anglin, 1978; Clark, 1973). Funnel, Hughes, and Woodcock

(2006) found developmental differences related to the type of information children know. In their study, younger children showed an advantage for naming, while older children showed an advantage for answering knowledge-based questions about objects. Almost all of these findings relate to children using or developing conceptual knowledge, suggesting that a ratings task would reveal developmental differences in conceptual knowledge.

Weirdness Ratings and Prototypicality Ratings

Although they both require participants to have a core conceptual representation, deviance (weirdness) ratings differ from the more commonly used prototypicality ratings. Ratings of prototypicality have traditionally been used as a window into the nature of superordinate category organization. The traditional outcomes of these ratings are typicality effects. For example, in a series of nine experiments, Rosch (1975) created categories of items that had a good, intermediate, or poor fit with a superordinate category. Rosch found that items that were good exemplars of superordinate categories facilitated priming more than items that were poor exemplars. Typicality effects have been found for adults such that items of higher typicality result in faster reaction times than items of lower typicality (e.g. Holmes & Ellis, 2006; Kiran & Thompson, 2003; Rosch & Mervis, 1975). Also, Crowe and Prescott (2003) report on findings that, in category production tasks, more prototypical items are produced earlier in the task and are produced by more people. However, prototypicality ratings are not as useful for determining whether particular features are part of a person's conceptualization of an item (e.g. does this particular chicken have the characteristics we would expect of chickens in general), or how richly encoded a person's representation of a concept is.

In contrast to prototypicality ratings, the 'weirdness' rating system used in the present studies asks people to rate images on a numerical scale that is anchored as 'normal' on one end and 'really weird' on the other. In order to perform this task, a person must evaluate an image in comparison to their underlying conceptual representation. The rater's internal conceptual representation, which is the benchmark for this rating system, is not necessarily based on a single image or exemplar in the way prototypicality ratings are. This cannot be the case because a mature representational system will allow for normal variation in the representation of any given class of items (e.g. both brown and white rabbits are possible). Rather, a mature conceptual representation is a compilation of common characteristics that are represented in the class of item.

The Role of Variation

Despite their reliance on individual prototypes of categories, experiments using prototypicality ratings have provided information suggesting variability plays a role in conceptual representation. Posner and Keele (1968) asked adults to learn abstract patterns, which included prototypes represented either by consistent or highly variable training exemplars. In addition to finding a typicality effect, the authors verified that the variability of training exemplars was important to developing a schema that was broader compared to a schema built upon prototypical exemplars only. When adults were trained with high variability exemplars, they were much better at recognizing highly distorted, but legal, exemplars of the pattern as valid members of the correct set.

The same phenomenon was found in work on the visual typicality effects in adults. Kiran (2007; Kiran & Thompson, 2003) found that adults with aphasia who were trained on semantic features in order to treat lexical retrieval deficits were able to generalize lexical labels to untrained exemplar. However, they only found generalization when participants were trained on items that were thought to be atypical for a category. Those patients trained on typical items did not generalize. This atypical exemplar advantage was thought to be due to the complexity of the stimuli, with the idea that the atypical items trained participants on category-central features plus the atypical additions, whereas the typical items only trained category-central features.

The role of variation appears to have a developmental trend. Bjorkland et al. (1983) found that the words younger children omitted from category sets were those that were more atypical. Thus, younger children's categories permitted less variability. Rhodes and colleagues (2008) verified that younger children (age 6) were less able to use diversity in their categorizations than adults. Hence, the ability to accept legal, but significant variability in a category may be a signal of a broader and more conceptually developed schema. Therefore, a mature conceptual schema should lead to acceptance of a standard image as 'normal', even if other variants of the category would also be considered 'normal'. To relate this to our rating task, if people have a conceptual representation that is rich enough to handle variants, then if one of those variants is presented as a standard image (e.g. either a brown or white rabbit), they should be able to accept the variant as a standard.

The Role of Language

Work on prototypicality also highlights the influence of language on how children respond to exemplars. For example, Southgate and Meints (2000) found a developmental trend in which 2-year-olds would accept a less frequent exemplar of an animal (e.g. ostrich v. sparrow), but only when the category name (i.e. 'bird') had been mentioned. Capone Singleton (in press) found that 2–3-year-old children were less able to generalize names to new exemplars of a novel category when those new exemplars differed from the original exemplars initially provided to define the category. Therefore, the language used to define items appears to be related to the typicality of the exemplars of that category. Despite the fact that their model of word learning intentionally did not provide for interactions between language and categorization, Mayor and Plunkett (2010) acknowledge the evidence that "... labels can influence the process of categorization..." (p. 42). Some authors go even farther, and have found that language, specifically object names and functions, actually directed infants to learn new categories (Booth & Waxman, 2002). This makes it all the more important to control for language issues when trying to measure the conceptual knowledge of children with known language challenges.

Insight into Characteristics of Learning

The use of 'weirdness' ratings offers another advantage over prototypicality ratings; the former has the potential to provide insight into item characteristics that affect learning. Studies of word learning have clearly shown that the ease with which we learn certain words is tied to the statistical properties of the words themselves. For example, statistical differences in phonotactic probability (the probability of occurrence of a certain sound

sequence within a word or syllable) (e.g. Storkel, 2001; 2003), prosodic patterns (e.g. Mattys, Jusczyk, Luce, & Morgan, 1999), neighborhood density (the number of words that differ from a given/target word by only one phoneme) (e.g. Metsala, 1997), and semantic set size (the number of words that are meaningfully related to a word) (e.g. Storkel, 2009) have all been found to have effects on lexical learning or recognition. The regularity of the word's structure can determine how easy it is to learn. In some cases, similarity to the properties of the majority of other words in a language facilitates learning (e.g. Alt, 2011). In other cases, the relative novelty of a low probability word appears to prompt learning (e.g. Gray & Brinkley, 2011). The point is that a novel word's relative distance from typical can make it salient. This type of information about lexical stimuli has been used to assess learning differences in impaired populations (e.g. Alt & Plante, 2006; Munson, Kurtz, & Windsor, 2005) and across development from infants (e.g. Jusczyk, Luce, & Charles-Luce, 1994) to adults (e.g. Edwards, Beckman, & Munson, 2004). However, we currently do not have information about whether or not the distance from a visual standard would influence learning in the same way as distance from an auditory standard does.

What Prompts a Rating of 'Weird'?

One key difference between auditory and visual input is that we simply do not have information on what constitutes a meaningful difference based on variation in visual exemplars. While there is an enormous literature on psychophysical perception, these data are not relevant to the current question. Our assumption is that participants will be able to perceive differences between the images we present. We are currently interested in whether participants will be able to reliably discern a slight versus a significant distance from a visual standard. Once we answer that question, we can ask clinically-compelling questions like 'Will differences in distance from visual standard have a meaningful effect on learning?' In addition, we are interested in what types of manipulations affect a person's judgment of typicality. Bahrack, Clark and Bahrack (1967) created variants of prototypical drawings of items and had adults rate them. Indeed, adults did show a range of ratings relating to these variants, indicating that some items were found to be less typical than others and, importantly, items were not simply rated as typical or not. However, the variants were simple monochromatic line drawings that were described as varying details in the drawings. The nature of the variation was not analyzed. Before moving forward with using ratings to determine conceptual knowledge, one must first determine what types of variations people tend to accept as representative of the same concept. The extant literature does not address this issue. In this study, we ask participants to rate standard images and images that are systematically varied from the standard to determine the relative acceptability of a range of deviations from the standard image.

The Current Studies

Prototypicality tasks have provided important information about the nature of the organization of superordinate semantic categories. However, they provide very limited information about specific aspects of the rater's conceptual representations related to any given individual item and how typicality at this level relates to learning and information processing. A rating of an image's distance from standard, legal exemplars, rather than a rating of its goodness-of-fit to category membership, has the potential expand the knowledge

base concerning conceptual development and is a central element of the studies presented herein.

Experiment 1

The first experiment was designed to examine how participants rate images of animals that have variations in one of several types of features. Differences in ratings between adults and children would be consistent with the known developmental differences in conceptual knowledge. A demonstration of differences in ratings for slight versus significant image variants would support development of a visual analog to concepts like phonotactic probability in the auditory domain.

We hypothesized that if ratings for individual images reflect underlying conceptual knowledge, there should be a difference in the ratings for standard images for children versus adults, just as in the prototype-judgment literature. If this hypothesis is supported the method would provide a means to assess conceptual knowledge free of linguistic demands. Second, we expected that all participants would be able to discriminate between three levels of weirdness: standard (normal), slight variation (kind of weird), and significant variation (really weird). This finding would add to the literature by introducing a novel method to measure the effects of differences in visual dimensions on learning. This is important to semantic learning because visual referents provide the semantic basis of lexical items.

Methods

Participants

Adults: Fifty-nine adults (30 males) participated in the study. The average age of the men was 26 years with a range of 18–39 and the average age of the women was 21 years with a range of 18–37. Adults were recruited from the University of Arizona campus and the surrounding Tucson community.

Children: Fifty-nine typically-developing children (31 boys) aged 6–9 years participated in the study. The average age of the boys was 8;0 (years;months), with a range of 6;0–9;6, and the average age of the girls was 7;11 with a range of 6;5–9;7. Children were recruited from public, private, and charter elementary schools, as well as local libraries in Tucson, Arizona.

All participants reported that English was their primary language. Participants who spoke other languages but reported that English was their primary language were included in the study. Adult participants and parents of child participants responded to a questionnaire to confirm normal vision (or corrected-to-normal vision), and normal hearing, speech, language, motor, and cognitive skills. Data for participants who indicated a history of special education services, seizures, or brain injury were excluded.

Stimuli Design—In order to accurately rate an item's weirdness, a person should be expected to have had some experience with the item, thus allowing them to demonstrate their level of conceptual knowledge. Therefore, we chose 20 animals that a majority of toddlers were reported to recognize (Dale, 1996), and that were appropriate for visual

presentation (e.g. small creatures, like insects, are typically not examined in great detail). The expectation was that people, beginning at toddler age, would be building conceptual representations of these animals, and thus their judgments would reflect this knowledge. Images that served as the ‘standard’ image were photographs of live animals that were available on the internet or were purchased from photographers that we used to construct the actual stimuli used in the experiment.

All standard images had to be full color and of high photographic quality. Specifically, the animals had to be in sharp focus, have good contrast and have details visible in both the highlights and shadows. The minimum size of any starting image had to be approximately 1,000 pixels $\times$ 1,000 pixels and have a minimum luminance resolution of 8 bits in each of the red, green and blue channels. Standard images had linear pixel resolutions ranging up to three times the minimum. Other considerations were equally important. The subject animal had to look normal or standard in size, shape, coloration and pose, as judged by the research team, who took environment into account (e.g. brown rabbits are more common in Arizona than white ones). Recall that we were interested in the effects of variation on an animal thus we selected images that had many of the typical characteristics for that animal. However, it was crucial to ensure that participants could identify the standard images as normal. We planned to test our judgments by examining the ratings for the standard images; those that were outliers would be removed from the analysis. Pose, in particular, was very important. The pose had to show all major features of the animal and a significant portion of the animal’s face. Therefore, a photograph somewhere in between a side view and a head-on view of a stationary animal was sought. Selected images were first cropped to make the animal the focus of the composition and then adjusted for uniform image quality.

Variations from Standard—We chose to study variations from the standard image in color, pattern, shape and facial configuration. In all cases the variants used were created from manipulations made to, or derived from (in the case of shape variation), the image of the standard animal. Seen as a collection, the standard animal and all its variants would be judged to be exemplars of the same standard animal. A slight variation and a significant variation were created for each concept – color, pattern, shape, and face—resulting in 8 variants of the standard for a total of 9 stimuli per animal. The research team chose the details of each variation and decided if it should be classified as a slight or significant variation, based on consensus. Some variations implemented relatively easily with certain animals and there was a great temptation to use it on many animals, but we did not to avoid confounding specific types of manipulations with levels of difference from the standard.

Color variation was produced by changing the color of one or more parts of the animal’s body. This might change a standard brown bear into a really weird bear, for example, if its fur were changed to blue. The color of the animal’s fur was changed without obscuring the hair detail or fur texture. In addition to uniform colors, we also utilized multicolor appearances. In some cases only part of the animal was given non-typical color, like an otherwise ordinary elephant with bright yellow tusks and tail. See Appendix A for an example of a typical animal and its variants.

Pattern variation is a bit more complicated to describe, but we are all familiar with exemplars of natural pattern in typical animals: smooth short hair, tiger stripes, feathers, geometric shells and giraffes' irregular, honey comb-like markings. Patterns were changed and made to blend in with the otherwise natural appearance of the animal. In all cases of pattern variation, regular, geometric patterns were used to suggest that this was not one of natural origin.

Shape variation in this study refers to a change in the outline or silhouette of the standard animal. An example of a shape variant is a duck with three legs. In every case variation represents a distortion of the standard shape. Proportionally scaling an animal larger or smaller would not distort and was not used as a shape variation.

Face variations are changes to the animal's face, which leave the outline shape or silhouette unchanged. In other words, all facial changes are interior to the outline of the standard animal. A deer looking directly at you may have the size of its eyes tripled creating a face variation while keeping the skull size constant. The additional constraint of keeping the animal's outline constant is included to isolate the subject's shape processing capabilities.

Procedures—The stimuli were presented one image at a time on a laptop computer using DirectRT software (Jarvis, 2006). Prior to beginning the task, participants were provided with a brief recorded training regarding the use of the rating scale. Participants were instructed to use numbers 1 through 9 on the keyboard to rate the image's weirdness. As anchor points on the scale, a rating of 1 was equivalent to "normal", a rating of 5 indicated the image was "kind of weird" and a rating of 9 was regarded as "really weird". Participants had a rating scale in front of them for the duration of the experiment in order to remind them of the anchor points. (See Appendix A.) Participants also heard instructions regarding a second task not reported on in this manuscript, which they were asked to do between rating images. In total, participants saw 220 images and rated 180 images, nine variants of each of the 20 animals. The additional 40 images included one typical image and one variant of each animal. This means the participants saw 40 of the images twice, but rated each image only once. Stimuli were shown in random order, and images to be described were randomly interspersed among the images to be rated. After completing the rating task, participants were shown the standard image for each animal and asked to name it.

Results

Before attempting to answer our main questions, we needed to ensure that our standard images were functioning as we expected. We omitted the data for the owl for all participants, because the standard owl was an outlier in terms of its mean score of 3.37 ($SD = 2.68$) compared to the group mean of 1.88 ($SD = .88$). We also omitted individual ratings for any animal that participants were not able to accurately name. This accounted for 6 out of 1,180 data points for the adults (e.g. llama for sheep) and 36 out of 1,180 data points for children (e.g. tiger for lion).

Gender—We had no specific predictions about gender, but wanted to be sure it would be safe to collapse across gender. Therefore, we ran an ANOVA with age and gender as the between subject measures and distance from standard (standard, slight variation, significant

variation) as our within subject measure. There was no main effect for gender ($F(1, 114) = .008, p = .92, \eta_p^2 < .001$) however there were interactions between age and gender ($F(1, 114) = 6.94, p < .001, \eta_p^2 = .05$). Tukey post hoc testing revealed no significant between-group differences for gender at $p = .05$ for any of these interactions. The interaction appeared to be driven by rating differences between women and girls. Therefore, we collapsed across gender for the remainder of the analyses.

Ratings—First, we were interested in whether participants would rate the modified images significantly differently from the standard images. We also wondered if there would be differences in ratings for the levels of variation (slight v. significant). If adults performed as the research team expected, that would validate our choice of stimuli. Finally, we were interested in whether or not there would be developmental differences in ratings. To answer these questions, we ran a mixed ANOVA with age as the between group variable (adult, child) and distance from standard (standard, slight variation, significant variation) as the within group variable. There was no significant difference for age ($F(1, 116) = 2.86, p = .09, \eta_p^2 = .024$), but there was a difference for distance from standard ($F(2, 232) = 1380.54, p < .001, \eta_p^2 = .922$). Post hoc t-tests showed that, at $p < .01$ with corrections for multiple comparisons, each rating was significantly different from the other (Standard $M = 1.88, SD = .88$; slight variation $M = 5.81, SD = 1.13$; significant variation $M = 6.24, SD = 1.16$). We only reported the combined means because this difference in variation was significant for both children and adults. Recall that a rating of 1 corresponded to normal and a rating of 9 corresponded to really weird. Thus, all participants were able to correctly identify the standard images as standard, and could also identify differences between slight and significant variations from the standard.

There was no significant interaction between age and distance from standard ($F(2, 232) = 1.716, p = .18, \eta_p^2 = .014$), however, given that we expected developmental differences, we tested each variation level with planned comparisons. The only significant difference was for rating the standard images ($F(1, 116) = 8.68, p = .003, \eta_p^2 = .069$). This F-test was significant even using statistical corrections to accommodate multiple tests. Children rated the standard image as less standard than the adults did. Please see Figure 1.

We were also interested in which specific variables may have been driving the differences in the levels of variation, and if adults and children reacted differently to any of the variables. To answer these questions, we ran a mixed ANOVA with age as the between-group variable (adult, child) and image type (color, shape, pattern, face) and variation level (slight variation, significant variation) as the within-group variables. There was no main effect for age ($F(1, 116) = .79, p = .37, \eta_p^2 = .006$), but there were main effects for image type ($F(3, 348) = 3.85, p = .009, \eta_p^2 = .032$), and variation level ($F(1, 116) = 166.28, p < .001, \eta_p^2 = .589$). There was a significant interaction between age and image type ($F(3, 348) = 12.59, p < .001, \eta_p^2 = .097$), but Tukey post hoc testing with an alpha of $p < .01$ did not reveal any between group differences directly related to a specific image type. Rather, there were within-group differences between images driving this interaction. Post-hoc t-tests with an alpha of $p < .01$ showed that there were no within-group differences related to image type for adults, but children rated shape ($M = 6.35, SD = 1.06$) as significantly weirder than color ($M = 5.69, SD = 1.27$), pattern ($M = 5.79, SD = 1.31$), and face ($M = 5.90, SD = 1.11$).

There was also a significant interaction between image type and variation level ($F(3, 348) = 15.27, p < .001, \eta_p^2 = .116$). Post hoc t-tests ($p < .01$) revealed that there were differences in ratings for slight and significant variations of each image type with the exception of shape. Slight and significant differences in shape were, statistically, not differentiated by adult or child raters.

Brief Discussion

This paradigm was designed to determine whether we could use weirdness ratings to measure differences in levels of conceptual development. The significant interaction between age and distance from standard found using planned comparisons for standard images supports this premise. As hypothesized, children rated standard images as less standard than adults. However, if children simply had a different concept of the rating scale, then we would expect all of their ratings to be significantly different than the adults for all ratings. This was not the case. Animal variants were not rated as more or less weird by children than adults. Therefore, we interpret the fact that children rated the standard items as weirder than adults did as evidence that children have less-developed conceptual knowledge of animals than adults do.

We feel confident that these ratings are based on conceptual rather than linguistic knowledge given that participants were not required to use language in any way to rate the items. It is also not surprising that there were not significant between-group differences for the variants of the animals. Although children may not know as many details about animals as adults do, developmentally, their general concepts of animals have been shown to have “underlying continuity” (Crowe & Prescott, 2003, p. 476) as measured through category production tasks.

Our second question was whether or not participants would rate slight and significant variations to the original images differentially or group all changes into a single “changed” rating. Both children and adults made statistically significant different ratings on slight versus significant variations, with the exception of shape. Interestingly, all of the features varied in this study were roughly equivalent in terms of cueing participants in to differences from the standard image. Only children found shape differences to be ‘weirder’ than the other features. The fact that participants attended to these features may not be surprising given that preschool children, even those with language impairment, can fast-map these features (Alt et al., 2004).

All participants provided statistically significantly different ratings for the slight vs. significant variants. Although the actual difference between the mean ratings for slight and significant variants was small (.49 points for adults and .38 points for children), it is actually greater than the differences typically found between high probability and low-probability stimuli in the auditory domain. For example, the recent difference between rare ($M = .08$) and common ($M = .15$) sounds sequences in an auditory study was .07 (Storkel & Lee, 2011). Further research is needed to determine whether these differences in visual stimuli are meaningful in terms of how they affect what people learn. Certainly, Kiran and colleague’s (2003 (2007) work suggests that strategically choosing stimuli for semantic learning tasks has the potential to improve learning. However, whether these same principles

apply to within-item differences versus within category differences is unknown, as is the effect of typicality for children who are in the process of developing their conceptual and semantic systems. The basic principles at play in the auditory domain (e.g. familiarity and practice lead to quicker production; competition effects slow down processing) could, conceptually, transfer to learning through the visual domain.

Experiment 2

Our findings from Experiment 1 suggested that ratings of standard images might be sensitive to the level of conceptual knowledge of the rater. If this hypothesis is accurate, then we should also see ratings differences between typically-developing children and children with Specific Language Impairment (SLI), who are known to have sparse semantic representations compared to peers. Specifically, if the weirdness rating task is sensitive to conceptual knowledge, then we would predict differences in typicality ratings for children with SLI, who have less-detailed semantic, and thus conceptual, representations than peers. Specifically, we would expect the children with SLI to rate standard items as weirder than children with typical language, because their relatively impoverished conceptual representations will make them less likely to accept normal details of the standard images as normal. Despite the fact that our task is language-free, we would also predict that weirdness ratings would be correlated with measures of general language, vocabulary, and non-verbal cognition, given that these are all factors likely to be associated with conceptual development. If these hypotheses are supported, this work will offer a method that can be used to gauge the conceptual knowledge of those who do not have the language skills to explain what they know. This type of information is critical to creating appropriate intervention plans.

Methods

Participants—In order to test these hypotheses, we recruited 17 school-aged children (7;0–8;11) with SLI and 17 age- and gender-matched peers. Eighteen of the 34 participants were male and all children were native English speakers. Only one child from Experiment 1 was included in Experiment 2. That participant's original ratings were included in the data analysis, and additional inclusionary/descriptive testing was completed. We received parental consent and assent from each participant. All participants had to pass a color vision screening and near vision acuity screening to participate in the study. In order to be included in the SLI group, children needed a standard score of 85 or lower on the *Clinical Evaluations of Language Fundamentals –Fourth Edition* (CELF-IV). This test has a sensitivity of 100% and a specificity of 82% when this cut-point is used (Semel, Wiig, & Secord, 2003). Participants also needed to achieve a standard score of 75 or greater on the Nonverbal Index of the *Kaufman Assessment Battery for Children –Second Edition* (KABC-II; Kaufman & Kaufman, 2004) in order to rule out intellectual disability. To be included in the typically-developing group (TD), children needed to pass the vision screenings, achieve a standard score of 86 or higher on the CELF-IV, and achieve a standard score of 75 or higher on the K-ABC II. Children were also excluded if a parental questionnaire indicated that the children had any other diagnosed neurological condition or had received special education services. Descriptive testing was performed to assess children's receptive

vocabulary scores using the *Peabody Picture Vocabulary Test – Fourth Edition* (PPVT-IV, Dunn & Dunn, 2007). We also measured Maternal Level of Education (MLE) in years as a proxy of socioeconomic status. The groups were significantly different on this measure ($t = -4.27, p < .001$), so MLE was included as a covariate in the analyses. MLE was not significant ($F(3, 26) = .99, p = .40, \eta_p^2 = .013$). Details on participant characteristics are included in Table 1.

Procedures—The experimental procedure was the same as the one used in Experiment 1. However, the children in this study were involved in several days of testing in order to complete the standardized testing used for inclusionary and descriptive purposes. The experimental task was the last task administered to all participants. As in the first experiment, we omitted individual ratings for any animal that participants were not able to accurately name. This accounted for 21 out of 680 data points. The children with SLI made 11 naming errors and the children with TD made 10 naming errors. This rate of omission (3.08%) is nearly identical to the rate of omission for children in Experiment 1 (3.05%).

Results

To confirm whether or not children with SLI would rate standard images as less typical than peers with unimpaired language skills, we ran the same analyses we did for Experiment 1, with the addition of MLE as a covariate, allowing us to potentially replicate the results.

For the first analysis (Group (SLI, TD) $\times$ Distance from Standard (standard, slight variation, significant variation), we found no significant main effect for group ($F(1, 32) = .40, p = .52, \eta_p^2 = .012$). There was a significant effect for Distance from Standard ($F(2, 64) = 145.66, p < .001, \eta_p^2 = .810$) that was mediated by a significant interaction between group and Distance from Standard ($F(2, 64) = 4.34, p = .01, \eta_p^2 = .11$). At the level of the main effect for variation level, post hoc t-tests ($p < .01$, corrected for multiple comparisons) indicated that all children rated standard images ($M = 2.53, SD = 1.50$) significantly lower, (i.e., closer to standard) than they rated either variant. However, there were no differences between images with slight variations ($M = 6.00, SD = 1.20$) and significant variations ($M = 6.25, SD = 1.28$), although there was a trend in that direction. To examine the interaction effect, we used planned comparisons to test between group differences at each level. The only significant difference was for rating standard images ($F(1, 32) = 4.63, p = .03, \eta_p^2 = .126$). Children with SLI gave standard images significantly higher ratings than did children with unimpaired language skills (see Figure 1).

A second analysis was aimed at examining the effect of the type of image variation (color, shape, pattern, face) and the level of variation (slight difference, significant difference) on ratings by children with and without SLI. There was no main effect for group ($F(1, 32) = .191, p = .66, \eta_p^2 = .005$) or image type ($F(3, 96) = 1.83, p = .14, \eta_p^2 = .054$). There was a main effect for level of variation $F(1, 32) = 18.03, p < .01, \eta_p^2 = .36$. Post-hoc t-testing ($t = -.81, p = .41$) revealed that this was actually a non-significant trend towards difference between slight and significant differences, consistent with the findings in the first analysis. There were no significant interactions.

To test our hypotheses that vocabulary (PPVT-IV), language (CELF-IV), and nonverbal thinking skills (KABC-II) might be related to conceptual knowledge, we tested the correlations between these scores and participant's weirdness ratings. Using a Pearson Product Moment correlation, we found a significant negative correlation between receptive vocabulary scores ($r = -.41$, $p = .02$), language scores (CELF-IV; $r = -.39$, $p = .02$), nonverbal thinking skills (K-ABC-II, $r = -.44$, $p = .01$) and standard image ratings. In other words, the higher the receptive vocabulary, language, or nonverbal thinking scores, the lower score (or more standard) the standard images were rated. Even though the covariate for maternal level of education (MLE-our proxy for socioeconomic status) was nonsignificant, we also ran a correlation with MLE and weirdness ratings. It was nonsignificant ($r = -.31$, $p > .05$) indicating that the between group differences were unlikely to be due solely to socioeconomic differences.

Discussion

Insight into Conceptual Knowledge—The finding that children with SLI rate standard images as 'weirder' than unimpaired age-matched peers provides further support for the idea that image ratings tap into the conceptual knowledge of the rater. One needs to have rich knowledge of a concept in order to recognize an unaltered image as 'standard' even if that image is not the most representative in terms of the child's real-life experiences. The literature would suggest that the ability to include nonstandard exemplars into category membership suggests a more abstract, well-represented conceptual schema (e.g. Posner & Keele, 1968). In other words, adults who can correctly include legal, but less-standard exemplars of a category would be thought to have a broader conceptual knowledge of that category. If we go back to our rabbit example from the introduction, it is hard to argue that someone who only includes the all white New Zealand breed as members of the rabbit family is as knowledgeable about rabbits as someone who includes less common breeds like English Lop or French Angora. The ability to recognize the Lop and Angora as valid rabbit exemplars, despite the one's long ears and the other's long hair speaks to a conceptual representation that is detailed enough to recognize the essential features of 'rabbitness' and can handle variability in non-essential features.

Prototype Ratings and Weirdness Ratings: Two Sides of the Same Coin—Our findings are in line conceptually with those in the prototype literature. However, given that our task measures *deviation* from a standard and prototype ratings measure whether and item is thought to *belong* to a category, the effects of these two approaches are best measured on different item types (variants vs. standard images). In prototype rating tasks, conceptual knowledge is anchored to the prototype, and increased richness of the conceptual representation leads to acceptance of more diverse or atypical variants. Because even someone with the most basic knowledge of a concept should recognize a prototype, it is the atypical items that are most likely to reveal between-group differences in the conceptual richness of the rater's representation. In contrast, weirdness ratings, as we have used them, assess conceptual knowledge by determining if a person's conceptualization is rich enough to promote acceptance of any valid image, even if other variants are also valid. Therefore, if a rater's conceptual idea of the animal was overly narrow or simply different than that represented by an image, any given detail found in the standard picture could provoke the

rater to indicate that the image was slightly weird to them. Our nonstandard variants were clearly recognizable as the animals they portrayed, but not natural. No one would expect to see a bright purple rabbit anywhere in nature. Therefore, it is not surprising that these were not seen as standard by all our raters. Instead, the ratings for the standard items best represent the flexibility of the rater's conceptualizations. It is not surprising that these items produced the between-group differences.

Role of Variability—Our findings suggest a developmental trend in weirdness ratings of standard images such that adults rate standard images as more 'normal' than do children. This fits in well with findings from the prototypicality literature that older children and adults are better able to handle variability than younger children are. For example, Rhodes et al. (2008) asked 6-year-olds, 9-year-olds, and adults which type of information people would most benefit from to learn about animals. Given an option of sets of pictures that were not variable (e.g. two different golden retrievers) or variable (e.g. two different breeds of dogs), adults reliably chose the diverse set, regardless of the typicality of the pictures in the set, thus indicating that variability is perceived as important for learning. Six-year-olds, on the other hand, were more drawn to the typicality of the images in a set. Diversity was not as important as typicality. Although children recognized that diversity could help (e.g. they chose diverse typical sets as the best option- a Labrador and a golden retriever), typicality was most important. Six-year-olds selected diverse atypical sets at a lower rate than chance. These choices imply that while children recognize the value of variability of stimuli, there is a limit to what can be varied and still be perceived as informative to them. This reluctance of children to accept variability in Rhodes et al. (2008) mirrors our results. Specifically, in our study, in order to rate the standard images as standard, one needed to have enough variability in their conceptual representation of an animal to recognize that, for example, while a pigeon might not be the most prototypical bird, it is a perfectly acceptable standard bird variant. If children are less inclined to accept variations as they are building their conceptual representations, it is not surprising that their ratings might reflect a decreased ability to accept valid variants.

Role of Language—The use of weirdness ratings, a completely non-linguistic task, as a means to tap into conceptual knowledge is innovative and holds promise for a field where we must often measure conceptual knowledge in individuals who are limited by speech and language challenges. The fact that the ratings were negatively correlated with both language and nonverbal cognitive measures strengthens the idea that the ratings are related to conceptual development. While we acknowledge that these correlations do not prove that the ratings were directly measuring conceptual knowledge, they provide converging evidence together with the evidence presented earlier for interpreting the ratings as an index of conceptual knowledge. As discussed earlier, language influences categorization (e.g. Capone Singleton, in press; Southgate & Meints, 2000), and our results suggest that even when subjects are permitted to respond nonverbally, underlying language may have an influence on performance.

This study offers the field an innovative approach to assess conceptual knowledge without requiring a participant to speak. This alleviates the need for raters to have intact speech and

language abilities and is appealing for any population in which expressive speech or language is compromised. This might include children with SLI, second language learners, or those with motor speech issues. This type of information could provide a method for learning more about a person's conceptual knowledge. If we could estimate conceptual knowledge, we would be in a better position to interpret phenomena like poor vocabulary in which the root cause could be linguistic or conceptual in nature.

Future work might address whether such a task could be used with younger children, who might or might not understand the task instructions. Likewise, extension to adults with developmentally impaired language would provide further evidence of sensitivity of the method to conceptual knowledge. It would also be important to see if this effect could be found for other types of categories like family members or food. Finally, we would want to assure that any of the other tasks used with the experiment (e.g., picture descriptions) did not affect raters' performance on the rating task. Despite the need for additional work, the idea of gaining a gross measure of conceptual knowledge through a button press activity is an enticing one, particularly considering the lack of options currently available to assess conceptual knowledge in populations with limited expressive language.

Tools for Exploring Characteristics of Learning—It was also encouraging to know that participants rated levels of variations of color, pattern, and facial morphometry differently. This finding will allow us to pursue the question of how the visual image's distance from a standard might affect learning in people in whom conceptual development is a work in progress. It is important to note that there was not a significant difference between slight and significant variations in Experiment 2, although there was a trend in this direction. The sample size for the second experiment was fairly small, thus this may be an issue of power. However, a significant difference would need to be verified in any future stimuli used to pursue testing related to visual rating effects.

Some may be uneasy with the idea that the variants used for the weirdness ratings were made based on the personal judgments of the experimenters rather than strict psychophysical measurements. However, several factors mitigate this concern. First, if the research team's perceptions were not reflective of the larger population, there would not have been a statistically significant difference in ratings for the levels of variation. Second, it is often the case that naïve judgments match up well with the actual statistical properties of the stimuli. For example, in the auditory domain, ratings of word typicality have been found to correlate highly with actual ratings of phonotactic probability (Eukel, 1980). Recall that we expected all of the variations to be perceived by all participants. Our question was whether participants would subjectively rate differences in, generally, the same way. Because they did, we now have the possibility of using these differences in the degree of variation to test for typicality effects. However, there is more to learn about what other types of features people might attend to when making rating judgments, or whether more subtle differences might yield the same types of ratings. Recall that we intentionally made variations that we expected participants to be able to notice. Our stimuli were not subtle, and thus, with the exceptions of the standard images, were not representative of animals that anyone would be likely to find in nature. Although we did not find a general rating bias for a particular feature (with the exception of shape for children), it is possible that varying the animacy of the

objects to be rated would lead to different profiles of feature saliency (e.g., Jones, Smith, & Landau, 1991). Many other possibilities that could affect saliency remain unexplored (e.g. those affecting animate vs. inanimate objects). Clearly, more research needs to be done.

Conclusion

One challenge with assessing conceptual knowledge is that an examiner cannot see inside someone's head to know what that person is thinking. The best we can do is gather converging evidence that conceptual knowledge is actually being assessed. The weirdness ratings, a non-linguistic task, showed a developmental difference consistent with the idea that conceptual knowledge develops over the course of childhood. Ratings were lower for children with SLI, a condition associated with weak conceptual knowledge of vocabulary. Finally, there was a predictable relationship between vocabulary and other cognitive skills thought to relate to conceptual knowledge and the ratings on this task.

In addition to the converging evidence that weirdness ratings were useful for measuring conceptual knowledge, we also found that people were able to differentiate between gradations of variations of an image of an animal's color, pattern and facial features. This ability to differentiate allows for the possibility of future tests of the impact of a visual image's distance from standard on learning.

An important limitation to this initial study is that we are only measuring a part of the underlying conception of what any image or animal should be. A measure of one type of knowledge does not encompass someone's entire representation. For example, Funnel, Hughes, and Woodcock (2006) make the point that knowing that a tapir's nose is short and flexible "...does not tell you about the relative proportions of the nose to the head or whether the nose is relatively fat or thin." (p. 286). Therefore, detection of individual features does not indicate how a person integrates them into the larger representation. However, this experiment provides evidence of the promise of a new technique for examining semantic and conceptual development, particularly for people with expressive language limitations.

Acknowledgments

The work presented in this paper was funded by a National Institutes of Deafness and Other Communication Disorders Research Grant R03 DC006841 to the first author and support from the Ruth L. Kirschstein National Research Service Award (NRSA) Institutional Research Training Grants (T32DC009398) from the National Institute for Deafness and Other Communication Disorders to the second author. We would also like to acknowledge all the participants who took part in the study, the members of the L4 Lab for their help with data collection, and Elena Plante for her insightful comments on an earlier version of this manuscript.

References

- Alt M. Phonological working memory impairments in children with specific language impairment: Where does the problem lie? *Journal of Communication Disorders*. 2011; 44:173–185. [PubMed: 20943232]
- Alt M, Gutmann ML. Fast mapping semantic features: Performance of adults with normal language, history of disorders of spoken and written language, and attention deficit hyperactivity disorder on a word-learning task. *Journal of Communication Disorders*. 2009; 42:347–364. [PubMed: 19439319]

- Alt M, Plante E. Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2006; 49:941–954.
- Alt M, Plante E, Creusere M. Semantic features in fast-mapping: Performance of preschoolers with specific language impairment versus preschoolers with normal language. *Journal of Speech, Language, and Hearing Research*. 2004; 47:407–420.
- Anglin JM. From reference to meaning. *Child Development*. 1978; 49:969–976.
- Bahrick HP, Clark S, Bahrick P. Generalization gradients as indicants of learning and retention of a recognition task. *Journal of Experimental Psychology*. 1975; 75:464–471.
- Bjorkland DF, Thompson BE, Ornstein PA. Developmental trends in children's typicality judgments. *Behavior Research Methods and Instrumentation*. 1983; 15:350–356.
- Booth AE, Waxman S. Object names and object functions serve as cues to categories for infants. *Developmental Psychology*. 2002; 38:948–957. [PubMed: 12428706]
- Capone NC, McGregor KK. The effect of semantic representation on toddler's word retrieval. *Journal of Speech, Language, and Hearing Research*. 2005; 48:1468–1480.
- Capone Singleton N. Can semantic enrichment lead to naming in a word extension task? *American Journal of Speech Language Pathology*. (in press). 10.1044/1058-0360(2012/11-0019)
- Clark, EV. What's in a word? On the child's acquisition of semantic in his first language. In: Moore, TE., editor. *Cognitive development and the acquisition of language*. New York: Academic Press; 1973. p. 65-110.
- Crowe SJ, Prescott TJ. Continuity and change in the development of category structure: Insights from the semantic fluency task. *International Journal of Behavioral Development*. 2003; 27:467–479.
- Dale PS. Lexical development norms for young children. *Behavior Research, Methods, Instruments, & Computers*. 1996; 28:125–127.
- Dunn, LM.; Dunn, DM. *Peabody Picture Vocabulary Test*. 4. Minneapolis: Pearson; 2007.
- Edwards J, Beckman ME, Munson B. The interaction between vocabulary size and phonotactic probability effect on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*. 2004; 47:421–436.
- Funnell E, Hughes D, Woodcock J. Age of acquisition for naming and knowing: A new hypothesis. *The Quarterly Journal of Experimental Psychology*. 2006; 59:268–295. [PubMed: 16618634]
- Gray S, Brinkley S. Fast mapping and word learning by preschoolers with specific language impairment in a supported learning context: Effect of encoding cues, phonotactic probability, and object familiarity. *Journal of Speech, Language, and Hearing Research*. 2011; 54:870–884.
- Gray S, Plante E, Vance R, Henrichsen M. The diagnostic accuracy of four vocabulary tests administered to preschool children. *Language, Speech, and Hearing Services in Schools*. 1999; 30:196–206.
- Holmes SJ, Ellis AW. Age of acquisition and typicality effects in three object processing tasks. *Visual Cognition*. 2006; 13:884–910.
- Jarvis, BG. *Direct RT Precision Timing Software* [computer software]. New York: Empirisoft; 2006.
- Jones SS, Smith LB, Landau B. Object properties and knowledge in early lexical learning. *Child Development*. 1991; 62:499–516. [PubMed: 1914622]
- Juszyk PW, Luce PA, Charles-Luce J. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*. 1994; 33:630–646.
- Kaufman, A.; Kaufman, NL. *Kaufman Assessment Battery for Children*. 2. Circle Pines MN: AGS Publishing; 2004.
- Kiran S. Complexity in the treatment of naming deficits. *American Journal of Speech-Language Pathology*. 2007; 16:18–29. [PubMed: 17329672]
- Kiran S, Thompson CK. The role of semantic complexity in treatment of naming deficits: training semantic categories in fluent aphasia by controlling exemplar typicality. *Journal of Speech, Language, and Hearing Research*. 2003; 46:773–787.
- Kovack-Lesh KA, Oakes LM, McMurray B. Contributions of attentional style and previous experience to 4-month-old infants' categorization. *Infancy*. 2010; 17:324–338. [PubMed: 22523478]

- Mattys SL, Jusczyk PW, Luce PA, Morgan JL. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*. 1999; 38:465–494. [PubMed: 10334878]
- Mayor J, Plunkett K. A neuro-computational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*. 2010; 117:1–31. [PubMed: 20063962]
- McGregor KK, Newman RM, Reilly RM, Capone NC. Semantic representation and naming in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2002; 45:998–1014.
- McGregor KK, Rost GC, Guo LY, Sheng L. What compound words mean to children with specific language impairment. *Applied Psycholinguistics*. 2010; 31:463–487. [PubMed: 22611295]
- Meints K, Plunkett K, Harris PL. When does an ostrich become a bird? The role of typicality in early word comprehension. *Developmental Psychology*. 1999; 35:1072–1078. [PubMed: 10442875]
- Metsala JL. An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition*. 1997; 25:47–56. [PubMed: 9046869]
- Munson B. Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research*. 2001; 44:778–792.
- Munson B, Kurtz BA, Windsor J. The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2005; 48:1033–1047.
- Posner MI, Keele SW. On the genesis of abstract ideas. *Journal of Experimental Psychology*. 1968; 77:353–363. [PubMed: 5665566]
- Rhodes M, Brickman D, Gelman SA. Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*. 2008; 108:543–556. [PubMed: 18436200]
- Rosch E. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*. 1975; 104:192–233.
- Rosch E, Mervis CB. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*. 1975; 7:573–605.
- Sandberg C, Sebastian R, Kiran S. Typicality mediates performance during category verification in both ad-hoc and well-defined categories. *Journal of Communication Disorders*. 2012; 45:69–83. [PubMed: 22261305]
- Semel, E.; Wiig, EH.; Secord, W. *Clinical Evaluation of Language Fundamentals*. 4. San Antonio, TX: The Psychological Corporation; 2003.
- Sheng L, McGregor KK. Lexical-semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2010; 53:146–159.
- Smith LB. Learning to recognize objects. *Psychological Science*. 2003; 14:244–250. [PubMed: 12741748]
- Southgate V, Meints K. Typicality, naming, and category membership in young children. *Cognitive Linguistics*. 2000; 11:5–16.
- Storkel HL. Learning new words: phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*. 2001; 44:1321–1337.
- Storkel HL. Learning new words II: phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research*. 2003; 46:1312–1323.
- Storkel HL. Developmental differences in the effects of phonological, lexical, and semantic variables on word learning by infants. *Journal of Child Language*. 2009; 36:291–321. [PubMed: 18761757]
- Yoshida H, Smith LB. Linguistic cues enhance the learning of perceptual cues. *Psychological Science*. 2005; 16:90–95. [PubMed: 15686573]
- Xu F, Tenenbaum JB. Word learning as Bayesian inference. *Psychological Review*. 2007; 114:245–272. [PubMed: 17500627]

Appendix A. Examples of Stimuli and Weirdness Rating Scale

Figure 1.

Mean group ratings of images on a scale of 1 – 9 by level of variation with standard error of measure for both experiments

Table 1

Participants' Demographic Features and Means and Standard Deviations on Inclusionary and Descriptive Testing for Experiment 2

	SLI		NL	
	Mean	Standard Deviation	Mean	Standard Deviation
Age (years; months)	8;0	0;6	7;9	0;6
* Maternal Level of Education	13.43	1.15	15.73	1.79
* CELF-IV	74.25	6.49	102.00	8.32
* PPVT-IV	88.47	10.05	111.94	10.05
* K-ABC-II	95.64	13.31	108.58	17.39

* groups are significantly different from each other at $p < .05$

Note: CELF-IV= Clinical Evaluation of Language Fundamentals-4th edition; PPVT-IV= Peabody Picture Vocabulary Test-4th edition; K-ABC-II = Kaufman Assessment Battery for Children-2nd edition.