Weight matters: The role of physical weight in non-physical language across age and culture

Tomer D. Ullman (tomeru@mit.edu)
Brain and Cognitive Sciences, MIT

Santiago Alonso-Diaz (salonsod@ur.rochester.edu)
Brain and Cognitive Sciences, University of Rochester

Stephen Ferrigno (sferrigno@rcbi.rochester.edu)
Brain and Cognitive Sciences, University of Rochester

Sarina Zahid (szahid@u.rochester.edu)
Brain and Cognitive Sciences, University of Rochester

Celeste Kidd (celestekidd@gmail.com)
Brain and Cognitive Sciences, University of Rochester

Abstract

Languages commonly use physical properties to discuss distinctly non-physical states and events in the world (e.g., "I'm not a *huge* fan of licorice"). Here, we investigate the degree to which associations between physical properties and abstract concepts are culturally specific constructs. To do this, we tested three distinct populations—US adults, US children, and adults from an indigenous group in the lowlands of Bolivia, the Tsimane'—on their associations between the physical concept of weight and a variety of abstract attributes (e.g., importance, emotional state, moral worth). We find a strong relationship between the associations of US and Tsimane' adults, but little-to-no relationship between US children and either adult population. These results suggest that the property of weight plays a similar role in everyday thought across cultures, but that it takes time to develop. Further, we found that these associations could not be recovered from a simple semantic embedding analysis, suggesting that the cross-culturally shared connections between physical and abstract attributes may be learned through more complex experiences than language alone.

Marty: Are you trying to tell me that my mother has got the hots for me?
Doc: Precisely!
Marty: Whoa, this is heavy.
Doc: There's that word again: "heavy." Why are things so heavy in the future? Is there a problem with the Earth's gravitational pull?
(*Back to the Future*. Dir. Robert Zemeckis)

Introduction

Physical notions weigh in on everyday conversation. We say a person *forced herself* to meet a deadline, as though she is pushing a cart uphill. We say a deadline is *fast approaching*, as though an actual train hurtling towards our location in time.

Our concepts of force, causality, space, and substance seem to shape how we talk about the world (Talmy, 1988; Pinker, 2007). Certainly some abstract thoughts rely on a universal understanding of the physical world. If a friend describes writing a paper as 'I'm banging my head against the wall', we can understand they are frustrated, and not literally writing a paper on the effects of head-banging (Figure 1, right). In the reverse direction, our language also shapes how we think about basic concepts, such space and time (e.g. Boroditsky, 2001; Núñez & Sweetser, 2006). And some thoughts are culture- and language-dependent in their meaning. If a friend says writing feels like carrying the day out in a basket, the thrust would not be universally recognized (Figure 1, left).

Figure 1: Details from *Flemish Proverbs* by Pieter Bruegel the Elder, 1559. **Left:** Carrying the day out in a basket, i.e. wasting time **Right:** Banging one's head against the wall.

The purpose of this paper is not to untangle the knot of development, culture, language, and physical concepts. Rather, we mean to pick up one strand of thought as it relates to an understanding of a physical quality, and to tug on it gently. In particular, we consider the concept of weight (heavy and light), which has received less attention in terms of its impact on thought, compared to concepts like force, space, and time [1]. Intuitively, we seem to associate weight with worth: In 2015, a technical teardown of Beats headphones found that 30% of their weight was accounted for by metal objects that add no function, but make them feel 'solid and valuable' (Einstein, 2015). When a character in *Romeo and Juliet* exclaims 'O heavy day!', we recognize that as an expression of dismay at an unfortunate event. But is all this because of the quirks of the English language, and our own WEIRD makeup (Henrich et al., 2010), or something deeper about the way weight ties in with concepts such as worth and sadness?

[1] Weight, mass, and density as physical notions have certainly been studied in development, across infancy (e.g. Baillargeon, 2004), childhood (e.g. Carey, 1999, 2009), and adulthood (e.g. Hamrick et al., 2016). But there the concern is with questions such as 'When do infants realize big things move small things' and 'When do children understand weight and density are separate', and 'Can adults tell which block is heavy', not 'Do children think being weighed down relates to being sad'.

To examine this question, we asked three groups (US adults, US children, and adults from the indigenous Tsimane' of Bolivia) to pick which of two differently weighted, visually identical boxes was better described by various attributes (external, internal, mental and non-mental). We reasoned that if the three groups show a systematic bias for some attributes within the group, but no relationship is found between the judgments of these three groups, then the use of a weight concept in our everyday thinking outside a strictly physical context is more likely to be a cultural construct. If the three groups all show similar judgments, then this is evidence in favor of an early shared conceptual organization. If the Western adults show similar judgments to Tsimane' adults, but are not similar to children, this would suggest a shared conceptual organization, but one that takes time to develop. A final option is that all groups will show random behavior, failing to associate any attribute with the boxes in a systematic way, which would be evidence of certain poor decisions about experiment design or a problem with the fundamental research question. The authors were agnostic about the most likely outcome out of the ones just listed.

Experiment 1: US Adults

Participants

Participants ($N = 100$, 42 female, median age 32.0 years) were recruited through Amazon's Mechanical Turk service (Crump et al., 2013) and paid a monetary sum for their participation, equivalent to $9 per hour. Participants were restricted to those living in the United States.

Materials and methods

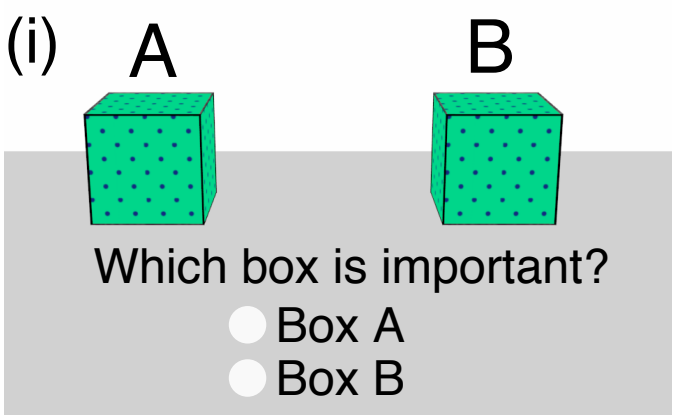
Participants were presented with an image of identical boxes marked A and B (see Figure 2, top), and asked to imagine that there were two boxes before them, as in the image. Participants were asked to imagine lifting up the boxes and discovering that one of the boxes is much heavier (the identity of the heavy box was randomized across participants).

Participants read 12 descriptions in succession, choosing the box that best fit the description. For each description, participants were reminded which box was heavier, and then given a prompt as follows: "Which box is [*attribute*]?", where the attribute varied from one question to the next. Participants indicated their response using a radio button. The 12 attribute adjectives were presented in random order, chosen from a list of 24 possible attributes that reflect inner traits (e.g., good/bad), external qualities (pretty/ugly), emotions (sad/happy), and external evaluation (cheap/expensive, important/unimportant). For a full list, see Table 1.

Each participant saw only 12 attributes, rather than the full list of 24, to prevent cognitive fatigue. Participants always saw only one of a possible antonym pair. In total, this meant there were 50 individual ratings per attribute. Following the attribute questions, the participants supplied basic demographic information, and were invited to share any comments they may have.

Important* / Unimportant* [Not Important]
Valuable* / Cheap* ; Old / Young
Serious* / Funny* ; Sad* / Happy*
Ugly* / Pretty* ; Interesting* / Boring*
Mean / Nice ; Smart* / Stupid* [Not Smart]
Good* / Evil* [Bad] ; Angry / Calm
Brave / Coward [Scared]

Table 1: The 24 attributes applied to the boxes in Experiments 1 and 2, grouped into antonyms. Attributes in [parentheses] indicate a child-friendly replacement for the preceding word, used in Experiment 2. Asterisks indicate words used in Experiment 3.



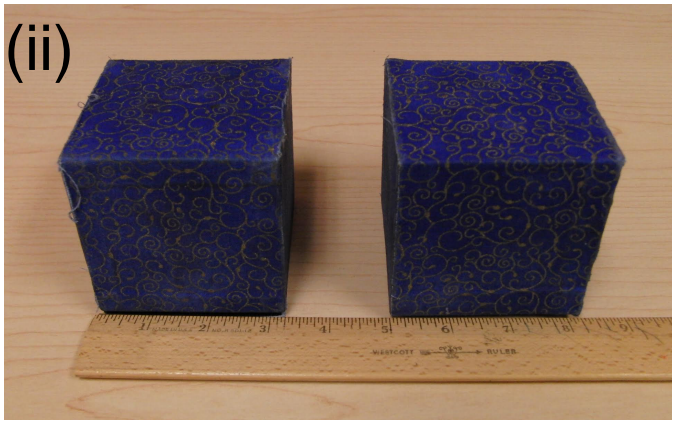


Figure 2: **(i)** Illustration of stimuli shown to participants in Experiment 1, for a specific attribute **(ii)** The boxes used in Experiment 2 with children, and in Experiment 3 with Tsimane' adults.

Results and analysis

Participants' ratings for each attribute were converted into the following measure: $\frac{\#\ participants\ that\ chose\ light}{\#\ total\ participants}$. This weight choice fraction (WCF) goes from 0.0 (all participants chose the *heavy* box for this attribute) to 1.0 (all participants chose the *light* box). The results are shown in Figure 3, with bars indicating 95% bootstrapped confidence intervals (CI, 1000 samples per attribute) around the WCF measure.

Of the 24 attributes, 15 had WCFs with CIs that do not overlap 0.5, indicating that participants considered these attributes as statistically significantly associated with heavy or light. The same result is obtained when using a two-tail binomial test at the $p = 0.05$ level. Such a result is highly unlikely to occur by chance: Using an additional bootstrap analysis that repeats the same procedure from the previous paragraph (counting the number attributes with WCF CIs that do not overlap 0.5), the median (and mean) expected number of attributes with a measure that does not overlap chance is 2. Also, the empirical distribution of participants' WCF is statistically significantly different from a distribution drawn from a random sample that assumes the same participant numbers, but with answers based on an unbiased coin flip (Kolmogorov-Smirnov two-sided test for 2 samples, $KS = 0.28$, $p < 0.05$).

The attributes significantly associated with *heavy* and *light* seem partially in line with intuition [2]. Heavier boxes are more likely to be seen as valuable, important, and interesting, as opposed to the cheap, unimportant, and boring lighter boxes. This is consistent with the marketing-driven decision by Beats to add superfluous weight to their headphones. This association makes sense given that more weight may imply more "stuff", which could generally be considered more desirable [3]. Participants also associated more personality-type traits with the boxes, in a way that is not accounted for by a simple positive-negative spectrum. Heavy boxes are more good and brave, but also mean and angry. Lighter boxes are more cowardly, but also more pretty. Presumably participants were able to anthropomorphize the boxes to some degree, seeing them as agents. For example, on this analysis a light agent is more likely to run away, and is likely to be younger. However, this does not account for the full pattern of results, such as seeing heavier boxes as more "good" and less "evil".

This pattern also cannot be recovered from a semantic embedding analysis. The analysis worked as follows: We embedded the attributes from Table 1, as well as the words *heavy* and *light*, in a high-dimensional semantic vector space, which was constructed using the co-occurrence statistics of several hundred-thousand words in a large corpus (Pennington et al., 2014). Specifically, we used 100-dimensional GloVe word vectors pre-trained on the Wikipedia 2014 + Gigaword 5 datasets. Such semantic embeddings have proved useful for measuring similarity between words, in the service of machine-learning applications such as sense-making, translation, and question answering (see for example Vedantam et al., 2015; Wolf et al., 2014; Yu et al., 2015). Intuitively, a shorter euclidean distance or larger cosine similarity between two points in this space indicate a larger degree of similarity between the words that those points represent. After embedding our terms, we measured the relative euclidean distance between the attributes and the terms *heavy* and *light* (that is, $distance(heavy, attribute_i) - distance(light, attribute_i)$). We found no correlation between participants' response and this distance. A similar analysis with cosine similarity also found no such relation. This suggests that while useful, basic semantic embedding does not necessarily capture association that is based in physical properties.

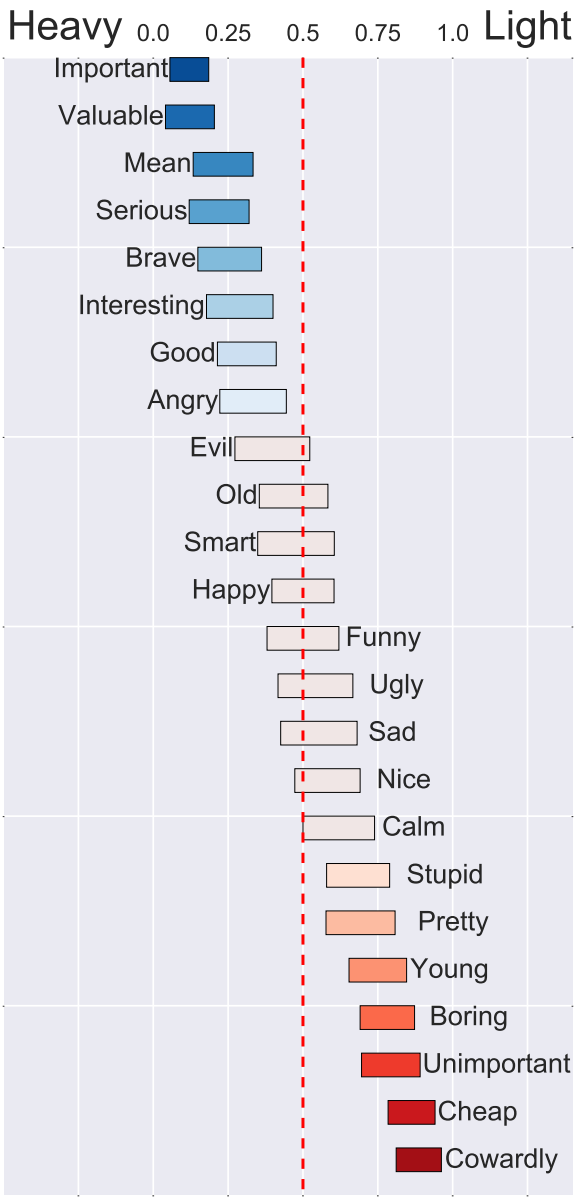


Figure 3: Results of Experiment 1: Participant responses per attribute are converted into WCF measure running from 0 (all participants chose the heavy box) to 1 (all participants chose the light box). Bars indicate 95% bootstrapped confidence intervals around the mean of this measure. Colors indicate the degree to which an attribute is associated with heavy (blue) or light (red). Beige indicates WCFs with CIs overlapping 0.5, indicating a random response or equal association.

[2] That is, with the intuition of the Western adult authors of this paper.

[3] Although, in contrast, consider the value and importance currently associated with slim technology products, or human figures. By the same logic, *slim* would imply less "stuff". Thus, this explanation is insufficient on its own.

While the pattern shown by US adults is interesting on its own, the original driving motivation was comparing this pattern to children and non-US cultures. With that, we turn to children.

Experiment 2: US Children

Participants

Fifty individuals were recruited from the Rochester Kid Lab participant pool (28 female, Median 4.0 years, range 3-6[4]).

Materials and Methods

Participants were tested in a designated room in the Rochester Kid Lab. Parents gave their informed consent, and generally did not accompany their children during the test, unless requested, or children expressed shyness. Parents who accompanied their children were explicitly advised not to encourage responses from their child. Families were compensated for their time and child participants were also given a small gift (a shirt or toy).

In the testing room, participants were asked to sit next to a table, where two boxes were laid out. The boxes were 3x3x3 inches, made of wood, and covered in blue fabric with a gold pattern (see Figure 2, bottom). The boxes were hollow, and inside one of them was a 200 gram metal weight, along with padding to prevent the weight from bouncing and rattling when the box was handled. The locations of the boxes with respect to a participant were randomized across children.

Participants were first asked if they noticed a difference in the boxes, based on visual appearance. Participants were then asked to hold the boxes, and to indicate if there was a difference (which they were able to verbally verify).

The participant continued to hold the boxes in each hand, as they answered the following question: "Which box do you think is [attribute]?", for a randomized set of 12 attributes taken from 24 attributes similar to Experiment 1 (and see Table 1). The study took a maximum of 10 minutes. Participants answered verbally or with a gesture, with the experimenter noting their response. Participants were also asked to explain their answers, but their reasons were scantly supplied and proved inconsistent across children [5]. Again, to prevent cognitive fatigue, participants were asked to judge 12 attributes rather than the full 24, with each participant seeing only one of a possible pair of antonyms. Thus there were 25 individual ratings per attribute.

Results and analysis

Participants' ratings for each attribute were again converted into the WCF measure used in Experiment 1 (with 0.0 indicating all participants chose heavy, and 1.0 indicating all chose light). In this case, however, only 3 attributes were different from chance, using a two-tailed binomial test at the $p = 0.05$ level). The empirical distribution of children's WCFs was also not statistically significant from a distribution drawn from a random sample, one that assumes the same participant numbers but with answers based on the flip of an unbiased coin (Kolmogorov-Smirnov two-sided test for 2 samples, $KS = 0.21$, $p = 0.22$). It is possible to conclude that children did not understand the task, either because of low-level explanations like inappropriate materials and framing, or because physical weight does not play a similar association role in their general thought as it does for US adults.

When correlating with the responses of adults from Experiment 1, we find there is a weak correlation ($r_s = 0.4, p = 0.05$, and see Figure 4). Taken in a positive light, this may indicate a fledgling understanding after all of the full adult association between the attributes used and physical weight. Still, this relationship is statistically tenuous. A median-split by age does not show a difference between younger and older children.

Is it possible US adults exhibit a culture-specific pattern of association with physical weight, one that requires years to acquire? In the last experiment, we consider a non-WEIRD adult population, an indigenous people of Bolivia.

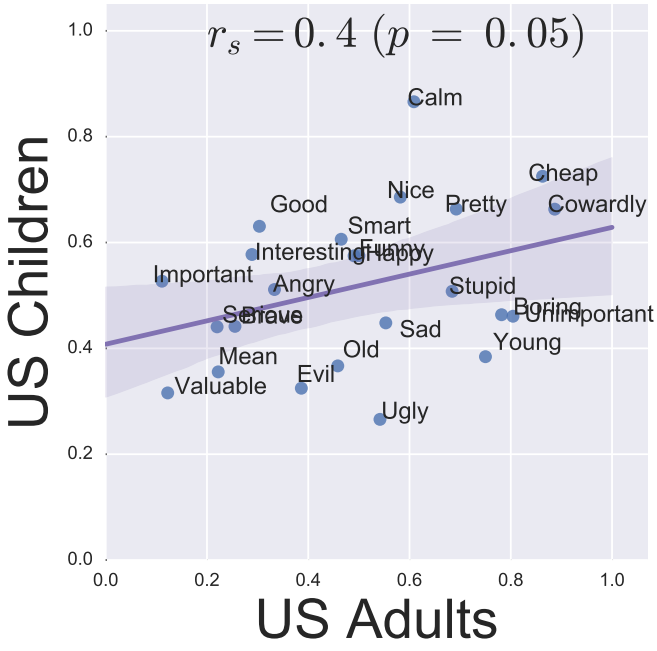


Figure 4: Comparison of adult responses from Experiment 1 with child responses from Experiment 2. The x and y axis both use the same weight-fraction measure, going from heavy to light. The shaded area indicates a 95% bootstrapped confidence interval on the linear regression model fit.

Experiment 3: Tsimane' Adults

The Tsimane are a native people of lowland Bolivia, consisting of several thousand individuals, who live in mostly small

[4] At this age children possess a sufficiently large vocabulary and can correctly point to a heavier object when prompted, but have not received much formal education.

[5] One participant designated the heavy box 'The Hulk' and the light box 'Captain America'. Captain America was funny, and The Hulk was not smart.

communities in the northeastern department of Beni. Traditional Tsimane' are farming-foragers who subsist off hunting, fishing, and some farming and trade. Members of the Tsimane' have highly variable education levels, and own few artifacts (Huanca, 2006; Reyes-García, 2001). As members of a relatively isolated non-industrial society, Tsimane' have been the topic of several previous studies, from market behavior (Reyes-García, 2001) to counting (Piantadosi et al., 2014), to color concepts (Cibelli et al., 2016), to notions of fairness (Jara-Ettinger et al., 2016).

Participants

Our final sample included fifty-five individuals (33 female, median age = 28.0 years, range 17-65) from twelve Tsimane' communities.

Materials and methods

Experiments took place in a community classroom, with a translator reading from a script, and a separate transcriber recording responses. The experiments were conducted in Tsimane', translated from a Spanish script. The translation was confirmed by a second Spanish-Tsimane' translator. Other people were present in the room, but could not see participant responses. Participants were compensated with gift bags equivalent to roughly $10 per hour. Participants completed other tasks in addition to the one in this study, with a total testing time of approximately one hour.

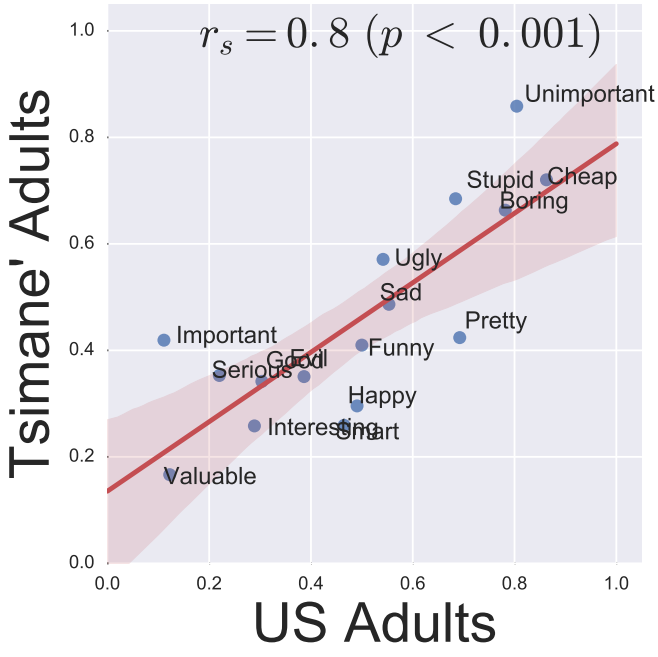


Figure 5: Comparing US adults from Experiment 1 with Tsimane' adults from Experiment 3. The x and y axis both use the same WCF measure, going from heavy to light. Shaded area indicates a 95% bootstrapped confidence interval.

Participants were presented with two identically marked boxes on a table. These were same boxes used in Experiment 2, and weighted as in Experiment 2 (see Figure 2, bottom). Participants were instructed to pick each box up before any questions were asked. Participants were then asked: "Which box is [attribute]?", and were instructed to point to a box. This was then repeated until all adjectives were covered. The order of the adjectives was randomized, as was the particular adjective from a given pair was randomized. Participants were allowed to pick up the boxes at any point. As participants saw only one word out of a possible pair, there were on average 21 individual ratings per attribute. In general, participants in this experiment went over a subset of 16 of the 24 adjectives in Table 1, due to translation difficulties.

Results and analysis

As in Experiments 1 and 2, participants' ratings for each attribute were transformed into a WCF measure. Four of the 16 attributes were different from chance, using a two-tailed binomial test at the $p = 0.05$ level). In addition, the empirical distribution of Tsimane' WCFs is statistically significant from that drawn from a random sample that assumes the same participant numbers, but with answers sampled from an unbiased coin flip (Kolmogorov-Smirnov two-sided test for 2 samples, $KS = 0.36$, $p < 0.05$).

We also correlated Tsimane' responses with those of US adults in Experiment 1. We found a significant correlation ($r_s = 0.8, p < 0.001$, and see Figure 5). As a final comparison, we correlated Tsimane' responses with those of US children in Experiment 2, and found no significant correlation ($r_s = 0.0, p = 0.94$). We next consider the general pattern of results.

Discussion

Thoughts weigh nothing, but they can weigh heavily on us. A man might feel lighthearted after dispensing with a heavy obligation. We can take matters lightly, but we should not take them too lightly.

Thoughts, obligations, and matters don't actually weigh anything, but we feel their press on us. Our language of thought cleaves the world into concepts that behave like objects with physical properties, located in space and acted on by force (Pinker, 2007). Conversely, our mental concepts can color our perception of the physical. In this paper we considered the particular physical notion of weight, and its relation to different non-physical qualities such as value, interest and seriousness.

We examined people's associations between weight and these different qualities in Western adults and children, and in members of the non-industrial Tsimane' society. We found a strong relation between the answers given by Tsimane' and Western adults, a tenuous relation between Western adults and children, and no relation between Tsimane' adults and Western children. Taken together, these findings indicate that weight acts a similar cross-cultural role in everyday thought, but that it takes time to fully get its act together. So, it may be language and culture-independent to think of important matters as physically weighing more, for example. However, a

fuller treatment would require relating the attributes to other measures beyond weight, such as imageability and affect.

Different alternative explanations can be put forward for why children provided responses that were inconsistent with one another. First, it is possible that children simply have not had the life experiences required to form strong, systematic associations between abstract attributes and physical properties like weight. Alternatively, it is possible they cannot anthropomorphize the boxes. This seems unlikely, as children can engage in pretend play with inanimate objects, but attributing metarepresentations may have required a more active signaling of the task as pretend play (Lillard, 1993). It may be that young children lack the basic physical skills associated with telling a heavy object from a light object and predicting their different behaviors, but previous research shows most of the basic intuitions are in place by the lower end of our age range, with young children predicting the effects of different masses interacting, and taking weight into account when planning actions (Baillargeon, 2004; Upshaw & Sommerville, 2015). Under these alternative explanations that posit children could have experienced confusion about the task, we would typically expect certain behavioral indicators of this state, such as failures, resistance or delays in providing responses. However, in our sample we observed no such indicators. Children were generally swift and willing to select a particular box for each attribute about which we inquired.

This study does not give a definitive final answer to questions of culture, development, and the constructs of thought, but it does shed light on a piece of the puzzle, in the form of weight. Also, it is not hard to think of other physical properties that our methodology could stretch to accommodate, roughly speaking.

Acknowledgments

We are grateful to Holly Palmeri for her support recruiting and scheduling children in the Rochester Kid Lab, to Tomas Huanca for logistical help with visiting the Tsimane' villages, to Ted Gibson and Julian Jara-Ettinger for provoding support in Bolivia, and to Dino Nate Aez, Robertina Nate Aez, and Salomon Hiza Nate for help translating. We would also like to thank Steve Piantadosi for doing some of the heavy lifting. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*, 89–94.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive psychology*, *43*(1), 1–22.

Carey, S. (1999). Knowledge acquisition: Enrichment or conceptual change. *Concepts: core readings*, 459–487.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The sapir-whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PloS one*, *11*(7), e0158725.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, jan). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

Einstein, B. (2015). *How its made series: Beats by dre.* Retrieved 2015-06-15, from `https://blog.bolt.io/how-it-s-made-series-beats-by-dre-154aae384b36`

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex physical scenes via probabilistic simulation. *Cognition*, *1*, 2.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Huanca, T. (2006). *Tsimane'oral tradition, landscape, and identity in tropical forest*. Tomás Huanca L.

Jara-Ettinger, J., Gibson, E., Kidd, C., & Piantadosi, S. (2016). Native amazonian children forego egalitarianism in merit-based tasks when they learn to count. *Developmental Science*, *19*(6), 1104–1110.

Lillard, A. S. (1993). Pretend play skills and the child's theory of mind. *Child development*, *64*(2), 348–371.

Núñez, R. E., & Sweetser, E. (2006). With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive science*, *30*(3), 401–450.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *EMNLP*, *12*, 1532–1543.

Piantadosi, S. T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, *17*(4), 553–563.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Reyes-García, V. (2001). *Indigenous people, ethnobotanical knowledge, and market economy. a case study of the tsimaneamerindians in lowland bolivia.*

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive science*, *12*(1), 49–100.

Upshaw, M. B., & Sommerville, J. A. (2015). Twelve-month-old infants anticipatorily plan their actions according to expected object weight in a novel motor context. *Frontiers in public health*, *3*.

Vedantam, R., Lin, X., Batra, T., Lawrence Zitnick, C., & Parikh, D. (2015). Learning common sense through visual abstraction. In *Proceedings of the ieee international conference on computer vision* (pp. 2542–2550).

Wolf, L., Hanani, Y., Bar, K., & Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. *International Journal of Computational Linguistics and Applications*, *5*(1), 27–44.

Yu, L., Park, E., Berg, A. C., & Berg, T. L. (2015). Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the ieee international conference on computer vision* (pp. 2461–2469).