COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2014) 1–30

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12122

Overestimation of Knowledge About Word Meanings: The "Misplaced Meaning" Effect

Jonathan F. Kominsky, Frank C. Keil

Department of Psychology, Yale University

Received 18 February 2013; received in revised form 4 September 2013; accepted 9 September 2013

Abstract

Children and adults may not realize how much they depend on external sources in understanding word meanings. Four experiments investigated the existence and developmental course of a "Misplaced Meaning" (MM) effect, wherein children and adults overestimate their knowledge about the meanings of various words by underestimating how much they rely on outside sources to determine precise reference. Studies 1 and 2 demonstrate that children and adults show a highly consistent MM effect, and that it is stronger in young children. Study 3 demonstrates that adults are explicitly aware of the availability of outside knowledge, and that this awareness may be related to the strength of the MM effect. Study 4 rules out general overconfidence effects by examining a metalinguistic task in which adults are well calibrated.

Keywords: Knowledge; Lexicon; Word learning; Overconfidence; Metacognition

1. Introduction

Confidence about understanding the meanings of many familiar words may be misplaced. In particular, when "meaning" is understood as personally knowing the specific features that guide reference, one may be very overconfident. People may use words freely and refer successfully in communications with others and therefore assume they have rich mental representations of their meanings. Yet, in many cases, these meanings may be largely represented in other minds or distributed throughout a community. The idea that "meaning ain't in the head" is not a new one (Putnam, 1975). According to the proposed division of linguistic labor (see section 1.1.), much of the "meaning" of a word is distributed across a language community, rather than in the head of any given speaker. Here we document that, when meaning is not in the head, people often believe that it is

Correspondence should be sent to Jonathan F. Kominsky, Department of Psychology, Yale University, 2 Hillhouse Ave, New Haven, CT 06520. E-mail: jonathan.kominsky@yale.edu

all the same. Furthermore, we show that this bias is powerful and emerges early in development, and we suggest that it may be critical for early word learning. We argue that this misplaced sense of meaning may serve a functional role in adults as well, given that it may more legitimately represent when a pathway to meaning is available through the division of linguistic and cognitive labor, and when it is not. In other words, it may allow adults to understand when consulting an expert source will yield greater understanding and when it will not.

1.1. The division of linguistic labor

Being able to mentally represent detailed meanings for every word in a person's vocabulary would be an extraordinary achievement. Consider that the vocabulary of an educated native English speaker approaches 20,000 words (Goulden, Nation, & Read, 1990). Equally impressive is how adults attain such a level; children learn words at rates that, for some periods, can average more than eight per day (Bloom, 2000). Because a person with a vocabulary of 20,000 words would normally be considered capable of successfully producing and comprehending discourse with each of those words, that person might seem to have internally represented each of those meanings in fine-grained detail such that each non-synonym could be distinguished from every other based on mentally represented contrasting features. Yet successful use of words may not entail such representations. For example, Hilary Putnam, in analogy to the long-accepted divisions of cognitive and physical labor (e.g., Smith, 1776), stated the hypothesis of the division of linguistic labor:

> Every linguistic community ... possesses at least some terms whose associated "criteria" are known only to a subset of the speakers who acquire the term, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets. (Putnam, 1975, pp. 145–146)

According to Putnam, when people "acquire" a term, that is, add it to their vocabulary, much of the meaning and what distinguishes it from any other term is only available to the speaker from outside sources. One has the meaning only by virtue of being able to access relevant experts to disambiguate meanings. By this account, speakers are constantly embedding themselves in networks of deference that ground what may be very incomplete meanings in their heads. However, this entire process may be largely tacit and overlooked by most people as they use words in their daily lives. Success at achieving reference may be mistakenly attributed to knowing the complete meaning of a word directly when in fact one only knows it by virtue of knowing, or at least believing in, a chain of access to experts. To use one classic example from Putnam, most adults believe they fully understand the meaning of "gold" and may indeed refer successfully to gold and know some of its properties, yet they may only succeed because they trust that others with greater expertise in chemistry and atomic structure could always tell the difference between gold and other substances. In some cases this division of linguistic labor may be

explicit. For example, work in linguistics has found that Americans can list many different types of tree, but only report being able to identify around 50% of them by looking at them (Gatewood, 1983). However, in the following experiments we suggest that participants' knowledge may be even more limited than they realize.

1.2. Illusions of understanding

People are often unaware of the shortcomings in their own knowledge, and these gaps can depend on quite specific features of that knowledge. Consider, for example, the "Illusion of Explanatory Depth" (IOED) (Rozenblit & Keil, 2002). When asked to rate their understanding of a mechanical or biological system, adults will often rate themselves quite highly. However, when they provide an explanation, their explanations are often skeletal and incomplete, and the act of trying to produce it makes them aware of the gaps in their understanding.

Recent work has suggested that individuals vary in their susceptibility to the IOED based on how much they tend to deliberate, as measured by scores on the Cognitive Reflection Test (Fernbach, Sloman, Louis, & Shube, 2013), but the effect depends on the kind of knowledge involved and is not a matter of general overconfidence. People are quite accurate in gaging their own knowledge in certain domains, such as procedural or narrative knowledge (Rozenblit & Keil, 2002). Factors that make the IOED particularly strong for explanatory causal knowledge may include confusions with functional and mechanistic understandings, the ability to recover information from systems when presented with them, and misattributing that real-time recovery from inspection to having internally represented the explanation.

Recent work has extended this idea further. Alter, Oppenheimer, and Zemla (2010) argued that "IOEDs are likely to emerge when people mistake their mastery of the abstract characteristics of the concept for a belief that they understand the concrete aspects of the concept much more deeply than they actually do" (p. 437). Thus, an illusion of understanding may not be exclusively bound to explanations or causal systems, and in fact an analogous illusion can be found in people's intuitions about their ability to justify arguments in far more detail than they really can (Fisher & Keil, 2014). Here, we propose that these illusions extend to vocabulary, and furthermore that they result from a similar process to the one proposed by Alter et al. (2010) for the IOED. In the context of word meaning, "abstract" versus "concrete" is not the most apt way of describing this contrast. The relevant aspect of the "abstract" versus "concrete" distinction concerns levels of detail. "Abstract," in the context of the IOED, refers to a coarse level of detail, and "concrete" to a more fine level of detail. For word meaning, we propose that a similar effect occurs across two different levels of detail, which we call "common" and "distinctive." "Common" aspects of word meaning encompass details of meaning that are shared by many similar terms (e.g., "it refers to a species of animal; species are differentiated by certain intrinsic biological properties") and very general metalinguistic information (e.g., "word X is not the same thing as word Y"). "Distinctive" aspects of word meaning are those that distinguish one particular word from every other, even words with

very similar meanings (e.g., the specific intrinsic biological properties that distinguish one species from another). If people have knowledge of the common aspects of a word's meaning, or pathways to distinctive aspects of a word's meaning, they may mistakenly believe they possess the distinctive aspects of that word's meaning in their own minds.

One particularly relevant type of common information about word meaning is the knowledge that a word *has* a distinct meaning. Indeed, from a young age we are strongly inclined to assume that novel words refer to novel referents. This basic assumption has been researched as the Mutual Exclusivity Principle (Markman & Wachtel, 1988), the Principle of Contrast (Clark, 1983, 1987), and the Novel Name—Nameless Category Principle (Mervis & Bertrand, 1994). While all three principles refer to different concepts and make somewhat different predictions, they all start with the same fundamental idea: When people are exposed to a novel word, they assume it refers to something different than words that they already know. With this sense that two words have different referents, people may then assume that they know something about what distinguishes those referents, even when they do not.

1.3. The Misplaced Meaning effect

We therefore propose the "Misplaced Meaning" (MM) effect. To achieve successful reference, there is normally a subset of speakers in any language who should know the distinctive differences between a given pair of words. However, speakers who are not members of that subset, who may only possess the common-level knowledge that some differences exist, may erroneously believe that they also possess more distinctive-level knowledge when in reality they are only able to access it from an outside source. They may be able to access that outside source because of some sense of who the relevant expert is likely to be and what types of expertise they are likely to possess, as suggested by research on the division of cognitive labor (Keil, Stein, Webb, Billings, & Rozenblit, 2008). Speakers may mistakenly confuse knowledge of how to access distinctive aspects of meaning with actual knowledge of features that distinguish the two kinds.

The MM effect should be present, and probably stronger, in children. The assumption that novel words have novel referents has been studied most often in the context of language acquisition, and in some cases has been shown to moderate with age (Markman, 1991). Given that children seem to employ this assumption as a learning strategy and will let it dominate other strategies early on, they should very readily acquire the sense that two words mean different things. At the same time there is no reason to expect that they should know the details of that distinction when they first hear a novel word. Furthermore, children frequently overestimate their capabilities or knowledge (e.g., memory: Flavell, Friedrichs, & Hoyt, 1970) and the IOED is stronger in younger children (Mills & Keil, 2004).

In addition to general overconfidence, there are other reasons we might expect a greater MM effect from children. Children often think they have known all along information that they have just learned (Taylor, Esbensen, & Bennett, 1994), which may be part of a larger set of difficulties with source monitoring (Roberts & Blades, 2000). A

strong MM effect in young children may be a form of source monitoring difficulty, namely that younger children have an illusion of competence that arises from misplacing the source of knowledge that enables them to successfully refer. Children assume their success comes from their own knowledge, when in fact it exists through networks of deference. As soon as they start to use a new word, they may assume they knew it all along in their heads, when really they only "knew it indirectly." Successful use may cause them to misattribute the indirect source of information to one that is directly in their own minds.

From these two lines of argument, we predict that young children will show a stronger MM effect. Putting aside the broader cognitive bases of this prediction, if the MM effect is indeed stronger in young children, it may either be because young children think that they know even more distinctive aspects of meaning than adults or that they actually know fewer, or a combination of the two. In the studies that follow, these alternatives are teased apart. We also consider why this MM bias might be an adaptive way of coping with the enormous cognitive demands of learning new words.

Given that our predictions are strongly grounded in the idea of deference, an intuitive prediction might be that the MM effect should gradually emerge with development as children become more immersed in their culture and learn about expert sources and how to rely on them. In other words, one might expect that children might not even be aware of the necessity of deference to outside, expert knowledge, or inexperienced in using it. On the contrary, recent work has found that children are conscious of and reliant on networks of deference from a very young age. Even preschoolers have a sense of different domains of expertise with different bodies of knowledge that can be accessed (Keil, Lockhart, & Schlegel, 2010; Lutz & Keil, 2002). Furthermore, young children are intelligent users of these networks of deference, employing surprisingly sophisticated tools in evaluating the quality of expert sources (e.g., Koenig & Harris, 2005).

To further emphasize the role of deference, it is important to note why we call this the *Misplaced* Meaning effect, rather than the *Missing* Meaning effect. A "Missing Meaning" effect would suggest that adults simply think they know things that they do not. Rather, our argument is that they think they know things themselves *that they can access from outside sources.* Thus, the meanings do exist, but they are "misplaced" in the minds of others.

There might seem to be a tension between the view that young children are aware of and use deference and the claim even adults "misplace" knowledge in the minds of experts. We argue here that the tension is only illusory. As we will demonstrate, the awareness of the division of linguistic labor may in fact enhance the MM effect. The basis of the illusion we have proposed is that the availability of expert knowledge causes people to confuse some portion of accessible knowledge with possessed knowledge. We are not suggesting that they should confuse *all* accessible knowledge with possessed knowledge. Rather, we suggest that the more knowledge people think is accessible through experts, the more they think they must possess as well. Thus, people should still expect experts to know more differences than themselves, even while overestimating their own knowledge.

1.4. The current studies

In the four studies reported here, we investigate the MM effect. In Study 1, we demonstrate the MM effect in adults. In Study 2, we test the MM effect in children in kindergarten, second, and fourth grade, and investigate whether it is stronger in young children. In Study 3, we examined whether adults recognize that experts should know more than they themselves (i.e., explicitly acknowledge the division of linguistic labor), even while overestimating their own knowledge. We further investigated the relationship between expected expert knowledge and the magnitude of the MM effect. Finally, Study 4 investigates whether the MM effect in adults is a result of general metalinguistic overconfidence, and provides further insight into the role of common aspects of meaning in the MM effect.

2. Stimulus pre-testing

We picked 45 pairs of words to test. Twelve of the 45 were true synonyms, defined by dictionary and thesaurus listings. These were originally included as a control for a blind overconfidence effect. However, we made no predictions about whether children would be able to recognize them as synonyms, allowing for the possibility that early learning biases such as mutual exclusivity and the contrast principle might make the idea of true synonyms less appealing to younger participants.

The remaining 33 items were selected pairs of words that referred to similar but not identical things, and specifically did not include pairs of words that referred to extremely different things. There are many reasons for adding this constraint to our stimuli. One is purely practical: If we used pairs of extremely dissimilar words (e.g., "church" and "daffodil"), the sheer number of differences that a participant might know would take an extremely long time for them to write down, which is both impractical and introduces the risk that participants' actual knowledge would be underrepresented by them not having time to record all the differences they knew, which would be a source of type-I error for testing our hypotheses. There are further reasons to expect that participants might have difficulty listing differences between extremely different word pairs that would reflect problems with our design rather than an accurate assessment of their knowledge. The "structure-mapping theory" of similarity holds that there are "alignable differences" and "non-alignable differences," and recent work has supported the prediction that alignable differences are more salient (Sagi, Gentner, & Lovett, 2012). Thus, word pairs that are very different are likely to have more non-alignable differences, which would be more difficult for participants to report even if they possessed knowledge of the differences.

On the basis of informal piloting, we selected some word pairs that had specific differences that almost everyone knew and other word pairs that did not. As explained below,

the variation in magnitude or frequency of the MM effect across these two classes would inform different accounts of the effect. We then conducted two pilot studies and divided the words into pairs with "known" differences and pairs with "unknown" differences. The first study asked what differences were common knowledge. For each non-synonym item pair, 10 participants from the Amazon Mechanical Turk online survey system were asked to write down all the differences they could think of without using outside sources. We verified or debunked all of the provided differences with external sources, and then created a four-item true-false test for each pair of words using both the correct and incorrect answers provided by participants.

In the second pilot experiment, we excluded everyone who had participated in the first pilot experiment and asked 10 new participants from the same population to take the true/false test. Any word pair where participants were more than 60% accurate overall and had no "true" items at chance (i.e., across all participants the accuracy on each of the true items was significantly greater than .5) was identified as a word pair with well-known differences ("Known" pairs), and all others were identified as word pairs without well-known differences ("Unknown" pairs). Synonyms were not tested because they were drawn from definitions in a widely used American dictionary. The breakdown of word pairs based on these results can be seen in Table 1.

Table 1
Stimuli used in Study 1

Word Pairs with Differences That Are WELL KNOWN ("KNOWN" Items)	Word Pairs with Differences That Are NOT WELL KNOWN ("UNKNOWN" Items)	Word Pairs That Are SYNONYMS
**butterfly–moth	asteroid–meteor	baby–infant
church–chapel	baking soda–baking powder	car–automobile
condominium–apartment	**blackbird–starling	dirt–soil
dog–wolf	coyote–jackal	expressway–freeway
**donkey–mule	**cucumber–zucchini	gasoline–petrol
fruit–vegetable	dinner–supper	grade school–elementary school
gecko–newt	**disease–syndrome	inoculation–vaccination
**jam–jelly	elm–beech	jewel–gem
nail–bolt	**ferret–weasel	redwood–sequoia
opossum–wombat	government resolution–government bill	sofa–couch
rabbit–hare	grasshopper–cricket	soda–pop
**rowboat–canoe	**pine–fir	student–pupil
**seal–walrus	porcupine–hedgehog	
shears–clippers	**shrew–mole	
silver–pewter		
tornado–hurricane		
town–village		
tweezers–tongs		
**wool–silk		

Note. Starred items were used in the list task.

3. Study 1

In the initial experiment, we sought to test whether the MM effect exists in adults. We took the direct approach of asking adults to estimate how many differences they could list for each of the 45 word pairs, then asked them to actually produce lists for a subset of the non-synonym pairs. There are three distinct measures one can examine from this procedure: (a) the initial estimates, (b) the number of differences provided in the list task, and (c) the difference between the initial estimates and the number of differences provided, which is a direct measure of the MM effect.

There are distinct predictions for each measure. We propose that adults have the common knowledge that two words mean different things but fail to recognize that they must defer to experts in the language community to access most of the distinctive features. Therefore, we predicted that there should be no difference in initial estimates for items that they knew had different meanings (Known and Unknown items). Yet both of those item types should be distinct from items that they know are not different or have very few differences (Synonym items). However, according to one alternative prediction, if adults have a relatively accurate sense of their own knowledge and are not deceived by their common knowledge, then we should see lower ratings for Unknown than Known items. Finally, if adults are blindly overconfident about knowing the differences between different words, we should see no distinction between the three item types.

For the provided differences, our "Known/Unknown" pilot explicitly predicts that adults should actually provide fewer differences in Unknown pairs than Known pairs. If we did not find this pattern, it would indicate that our pilot study was flawed, as it was supposed to be a direct measure of what adults should be able to provide.

For the difference between the estimates and the number of provided differences, hereafter referred to as the MM effect, our prediction is straightforward: We should see a consistent MM effect across items and individuals. We were primarily interested in the frequency of the MM effect rather than its magnitude—a significant difference in means between provided differences and initial estimates could indicate that a minority of individuals overestimated their knowledge, but to a large degree. Our hypothesis states that the broad population of speakers should mistakenly assume they possess knowledge of distinctive features of meaning when in fact they must defer to acquire them. Therefore, the strongest support of the theory is to demonstrate that overestimation is very common within the population across most items, not simply that a few people overestimate by some large margin. However, with regard to magnitude, we predicted that we would see a difference between Known and Unknown items. If our predictions for the initial estimates are correct, they should provide equally large estimates for Known and Unknown items. If our prediction for the provided differences is correct, they should provide fewer differences in Unknown items. Therefore, by failing to distinguish Known and Unknown items in their initial estimates but providing fewer differences in Unknown items, the magnitude of the MM effect should be greater for Unknown items.

3.1. Methods

3.1.1. Participants

Participants were adults (N = 36, 13 male, 19 female, four did not report) drawn from the local population and the university's Introductory Psychology Subject Pool. Participants received $10 or course credit for their participation.

3.1.2. Apparatus

For all participants, stimuli were presented and data were collected on an Apple MacBook™ laptop using the PsyScope stimulus presentation software (Cohen, MacWhinney, Flatt, & Provost, 1993). Participants responded on a USB keyboard attached to the laptop.

3.1.3. Materials and procedure

The study consisted of three tasks: an initial rating task, a distracter task, and a list task. In the initial rating task, participants were instructed to type in how many differences they thought they could list between pairs of words. They were informed that these differences had to be intrinsic to the meaning of the words and could not involve how the words were spelled, used pragmatically (e.g., "this word is more high-class than the other one"), or personal preferences. Examples of acceptable and unacceptable differences were provided for a pair of words that were not used in the actual study, "Cat-Dog." An example of acceptable difference was "Dogs bark and cats meow," and examples of unacceptable differences were "Cat starts with 'c' and Dog starts with 'd'" and "I personally prefer cats to dogs."

The words were presented in the center of the screen. Participants were told they had 8 s to report how many differences they thought they could list between each pair, and a countdown was displayed on the screen during the task. The time limit was used to prevent participants from composing a list of all the differences they knew internally before responding. After 8 s, the program automatically advanced to the next item. Participants responded using the number pad on a keyboard. If they failed to respond in time, the item recorded blank data, and if it was an item later used in the list task, that item was excluded from further analysis.

The distracter task was an unrelated task where participants had to rate the usefulness of various facts. This distracter had no words that were used in the rating task. The purpose of the distracter task was to reduce the influence of memory of the initial estimates on the subsequent list task.

In the list task, participants were instructed to make lists of all the differences they knew for a subset of the items from the rating task. They were instructed that the differences had to be real, about the meanings of the word, and could not involve spelling or subjective differences like personal preferences, mirroring the exact constraints of the rating task. The same examples of acceptable differences from the rating task were provided (see above). Twelve items were used, six from the "Known" category and six from the "Unknown" category. These pairs were selected based on two criteria, determined in piloting: First, the items did not have regional differences in meaning, as far as we were

able to determine. Second, the items had unambiguous, externally verifiable differences, to make coding tractable. Participants typed in their lists on the keyboard. Participants were told they had as long as they needed and were encouraged to list as many differences as they could think of.

3.2. Results

Six participants were excluded due to software failures. To reduce noise, we excluded participants who had average initial ratings greater than 30, far more than two standard deviations from the overall mean (M = 5.6, SD = 9.7). Only one participant was excluded based on this criterion, leaving a final N of 29.

The analyses cover three dependent measures: the initial estimates, the number of differences provided in the list task, and the difference between the provided differences and the ratings, or the MM effect.

3.2.1. Initial estimates

As predicted, Synonym items were distinguished from Known and Unknown items, but Known and Unknown items were not distinguished from each other. As Fig. 1 shows, participants gave significantly lower initial estimates for Synonym items (M = 1.810, SD = .665) than Known (M = 4.358, SD = 1.104) and Unknown (M = 3.681, SD = 1.003) items, repeated-measures ANOVA $F(2, 28) = 11.734$, $p < .001$, $\eta_p^2 = .442$; pairwise comparisons, ps < .01. However, pairwise comparisons showed they did not rate Known and Unknown items significantly differently, $p > .5$. This suggests that the availability of differences in Known items had no effect on initial estimates.

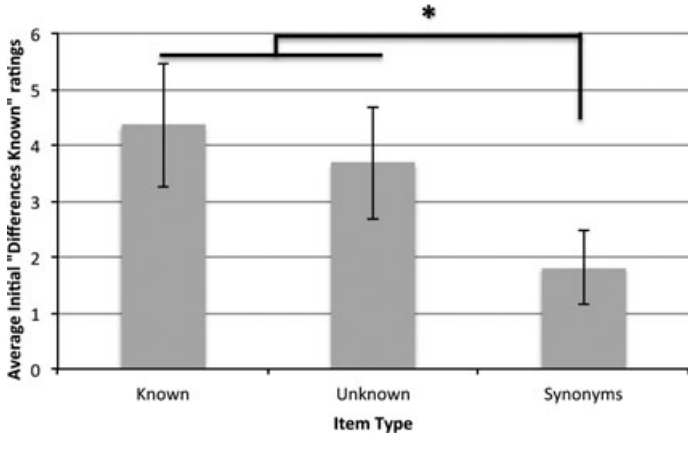

Fig. 1. Adults' initial estimates of differences they know between words, by item type, in Study 1. Error bars represent SEM. *Pairwise comparisons significant at $p < .05$.

3.2.2. Provided differences

To obtain an accurate measure of participants' knowledge, all provided differences were coded by one research assistant for accuracy, and then independently coded by a second research assistant to obtain interrater reliability. This coding ensured that participants could not simply fabricate items to lengthen their lists. Both coders were not blind to the hypotheses of the study, but they were blind to the initial ratings and therefore could not predict whether the coding of any given item would confirm or deny the hypotheses. Interrater reliability was analyzed with a Spearman Rank-Order Correlation across individual items and was good (r_s[383] = .884). The codes of the first coder were used for all analyses. Overall, 181 differences (28.5% of all provided) were coded as invalid across all 12 items and 29 participants, with a maximum of 31 excluded for any individual item (Cucumber–Zucchini). The exclusions were due to either factual inaccuracy, verified by external sources (e.g., "cucumber has seeds zucchini doesn't"), or failing to follow the directions regarding acceptable differences (e.g., "Jam can also refer to a sticky situation in which you are stuck.").

As we predicted, adults provided more differences in Known items (M = 1.856, SD = .866) than Unknown items (M = .656, SD = .761), $t(58) = 5.698$, $p < .001$. This validates the categorization from our pilot study for these 12 items.

3.2.3. The MM effect

Fig. 2 shows the proportion of participants that showed an MM effect on each item. As predicted, participants generally estimated that they would be able to list more

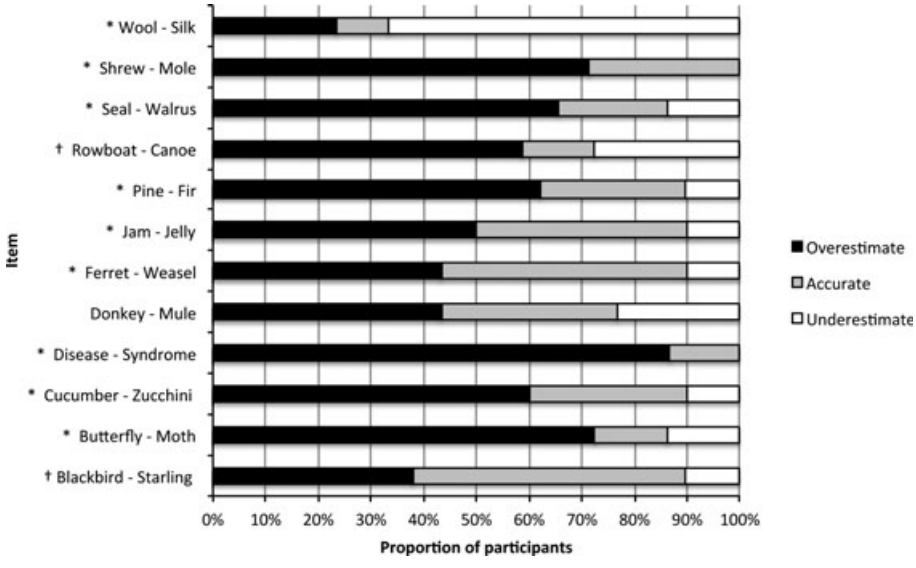

Fig. 2. The proportion of adults who overestimated their knowledge, underestimated it, and were accurate in Study 1. *significant sign test, †marginally significant sign test. Only one item was significant in favor of underestimating knowledge (Wool–Silk).

differences than they were actually able to (the "overestimate" bars in Fig. 2). The effect was analyzed in sign tests for each item, to better determine how common the MM effect was rather than how large it tended to be. Eight items were significant and in the expected negative direction (ps < .05), three Known items and five Unknown items. One item was significant in the opposite direction (Wool–Silk, $p < .05$). For the remaining three items, all were in the expected negative direction, but marginal (Blackbird–Starling, $p = .057$; Rowboat–Canoe, $p = .1$) or nonsignificant (Donkey–Mule, $p > .2$). While not completely uniform, it is still clear that in the majority of cases adults showed the MM effect.

Readers may notice that this effect could be explained as a by-product of the 181 provided differences that were excluded in coding. Participants' estimates may have reflected the number of things they felt they could list, not realizing that some of the knowledge they possessed was inaccurate. This would be a very different and much less interesting effect—it is no surprise that people are unaware that some of their knowledge is inaccurate (e.g., Fischhoff, Slovic, & Lichtenstein, 1977). Rather, according to the MM effect predicted here, people are actually misjudging the amount of knowledge they possess. To rule out this deflationary interpretation, a separate set of sign tests were conducted using *all* of the provided differences (regardless of accuracy or rule adherence). The MM effect was still significant for four items (ps < .05), marginally significant for three additional items (ps < .1), and only one item showed a significant effect in the opposite direction (Wool–Silk, $p = .007$). In short, while excluding inaccurate responses did make the MM effect more consistent, the effect exists even when accuracy is ignored. For all future analyses and studies, we elected to focus on accurate differences only, again to rule out the possibility that some participants might fabricate additional differences.

Turning to the magnitude of the MM effect rather than its frequency, there was no significant difference in the magnitude of the MM effect between Known ($M = -2.404$, $SD = 5.899$) and Unknown ($M = -3.057$, $SD = 5.649$) items, $p = .434$. This is unexpected, given that our predictions about the initial estimates and provided differences were correct. Furthermore, when examining the initial estimates for just the 12 items used in the list task, there is no significant effect of item type, $p > .5$, so the result cannot be explained by the estimates for these 12 items following a different pattern from the overall estimates.

3.3. Discussion

Adults commonly showed a clear MM effect for the majority of items tested. The differences that participants provided often revealed how few distinctive features they actually possessed. In one typical example, one participant estimated that he could name three differences between a cucumber and a zucchini. However, in the list task, he only had one thing on his list: "they are different kinds of vegetables."

In addition, adults clearly distinguished Synonym items from Known and Unknown items in their initial estimates, but gave equal estimates for Known and Unknown items. This suggests that availability did not impact the initial estimates but also that participants were not blindly overconfident, as they recognized that synonyms would have few or no differences between them. Participants also validated the Known–Unknown distinction by providing fewer differences in Unknown items overall.

There was no effect of item type on the magnitude of the MM effect. While it is difficult to interpret a null result, the MM effect is calculated on the basis of the individual difference between a single participant's estimate for a given item and the number of differences they are able to list for that item. As such, there is ample room for variation in the magnitude of the effect that is not captured by the independent averages of those two measures. This may indicate that, while on average participants do not distinguish between Known and Unknown pairs in their initial estimates, on an individual basis they may be well enough calibrated that the magnitude of the MM effect is no greater for Unknown items. However, that is not to say that they are well calibrated, only that the degree to which adults are overconfident is consistent across item types.

4. Study 2

In our second study, we investigated the effects found in Study 1 with children in grades K, 2, and 4. This study was motivated by a simple prediction, as outlined in the Introduction: The MM effect should be greater in magnitude and frequency for younger children. While this provides a prediction for the overall pattern of results, more specific predictions can be made about each measure.

There are two means of increasing the magnitude of the MM effect: Young children could give even greater estimates of the number of differences they know, or they could provide fewer differences in the list task. These are not mutually contradictory, and in fact we predict that we should see both. As noted in the Introduction, children are often overconfident about their knowledge, and at the same time, at the ages we are testing, children are still adding many words to their vocabulary. Therefore, we predicted that young children should give greater initial estimates and provide fewer differences than older children and adults, and as a result show a greater and potentially more frequent MM effect.

4.1. Methods

4.1.1. Participants

Kindergarteners (N = 41, 16 male, 25 female), second graders (N = 37, 13 male, 24 female), and fourth graders (N = 34, 18 male, 16 female) were recruited from elementary schools in southern Connecticut. Every effort was made to get a representative population from each school. Demographics roughly mirrored state population norms.

4.1.2. Apparatus

The same apparatus from Study 1 was used for Study 2, with a modified program that made the procedure more accessible to young children.

4.1.3. Materials

Twenty-two of the original 45 word pairs in the initial rating task in Study 1 were used in Study 2. The proportion of Known versus Unknown versus Synonym pairs was kept roughly the same, with seven Known items, nine Unknown items, and six Synonym items. The list of stimuli used with children can be seen in Table 2. This selection was based on expectations of children's exposure to these terms, based on literary material aimed at the ages tested.

4.1.4. Procedure

Participants were run in hallways or empty classrooms during school hours. Participants received a certificate of appreciation and a small toy on completion of the study.

The procedure was similar to that of Study 1 with a few key modifications. The experimenter read the instructions aloud to younger children, while older children read them to themselves. The same "acceptability" instructions were used, with the same examples. All children completed three practice items to get used to the task. The computer played a recording of the experimenter reading each pair of words. These recordings were standardized at 2 s, and the 8-s countdown started at the end of the recording. Kindergarteners saw the countdown but the program did not automatically advance if it reached 0. Older child participants responded using the number pad on a keyboard, while kindergarteners dictated their ratings to the experimenter, as in piloting the act of entering numbers on the keyboard proved too distracting for the youngest age group.

With children, the distracter task was to identify the actions being performed in a set of rapidly presented photographs. The photographs did not have any objects related to words in the rating task. On the list task, 6 of the original 12 items were used, again half Known items and half Unknown items. All children dictated their lists to the experimenter.

Table 2
Stimuli used in study 2, divided by item type

Word Pairs with Differences That Are WELL KNOWN ("KNOWN" items)	Word Pairs with Differences That Are NOT WELL KNOWN ("UNKNOWN" Items)	Word Pairs That Are SYNONYMS
**butterfly–moth	coyote–jackal	baby–infant
dog–wolf	**cucumber–zucchini	car–automobile
donkey–mule	dinner–supper	dirt–soil
gecko–newt	**ferret–weasel	jewel–gem
**rowboat–canoe	grasshopper–cricket	soda–pop
rabbit–hare	**pine–fir	sofa–couch
**seal–walrus	porcupine–hedgehog	
tornado–hurricane		
tweezers–tongs		

Note. Starred items were used in the list task.

4.2. Results

To compare the performance of children to that of adults, we conducted the following analyses using data from Study 1, but only for the items used in Study 2. The same average initial estimate exclusion criteria were used, which removed the data of one kindergartener, three second graders, and two fourth graders. These analyses therefore include data from 29 adults (Study 1), 30 fourth graders, 29 second graders, and 28 kindergarteners.

4.2.1. Initial estimates

Fig. 3 shows the initial estimates for each age group and item type. There were significant effects of grade and item type, as well as an interaction. All pairwise comparison *p*-values are Bonferroni corrected unless otherwise noted. As predicted, kindergarteners provided higher initial estimates than any other age group, repeated-measures ANOVA, $F(3, 116) = 5.376$, $p < .01$, $\eta_p^2 = .122$; pairwise comparisons $ps \leq .01$. Overall, participants gave lower estimates for Synonyms than other item types, $F(2, 115) = 6.913$, $p = .001$, $\eta_p^2 = .056$; pairwise comparisons $ps \leq .004$, but there was an interaction between grade and item type, $F(6,232) = 4.696$, $p < .001$, $\eta_p^2 = .108$. Further analysis revealed that only fourth graders and adults showed a significant effect of item type, 4G: $F(2,29) = 4.070$, $p = .028$, $\eta_p^2 = .218$; Adult: $F(2, 28) = 11.894$, $p < .001$, $\eta_p^2 = .459$. Fourth graders gave significantly lower estimates for Synonym items ($M = 2.343$,

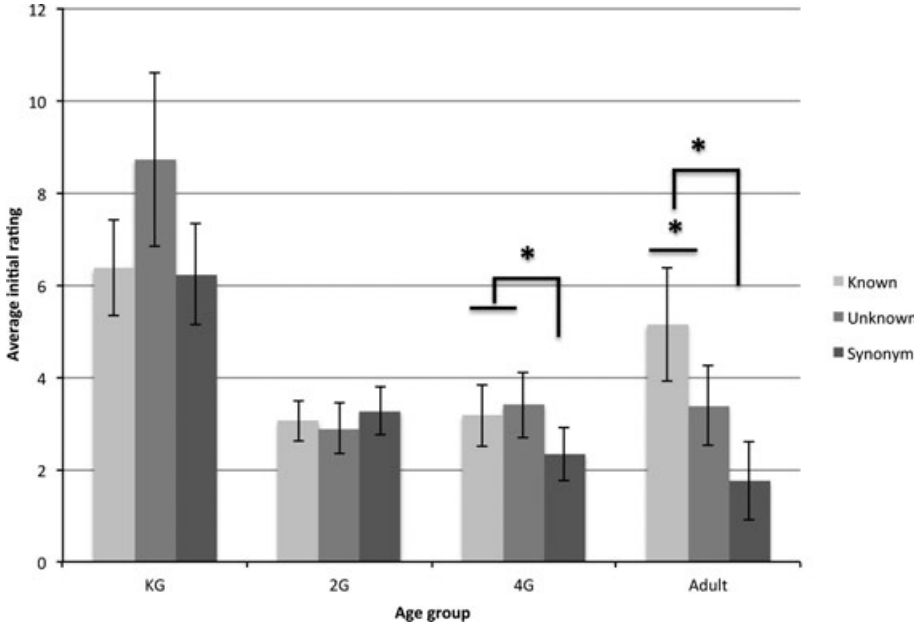

Fig. 3. Initial estimates of differences known by item type and age group in Study 2. Error bars represent SEM. *Pairwise comparisons significant at $p < .05$.

SD = 3.231) than Known (*M* = 3.176, *SD* = 3.716) or Unknown items (*M* = 3.411, *SD* = 3.979) (pairwise comparisons, *p*s < .05). With these 22 items, adults actually showed differences between all three item types, with Synonyms (*M* = 1.751, *SD* = 4.65) receiving lower estimates than Known (*M* = 5.159, *SD* = 6.776) or Unknown items (*M* = 3.388, *SD* = 4.771), and Unknown items receiving lower ratings that Known items (pairwise comparisons, *p*s ≤ .03). The difference between Known and Unknown items in adults is somewhat surprising, given that this effect was not significant in Study 1. However, this simply means that the effect only reaches significance in this smaller set of 22 items, and drowned out in the larger set.

4.2.2. Provided differences

Responses were coded by the same coder as Study 1 using the same coding guidelines. Overall, 137 KG responses (74%), 144 2G responses (61%), and 139 4G responses (52%) were coded as invalid. For adult responses from Study 1, 103 (29%) were coded as invalid across these six items.

For provided differences, there were main effects of grade and item type, as well as an interaction. As we predicted, kindergarteners (*M* = .230, *SD* = .215) provided significantly fewer differences than fourth graders (*M* = .640, *SD* = .368) and adults (*M* = 1.389, *SD* = .876), and adults provided more than all other grades, but there were no significant differences between second (*M* = .483, *SD* = .356) and fourth grade, and only marginally significant differences between kindergarten and second grade, repeated-measures ANOVA, $F(3, 116) = 27.376$, $p < .001$, $\eta_p^2 = .415$; pairwise comparisons, all significant differences *p*s < .01.

Further analyses were conducted to examine the interaction effect. Separate one-way ANOVAS were conducted to examine the effect of age group on Known and Unknown items. Both item types showed the significant effect of age, Known: $F(3, 116) = 25.889$, $p < .001$; Unknown: $F(3, 116) = 16.242$, $p < .001$. Post hoc tests of Known items showed that kindergarteners (*M* = .356, *SD* = .344) provided fewer differences than fourth graders (*M* = 1.054, *SD* = .603) and adults (*M* = 1.922, *SD* = 1.15), but not second graders (*M* = .711, *SD* = .493) (pairwise comparisons, *p*s < .01). Furthermore, second graders and fourth graders provided significantly fewer differences than adults (*p*s < .05). Second graders did not differ significantly from kindergarteners or fourth graders. For Unknown items, the same analyses revealed a very different picture. Kindergarteners (*M* = .103, *SD* = .157), second graders (*M* = .255, *SD* = .335), and fourth graders (*M* = .225, *SD* = .263) did not differ significantly from each other, but all three child groups differed significantly from adults (*M* = .855, *SD* = .791), pairwise comparisons, all *p*s < .001. This pattern of results can be seen in Fig. 4.

4.2.3. The MM effect

The sign test was significant for all six items in every grade at $p < .025$. Because the majority of children's provided differences were excluded, the possibility arises that the large number of excluded differences contributed to this result. We therefore conduced further sign tests with every provided response and found that the MM effect was still

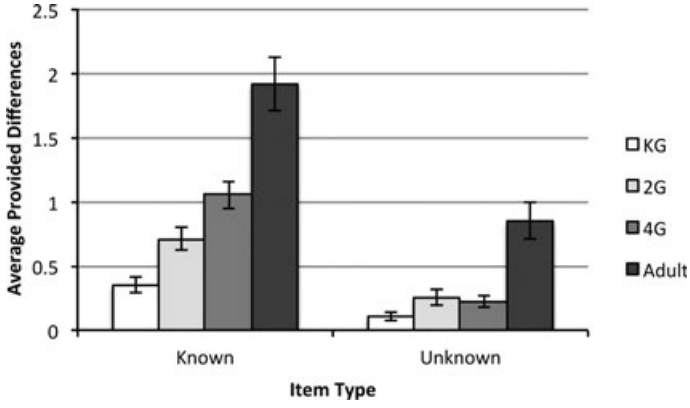

Fig. 4. Provided differences by item type and age group in Study 2. Error bars represent SEM.

present. For kindergarteners, every item still showed a significant MM effect. For second graders, all but one item showed a significant effect (Butterfly–Moth). Fourth graders showed no significant effects for three items, but the other three still showed a significant MM effect. For these six items with adults, two showed a marginal MM effect ($p < .1$) and one became nonsignificant, but the other three showed a significant MM effect. Even including every provided difference that was completely inaccurate or otherwise not in line with the stated rules, the MM effect persisted for, at a minimum, three out of six items for older participants and most or every item for younger participants. Going forward, we will focus on the coded data, with the invalid differences excluded.

Fig. 5 shows the magnitude MM effect for all age groups. There was a significant effect of grade on the magnitude of the MM effect, and as predicted kindergarteners ($M = -6.885$, $SD = 6.95$) showed a significantly greater MM effect than all three older groups (2G: $M = -2.278$, $SD = 1.72$; 4G: $M = -2.972$, $SD = 1.665$; Adult: $M = -2.973$, $SD = 1.656$), $F(3, 116) = 5.098$, $p = .002$, $\eta_p^2 = .116$; pairwise comparisons, all ps $< .05$. There was no effect of item type and no interaction.

To test for an effect of grade on the frequency of the MM effect, responses were recoded on the basis of whether or not they showed the MM effect (either 1 or 0). We then conducted the same Grade × Item Type analysis using the frequency of the MM effect (averaged by item type) rather than its magnitude. As predicted, there was a significant main effect of grade, repeated-measures ANOVA, $F(3, 116) = 3.892$, $p = .011$, $\eta_p^2 = .091$. Kindergarteners ($M = .851$, $SD = .285$) showed the MM effect more frequently than adults ($M = .600$, $SD = .284$), pairwise comparisons, $p = .006$. There were no significant effects for second graders ($M = .706$, $SD = .284$) or fourth graders ($M = .704$, $SD = .284$).

There was no main effect of item type ($p > .5$), but there was an unexpected interaction between grade and item type, $F(3, 116) = 4.253$, $p = .007$, $\eta_p^2 = .099$. We analyzed item type separately in each grade, and found that there was only a significant effect of item type for fourth graders, who showed the MM effect more often for Unknown ($M = .785$, $SD = .305$) than Known items ($M = .624$, $SD = .307$), paired-sample t-test,

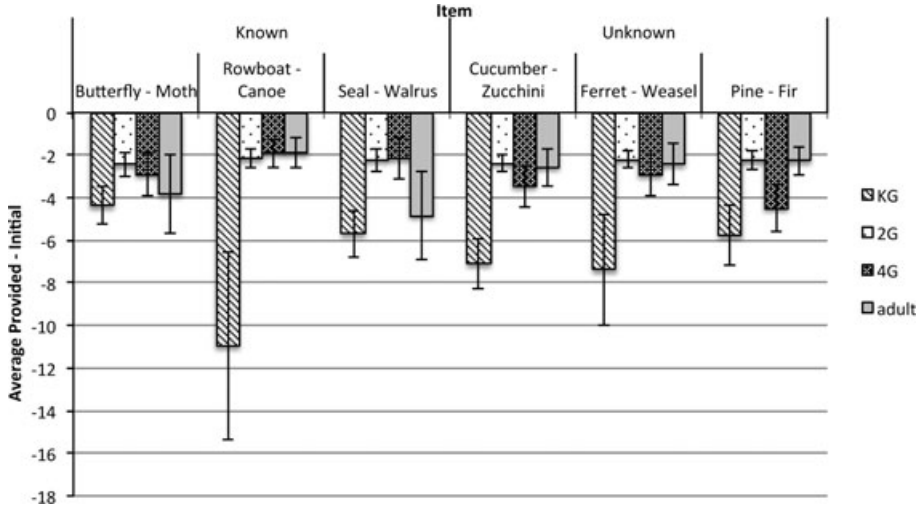

Fig. 5. The magnitude of the MM Effect for the six items used with all age groups in Study 2, by age group. Note that kindergarteners consistently showed a greater MM effect than other groups. Error bars represent SEM.

$t(30) = -2.54$, $p = .016$. While unexpected, this result does not have any bearing on the key questions of interest. However, future studies of the MM effect should attempt to replicate this finding and determine if in fact fourth graders uniquely distinguish between Known and Unknown items in the MM effect.

5. Discussion of Studies 1 and 2

In our first two studies, we found that adults and children in kindergarten, second, and fourth grade all show a frequent MM effect. Furthermore, young children (kindergarteners) showed a stronger and more frequent MM effect than older children and adults. This stronger MM effect reflects both greater initial estimates of the number of items they could list and being able to actually list even fewer items.

The age effects primarily show a difference between kindergarteners and older participants, which mirrors previous developmental patterns found with the IOED (Mills & Keil, 2004). However, one notable effect with older child participants is that neither kindergarteners nor second graders distinguished Synonym items from Known or Unknown items in their initial estimates. This may indicate that they are overapplying the assumption that distinct words have distinct referents (Clark, 1983; Markman & Wachtel, 1988; Mervis & Bertrand, 1994). Younger children might be so strongly biased to assume that novel words have different meanings from other words that they will immediately conclude not only that the words must be different but, based on that common knowledge, conclude that they must know some of the distinctive features as well. Indeed, previous

work has found that the strength of this bias seems to diminish with age, though it can still be found in adults (Markman, 1991).

These first two studies clearly support the presence of the MM effect and its greater strength in children. In Studies 3 and 4, we investigated the potential mechanisms and boundaries of the MM effect in adults. One account of the MM effect is that people confuse all the knowledge that they believe *exists* with the knowledge they actually possess. In other words, people are overconfident because they are completely unaware of the division of linguistic labor. Study 3 tested this possibility and further examined the role of perceived available expert knowledge on the MM effect.

6. Study 3

Study 3 aimed to establish whether adults were explicitly aware of the division of linguistic labor. This awareness is important to our interpretation of the MM effect. The MM effect may exist because people mistake some portion of the information they can access from expert sources for information they already posses. Thus, despite their overconfidence in their own knowledge, they should still expect that there exist expert sources of knowledge that they can access. An alternative interpretation of the MM effect would be that participants believe they possess all of the knowledge about a word's meaning or believe they have access to the same information that experts do (i.e., that experts possess very little knowledge as well, or no more knowledge than the participants believe they can access from direct observation), without acknowledging any division of linguistic labor. Our methods are a straightforward test of this prediction: We repeated the initial estimation task from Study 1, but participants were asked either how many differences they personally knew (as in Study 1), or how many differences they thought existed that an expert would know. No list task was needed to answer the question at hand, and therefore we did not include a list task in this paradigm.

If adults are cognizant of the division of linguistic labor, then the participants we ask to estimate the number of differences that exist that an expert would know should give much higher estimates than those asked to rate how many differences they personally know. However, this difference should be true primarily for Known and Unknown items. For Synonym items, as there should be few or no differences that exist in the first place, we predicted that adults would not expect experts to know more differences than themselves. The obvious alternative is that there would be no differences between adults' estimations of their own knowledge and experts' knowledge. This would suggest that the MM effect is actually driven by a complete lack of awareness of the division of linguistic labor, which would be extremely surprising.

Awareness of expert sources may also have a more subtle effect. If knowing that a group of experts has more details about the relevant word meanings than oneself causes the MM effect, the magnitude of that expert/novice difference might influence the magnitude of the MM effect. There are two possibilities here. One intuitive prediction is that knowing that experts know a great deal more than oneself might cause one to be more

conservative about one's own knowledge, thereby reducing the MM effect. Alternatively, according to our account of the causes of the MM effect, believing that experts know a great deal about a contrast might cause novices to assume they know more as well, if they confuse the available knowledge of external sources with their own internal representations. In other words, if the MM effect is the result of confusing some portion of available (but not possessed) knowledge for possessed knowledge, then the greater the gap between expected possessed knowledge and expected available knowledge (operationalized here as the between-subjects difference between estimated self differences known and estimated expert differences known), the greater the MM effect should be. One can visualize this as a kind of pressure equilibrium of meaning features—the greater the disparity between experts and novices, the more some of the expert features mistakenly "leak" into one's own inferred knowledge.

6.1. Methods

6.1.1. Participants

Study 3 was conducted online using Amazon Mechanical Turk. Participants were 44 anonymous "workers" from the Mechanical Turk worker pool, all over the age of 18. All participants were paid $0.75 for a roughly 5–7 min study, a rate comparable with similar tasks on Mechanical Turk.

6.1.2. Materials and procedure

The rating task in Study 1 was adapted to an online format using the Qualtrics online survey system, which was then embedded in a frame in the Mechanical Turk interface. Participants were randomly divided into two groups that received different instructions. Group A was given the exact instructions from Study 1. Group B was asked to enter the number of differences that existed that an expert would know. Everything else was identical between groups and to the rating task of Study 1. The distracter and list tasks from Study 1 were omitted.

6.2. Results

One participant was excluded due to failing to respond to more than half of the items. In the end, there were 21 participants in group A and 22 participants in group B. As Fig. 6 shows, participants were aware that experts should know more than themselves, with group B ($M = 5.07$, $SD = 2.77$) giving higher estimates than group A ($M = 2.06$, $SD = 2.77$), $F(1, 41) = 12.664$, $p = .001$, $\eta_p^2 = .236$. In addition, as in Study 1, participants gave lower estimates overall for Synonyms ($M = 1.98$, $SD = 2.35$) than Known ($M = 4.64$, $SD = 3.64$) or Unknown ($M = 4.16$, $SD = 4.19$) items, $F(2, 40) = 25.915$, $p < .001$, $\eta_p^2 = .564$; pairwise comparisons, $ps < .001$.

There was no interaction between group and item type ($p > .1$). Participants who were asked to estimate expert knowledge gave higher ratings for all three item types (*t*-tests, $ps < .05$), even while recognizing that the Synonym items were different from both

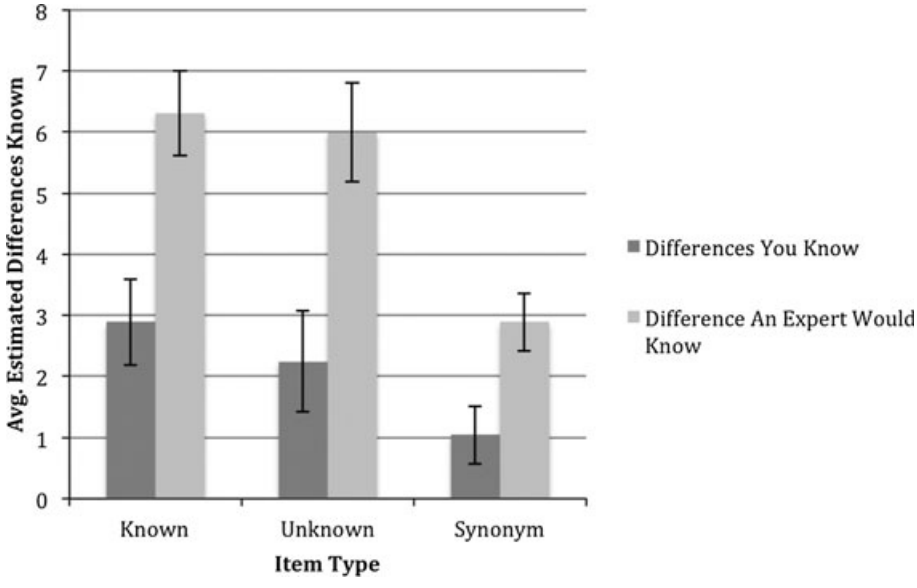

Fig. 6. Estimated differences known by group and item type in Study 3. Error bars represent SEM.

Known and Unknown items. This was unexpected, but potentially explained by the simple fact that estimates for Synonym items were not at floor. Adults apparently believe that the word pairs we classified as Synonyms have *fewer* differences than Known or Unknown pairs, but because they believe some differences exist, they still expect that an expert would know more than they themselves.

To test our hypothesis that a greater gap between perceived expert knowledge and perceived self knowledge would lead to a greater MM effect, we calculated the average difference in estimates between group A (self knowledge) and group B (expert knowledge) for the 12 items used in the list task of Study 1. We then conducted a linear regression of these averages with the average magnitude of the MM effect for the same 12 items. Because the group A–group B difference is between-subjects, we could not calculate these average differences scores on a subject-by-subject basis, so the regression ultimately only included the 12 averages from this study and the 12 average magnitudes from Study 1. Despite the low power of this regression, there was a significant relationship in the predicted direction. The greater the gap between estimated self knowledge and estimated expert knowledge for a given item, the greater the average magnitude of the MM effect for that item in Study 1, $F(1, 10) = 5.60$, $p = .04$, adjusted $R^2 = .295$. This pattern can be seen in Fig. 7.

6.3. Discussion

Study 3 showed that adults seem to be aware of the division of linguistic labor, despite their overconfidence in estimates of their own knowledge. Furthermore, we found

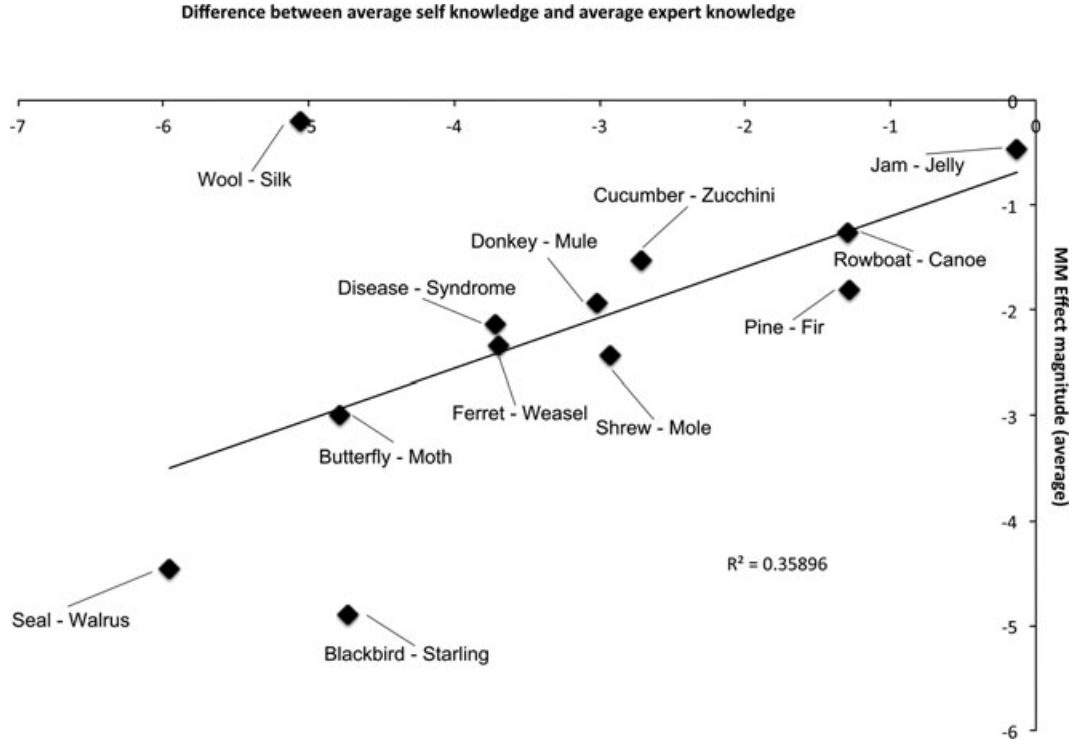

Fig. 7. Relationship between perceived self versus expert knowledge (Study 3) and the MM effect (Study 1).

evidence that the magnitude of the MM effect in adults is partially determined by the amount of expected available knowledge for a given distinction. This provides strong support for our account of the MM effect that it springs from mistaking some portion of available knowledge for possessed knowledge. The greater the available (but not possessed) knowledge, the greater the MM effect.

These findings provide some insight into the inner workings of the MM effect, but alone they do not rule out some alternative accounts. Given that participants seem aware of the division of linguistic labor but still overestimate their own knowledge, one simple account of the MM effect is that people are generally overconfident about metalinguistic knowledge. Indeed, there is a long history of prior work that has found that adults can be overconfident about many kinds of knowledge (e.g., Fischhoff et al., 1977). Study 4 tested this account with a task designed to only access common aspects of word meaning, which we propose would not require deference and therefore not generate an MM effect.

7. Study 4

At this point we have demonstrated the existence of the MM effect in children and adults and established that it coexists with explicit knowledge of the division of linguistic

labor at least in adults. However, an alternate interpretation of this pattern is that people are blindly overconfident about their metalinguistic knowledge, and it has nothing to do with the division of linguistic labor. This interpretation makes a clear prediction: We should see the same overconfidence effect in any metalinguistic task. In Study 4, we tested whether this is the case and further examined the role of common versus distinctive aspects of word meaning.

We have argued that the MM effect is related to the contrast between common and distinctive aspects of meaning. We have defined "common" aspects of meaning as superordinate category knowledge and non-specific metalinguistic comparisons, and "distinctive" aspects as fine-grained detailed knowledge that distinguishes one word from any other. We have suggested that the MM effect emerges from people possessing common aspects of meaning, but believing that they also possess the distinctive aspects. This makes a quite focused prediction: In a task where the common aspects of meaning are sufficient, there should be no overconfidence effect.

Based on our definitions of common aspects of meaning, one such task is simply to ask how many facts participants know about one of the words in the pairs we used in Study 1. While knowing the differences between two words obviously requires distinctive knowledge, knowing information about one word should only require common knowledge, such as that weasels are mammals. Study 4 therefore allows us to test two parts of our account. First, participants should be relatively well calibrated in their knowledge, and as such show that they do possess some common aspects of word meaning, and second that the deficits found in Study 1 are specific to distinctive aspects of meaning.

7.1. Methods

7.1.1. Participants

Study 4 was conducted using Amazon Mechanical Turk. Participants were 20 anonymous "workers" from the Mechanical Turk worker pool that had not participated in previous studies, all over the age of 18. Participants were paid $3.50 for a 20–30 min task, a rate comparable to similar tasks on Mechanical Turk.

7.1.2. Materials and procedure

One word from each pair used in Study 1 was selected for Study 4, except for the pairs with phrases instead of single words (e.g., "government resolution–government bill"), for a total of 41 items. As in Study 1, there was a rating task, a distracter task, and a list task. For the rating tasks, participants were instructed to estimate how many *facts* they could list about a given word, with the same rules as the "differences" task in Study 1, with examples adapted for single words rather than pairs of words and the same 8-second time limit.

The distracter task consisted of a standard mental rotation task. Participants were asked to indicate whether two three-dimensional shapes composed of joined cubes were two views of the same object or different objects, and then rate how confident they were in their answer on a 1–7 scale. There were 15 distracter items.

For the list task, participants were asked to list all the facts they could think of for one word from each of the 12 pairs in Study 1, using the same constraints as the list task of Study 1. The words chosen from each pair in this task were also used in the initial rating task. As in Study 1, participants were told to take as long as they needed and list as many words as possible, but were specifically requested not to refer to any outside sources (books or websites).

7.2. Results and discussion

7.2.1. Coding

Two new coders blind to both the hypotheses and the initial ratings coded the list task of Study 4. Coders were instructed to decide whether each fact was valid using the same criteria as participants and the coders of Study 1. Interrater reliability was very high (Spearman rank-order correlation, r_s = .851). Disagreements were settled by discussion to generate the final coding used in these analyses. In the final coding, 59 facts (9.3% of all provided) were omitted for inaccuracy or failing to follow the guidelines.

7.2.2. Results

The frequency of overconfidence among participants in Study 4 can be found in Fig. 8. Sign tests for every item were nonsignificant ($ps \geq .146$), indicating that, on the whole, participants were very well calibrated in this task.

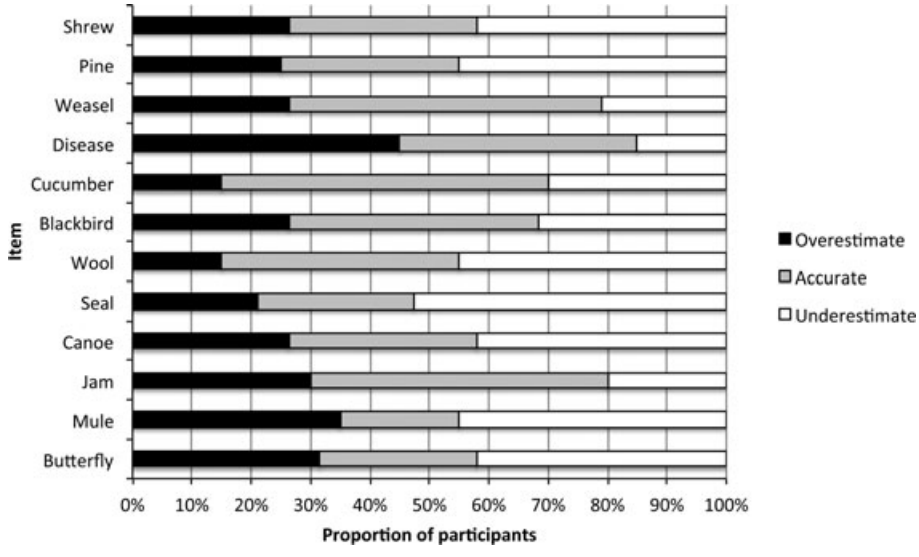

Fig. 8. The proportion of adults who overestimated their knowledge, underestimated it, and were accurate in Study 4. Note that participants were much more well calibrated to their knowledge compared to Study 1. *significant sign test, †marginally significant sign test.

This finding fits well with our account that participants should have access to common aspects of meaning but mistakenly believe they have access to distinctive aspects of meaning as well. If participants are basing their estimates on the common aspects of meaning to which they have access, then their initial estimates should be similar to those provided in Study 1. This result would also rule out an uninteresting explanation of why participants were better calibrated in this task—they could have provided lower ratings, making it easier for them to provide sufficient differences. On average, the initial estimates of participants in Study 1 (M = 4.00, SD = 5.22) did not differ significantly from the initial estimates of participants in Study 4 (M = 2.43, SD = 1.14), $t(48) = 1.315$, $p = .195$. This suggests that participants have relatively accurate access to common knowledge of word meanings but mistakenly believe that they have access to distinctive aspects of word meaning as well.

Far fewer items were excluded in Study 4 (9.3%) compared to Study 1 (28.5%). This cannot explain the calibration in and of itself, as even when every difference in Study 1 is examined, the MM effect is still present (see 3.2.3.). However, it does suggest that participants were more easily able to provide accurate information in this task, perhaps because common aspects of meaning were sufficient for providing a fact about a word, but not so for distinguishing between two words. For example, in Study 4 several participants provided some variation on "Shrews are mammals." This very abstract information about the meaning of the word "shrew" is sufficient to be a fact about shrews (and many other animals), but it would not distinguish between a shrew and a mole (as was required in Study 1).

8. General discussion

In four experiments, we have demonstrated the existence of the MM effect in adults (Study 1) and children in kindergarten, second, and fourth grades (Study 2). We found that children in kindergarten show a greater and more frequent MM effect, as they make even greater estimations of their own knowledge while actually knowing even less (Study 2). Furthermore, as we predicted, adults are aware of the availability of outside knowledge (Study 3), which may be a driving force behind the effect, a finding further supported by a positive relationship between the expert/self difference and the size of the MM effect. We have demonstrated that the MM effect is not due to broad metalinguistic overconfidence, as it occurs for distinctive aspects of word meaning, but not common aspects (Study 4).

8.1. Developmental importance of the MM effect

Although the MM effect is highly consistent across development, it is stronger in young children (Study 2). Kindergarteners both thought they knew more differences and were actually able to provide fewer differences than any other grade. The latter is to be expected. Children learn more differences between various items in the real world as they

grow older. More interesting is their larger overestimation of their personal knowledge. However, with the studies reported here we can draw no firm conclusions about how the MM effect changes between childhood and adulthood, and what specific mechanisms generate these changes. Further work is required to determine whether the MM effect occurs by the same mechanisms in children as it does in adults. In addition, more work is required to explore the sources of developmental change between kindergarten and adulthood in the magnitude of the MM effect.

However, there are some indications from other literatures that suggest that the MM effect may be adaptive for young children. Using mechanisms like the principle of mutual exclusivity (Markman, 1988), children are adding new words to their vocabulary every day, and in one sense they do "know" these words. They know of experts, or believe in the existence of experts, who know all the distinctive aspects of meaning between that word and others that they already possessed. This awareness then leads to the MM effect as it does in adults, in that they mistake the availability of that knowledge for its possession. However, if they realized that they did not actually possess that knowledge, they might not use those words, realizing that they do not understand how they contrast with other words. Because they are not aware of the gaps in their knowledge they are able to continue acquiring new words at a rapid pace, without getting "stuck" trying to learn all the nuances of a single word's meaning and the particular features that make it unique. This could lead to a stronger MM effect in younger children, who are especially under the sway of mutual exclusivity and related contrast principles and may need stronger support against being discouraged by their own ignorance.

Younger children are not completely ignorant about differences in meanings. They may not know any specific differences, but it is very likely that they do know what kinds of features or properties would count as specific differences and would be able to identify them as such. For example, based on earlier work on differences in how children think of the central features of artifacts and natural kinds (Brandone & Gelman, 2009; Keil, 1989, 2010), even preschoolers would be expected to know that intrinsic microstructural properties might be especially relevant to meaning contrasts for animals whereas functional nuances and features related to intentions of creators would matter more for artifact differences. A strong knowledge of probable kinds of differences may also be another source of the MM effect. This is a real form of common knowledge and it helps enable access to the right kinds of experts, and a sense of knowing associated with this form of knowledge may be confused with personally knowing distinctive details that differentiate some meanings. That knowledge, however, carries no information in itself about the difference between two closely related word meanings.

A striking developmental finding is younger children's beliefs that they know a large number of contrastive meanings for synonyms. We suspect that this belief may stem form an overzealous application of the mutual exclusivity principle or related contrast principles, combined with the adaptive value of assuming that one knows the meaning of a word well enough to use them in discourse. A related reason may have to do with early difficulty understanding that there are true tautologies and circularities (Baum, Danovitch,

& Keil, 2008; Osherson & Markman, 1975), as synonyms are tautologically equivalent in meaning. This phenomenon warrants further study beyond the MM effect.

8.2. Divisions of linguistic and cognitive labor

Our studies provide strong evidence for the cognitive underpinnings of Putnam (1975)'s division of linguistic labor. Study 3 in particular showed that adults are aware that they do not possess *all* of the criteria that differentiate two words, while the MM effect itself demonstrates that "meaning ain't in the head." Study 4 added an interesting twist to this story, by providing evidence that adults possess at least common aspects of meaning, and it is specifically more fine-grained distinctive components of meaning that they must defer to acquire. A further test of the division of linguistic labor would be to show that when participants are made aware of the fact that they do not possess those differences, they seek information from experts that they believe do. The act of deference would directly demonstrate the function of the division of linguistic labor in the real world, the ability to use terms with confidence because of the availability of information about them.

Deference is common in the world, and there is some suggestion that it is growing even more common as we become more reliant on tools such as the Internet as sources of information. Recent research also suggests that when people expect to have access to information, they will tend not to remember the information itself, but will remember where to access it (Sparrow, Liu, & Wegner, 2011). However, that study did not investigate whether people believed that they possessed that knowledge, nor did it focus on verbal knowledge. A future study might take these findings and ask whether the same pattern holds for knowledge of word meanings. If people expect to have access to outside knowledge about these words in the future, will they show an inflated MM-like effect, either because they overestimate their knowledge even more or retain even less (or both)?

A second question concerns whether people overestimate the availability of information. Both this work and the IOED literature have suggested that people overestimate their knowledge because they are aware of its availability from outside sources, but is that awareness itself accurate? Is there a secondary illusion where information they think they can access is in fact often out of their reach, perhaps because it is not known by anyone or is too hard to access? This would be even more problematic, as it would imply that people both think they have knowledge and access to knowledge that they simply do not. One can create cases where this would be true; the critical question is how often it occurs in more naturalistic contexts.

One final point to consider is that "outside sources" may encompass more than just experts. The distinctive details of a word meaning may be accessible in the mind in the expert, but they are sometimes accessible from the thing itself. Indeed, if participants in Study 1 were shown images of a ferret and a weasel during the list task, it is entirely plausible that they would have been able to list more differences than they did, and the MM effect would be smaller or non-existent.

In placing emphasis on the role of deference in the MM effect, we have spoken mostly about deference to experts, but deference to extra-linguistic context information could be included as well without substantive alteration to the account. People assume that they possess distinctive details of meaning, when in fact those distinctive details are only accessible from an outside source, "misplaced" in the broader context and the minds of experts. We focus on deference to experts specifically as the distinctive details that can be gleaned from direct observation are at best a subset of those available from an expert source. One would expect that an expert source would have access to all the details of meaning accessible by direct observation, but perhaps also other details not accessible by those means (for example, details of biology).

8.3. Conclusion

In short, meanings for some words "ain't in the head," even though we often think they are throughout development. Moreover, this may be an adaptive illusion that reflects real success in reference and gives us all the confidence to use terms that are frequently based on quite minimal understandings of contrastive details with other terms. Because the effect is rooted in deference, it is an illusion of misplaced meanings rather than of missing meanings.

Acknowledgments

This research was supported by National Institutes of Health Grant R-37-HD023922 to Frank C. Keil and THRIVE Center grant "Cognitive and Developmental Facets of Intellectual Humility" to Frank C. Keil and Kristi L. Lockhart. We thank the public school districts of Weston, Derby, Brookfield, Bethel, Montville, New London, and Meriden, CT, and Stepping Stones Children's Museum, Norwalk, CT, for their assistance with Study 2. The authors also wish to thank Kara Gaughen, Luke Berszakiewicz, Michael Pacer, Madhawee Fernando, Kate Doyle, Sarah Tornetta, and Alexander LaTourette for their assistance with the project.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436–451. doi:10.1037/a0020218.

Baum, L. A., Danovitch, J. H., & Keil, F. C. (2008). Children's sensitivity to circular explanations. *Journal of Experimental Child Psychology*, *100*(2), 146–155. doi:10.1016/j.jecp.2007.10.007.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Brandone, A. C., & Gelman, S. A. (2009). Differences in preschoolers' and adults' use of generics about novel animals and artifacts: A window onto a conceptual divide. *Cognition*, *110*(1), 1–22. doi:10.1016/j.cognition.2008.08.005.

Clark, E. V. (1983). Convention and contrast in acquiring the lexicon. *Concept Development and the Development of Word Meaning*, *12*, 67.

Clark, E. V. (1987). *The principle of contrast: A constraint on language acquisition. Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using macintosh computers. *Behavior Research Methods*, *25*(2), 257–271. doi:10.3758/BF03204507.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*, 1115–1131.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552. doi:10.1037/0096-1523.3.4.552.

Fisher, M., & Keil, F. C. (2014). The illusion of argument justification. *Journal of Experimental Psychology: General*, *143*, 425–433.

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology*, *1*(4), 324–340.

Gatewood, J. B. (1983). Loose talk: Linguistic competence and recognition ability. *American Anthropologist*, *85* (2), 378–387. Available at: http://www.jstor.org/stable/67632. Accessed October 9, 2012.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341.

Keil, F. C. (1989). *Concepts, kinds, and conceptual development*. Cambridge, MA: MIT Press.

Keil, F. C. (2010). The feasibility of folk science. *Cognitive Science*, *34*(5), 826–862. doi:10.1111/j.1551-6709.2010.01108.x.

Keil, F. C., Lockhart, K. L., & Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *Journal of Experimental Psychology: General*, *139*(1), 1–15. doi:10.1037/a0018319.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, *32* (2), 259–300. doi:10.1080/03640210701863339.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261–1277. doi:10.1111/j.1467-8624.2005.00849.x.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073–1084.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. *Child Development*, *65*(6), 1646–1662.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*(1), 1–32.

Osherson, D. N., & Markman, E. (1975). Language and the ability to evaluate contradictions and tautologies* 1. *Cognition*, *3*(3), 213–226.

Putnam, H. (1975) *The meaning of "meaning". In Philosophical Papers, Vol. 2: Mind, Language and Reality*. Cambridge, UK: Cambridge University Press.

Roberts, K. P., & Blades, M. (2000). *Children's source monitoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. doi:10.1207/s15516709cog2605_1.

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019–1050. doi:10.1111/j.1551-6709.2012.01250.x.

Smith, A. (1776). *The wealth of nations*. London: W. Strahan and T. Cadell.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776.

Taylor, M., Esbensen, B. M., & Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child Development*, *65*(6), 1581–1604. doi:10.1111/j.1467-8624.1994.tb00837.x.