Forthcoming in *Cognition*

Assessing Abstract Thought and its Relation to Language with a New Nonverbal Paradigm: Evidence from Aphasia

Peter Langland-Hassan*, Frank R. Faries*, Maxwell Gatyas*, Aimee Dietz**,

and Michael J. Richardson***

*Department of Philosophy, University of Cincinnati
**Department of Communication Sciences and Disorders, University of Cincinnati
***Department of Psychology, Macquarie University

Abstract: In recent years, language has been shown to play a number of important cognitive roles over and above the communication of thoughts. One hypothesis gaining support is that language facilitates thought about abstract categories, such as *democracy* or *prediction*. To test this proposal, a novel set of semantic memory task trials, designed for assessing abstract thought non-linguistically, were normed for levels of abstractness. The trials were rated as more or less abstract to the degree that answering them required the participant to abstract away from both perceptual features and common setting associations corresponding to the target image. The normed materials were then used with a population of people with aphasia to assess the relationship of abstract thought to language. While the language-impaired group with aphasia showed lower overall accuracy and longer response times than controls in general, of special note is that their response times were significantly longer as a function of a trial's degree of abstractness. Further, the aphasia group's response times in reporting their degree of confidence (a separate, metacognitive measure) were negatively correlated with their language production abilities, with lower language scores predicting longer metacognitive response times. These results provide some support for the hypothesis that language is an important aid to abstract thought and to metacognition about abstract thought.

Keywords: abstract concept, language, abstraction, aphasia, metacognition

1. Introduction

Most animals think, in some sense of the word 'think.' What, if anything, is distinctive of human thought? A common answer is that humans are uniquely capable of *abstract* thought, reflected in the abstract lexicon of our spoken languages. For example, in English, words like 'integrity,' 'philosophy,' 'love,' 'justice,' and 'mind' mark intuitively abstract notions. This invites the question: what is the relationship between abstract thought and the words we use to express it?

An early and still influential answer to this question assigns to language a merely communicative role, with words enabling the expression of thoughts, while being inessential to abstract thought itself (Fodor, 1975; Pinker, 1994). On this view, someone lacking language could think all the same thoughts as a fluent speaker without being able to put their thoughts into words—not even "in their head," as inner speech. Recently, however, a competing hypothesis sees language as not only a vehicle for communicating thoughts but as a resource that *supports* or *enables* certain forms of thought (Binder, Westbury, McKiernan, Possing, & Medler, 2005; Borghi et al., 2017; Condry & Spelke, 2008; Dove, 2014; Langland-Hassan, Gauker, Richardson, Dietz, & Faries, 2017; Lupyan & Bergen, 2016; Wang, Conder, Blitzer, & Shinkareva, 2010; Yee, 2019). In particular, language is seen by many as a crucial support or tool for *abstract* thought (Borghi et al., 2017; Boroditsky, 2001; Davis & Yee, 2019; Thibodeau, Hendricks, & Boroditsky, 2017). On some of these views, linguistic labels and word associations provide a kind of cognitive shortcut to abstract conceptual information (Barsalou, 2008; Wilson-Mendenhall, Simmons, Martin, & Barsalou, 2013). According to others, language facilitates thoughts about abstract relationships—especially when a relationship's holding between two entities can be difficult to grasp in purely sensory terms (Borghi et al., 2019; Boroditsky, 2001; Dove, 2018; Gentner & Boroditsky, 2001; Lupyan & Lewis, 2019; Reilly, Westbury, Kean, & Peelle, 2012). On other views, language serves a more fundamental role as a representational *medium*, providing an "internalized amodal symbol system" needed for abstract thought (Dove, 2014, 2019). These by no means exhaust the current menu of options (for review, see Borghi et al. (2017), Yee (2019), and Bolognesi & Steen (2018)).

The present study aims to build on this existing work by investigating the relation of language to a specific form of abstract thought—namely, the ability to abstract away from the perceptible features and salient thematic associations of two items to grasp some other commonality. To that end, we compare the performance of a group of people with aphasia ("PWA") to that of controls on a novel set of non-verbal stimuli, designed and normed for the purpose of assessing this kind of abstract thought. This paper reports both the norming process itself and the subsequent experiment with PWA. In the balance of this introduction, we will situate the study's conception of abstract thought with respect to other common conceptions in the literature, as a means to motivating the measure of abstract thought developed in the norming study. A general theme is that existing measures of abstract thought are typically tied to words in a way that makes them unideal for use in a setting where the relationship of abstract thought to language itself is being assessed. We then explain (Section 1.7) why we predict that a capacity for the kind abstract thought assessed here—and for metacognitive awareness of such abstract thought—would be disrupted by a loss of language abilities.

1.1 Defining "abstract thought"

A difficulty in studying abstract thought lies in specifying what qualifies a form of thought as "abstract" in the relevant sense (Gilead, Trope, & Liberman, forthcoming). If we adopt the common view that concepts are the building blocks of thought—with thoughts being composed of concepts—we can ask what it is that makes a concept abstract. This question belies a misunderstanding, however, as *all* concepts are abstract in a sense, insofar as grouping any set of individuals under a single concept involves *abstracting away* from properties they do not share to focus on ones they do (for review, see Borghi et al., 2019; Gilead et al., forthcoming; Yee, Jones, & McRae, 2018). To recognize both a black Poodle and a Golden Retriever as *dogs,* for instance, we need to abstract away from—i.e., ignore—their differences in color and shape to focus on their shared doghood.

1.2 Concreteness, imageability, and Sensory Experience Ratings

Nevertheless, a relative sense of abstractness, whereby some concepts are *more* abstract than others, can be defined. A common means for doing so is by appeal to the relative

concreteness or *imageability* of words associated with the concept (Brysbaert, Warriner, & Kuperman, 2014; Coltheart, 1981; Cortese & Fugett, 2004). Concreteness is typically understood as the extent to which a particular item or event can be experienced by the senses; a concreteness *rating* for a word is generated by averaging the values participants give when asked to assess how easy it is to perceive the item named by the word (Brysbaert et al., 2014; Medler & Binder, 2005). This way of defining abstract thought gives rise to the well-known "concreteness effect," whereby words with higher concreteness ratings are processed faster in lexical decision tasks and are associated with better performance in naming and recall (Begg & Paivio, 1969; Kounios & Holcomb, 1994; Schwanenflugel & Shoben, 1983). Comparable results have been found with respect to the related measure of imageability, where imageability is understood as the subjective ease with which a word gives rise to a related sensory-motor mental image (Cortese & Fugett, 2004; Paivio, 1971). (A third related, but less commonly used, measure is Juhasz & Yap's (2013) Sensory Experience Ratings (SER).[1])

A straightforward way to assign a degree of abstractness to a concept is to associate its abstractness with the concreteness, imageability, or SER rating given to the word that expresses the concept. A limitation of this method is that, in being tied to specific words, such ratings are of limited use in assessing the relation of abstract thought to words themselves. If words—with their associated ratings—are used in experimental stimuli, there is a risk of conflating a capacity to process linguistic items with a capacity for abstract thought. On the other hand, if the experimental stimuli exclude words, it can be difficult to assess which concepts are triggered by a stimulus—and, thus, which if any concreteness, imageability, or SER ratings should be used to rate the abstractness of the stimulus. The novel measure of abstractness and concreteness developed below (which we call "Trial Concreteness") aims to overcome these problems, while respecting motivations of these word-related rating systems in ways we will discuss.

1.3 Hierarchical notions of abstractness

[1] To assign Sensory Experience Ratings (SER) to words, Juhasz & Yap asked participants to "rate the degree to which each word evoked a sensory experience, on a 1 to 7 scale, with higher numbers indicating a greater sensory experience" (2013, p. 163). Juhasz & Yap propose that SER ratings may avoid-vision centric judgments encouraged by standard imageability rating-prompts, which ask participants to rate their ease in generating word-related "images" or "imagery."

A second commonly discussed notion of abstract thought derives from hierarchical relationships among categories, with concepts of superordinate (i.e. more inclusive) categories being rated as more abstract than concepts of subordinate (i.e. less inclusive) categories (Rosch, 1978; Yee, 2019). For instance, in the hierarchical sense of abstractness, the concept OBJECT is more abstract than the concept MISSILE, because the category of objects is superordinate to that of missiles: all missiles are objects, but not all objects are missiles. Concepts themselves can be seen as stores of semantic knowledge arrived at through processes of abstraction from regularities in sensory input and related motor outputs (Yee, 2019). A concept of a superordinate category—such as OBJECT—will tend to be more *abstracted* from such sensorimotor correlates than a concept corresponding to one of its subordinate categories (such as CHAIR), rendering the concept itself "more abstract." In short, the more abstracted a store of knowledge is from its sensorimotor correlates, the more abstract the concept is that constitutes that knowledge.

While it may be tempting to view concepts that are relatively abstract, in this hierarchical sense, as corresponding to words with low concreteness or imageability ratings, the relation between the two is not straightforward. Whether the members of a given category share many, or only a few, salient perceptible similarities—and thus whether that category is highly abstracted from related sensorimotor information—is not what subjects are asked to assess when providing imageability or concreteness ratings for a word. Following Paivio (1968), Cortese and Fugett solicit imageability ratings by informing participants that "any word that…arouses a mental image…very quickly and easily should be given a high imagery rating," and "any word that arouses a mental image with difficulty or not at all should be given a low imagery rating" (2004, p. 387). Similarly, in soliciting concreteness ratings, Brysbaert et al. (2014) explain to subjects that they should assign high a concreteness rating to a word to the extent that it "refers to something that exists in reality; you can have immediate experience of it through your senses," while low concreteness should be assigned to words that "refer to something you cannot experience directly through your senses or actions" (2014, p. 906). *Prima facie,* there is no reason it should be more difficult to form images of members of superordinate as opposed to subordinate categories, and no reason members of a subordinate category should be judged to exist in reality, and to be perceptible, to a greater degree than members of superordinate categories. After all, to form an image of (or to perceive) a chair is simultaneously to form an

image of (or to perceive) a member of the subordinate category 'chair' *and* the superordinate category 'object.'

Nevertheless, participants often appear to answer imageability and concreteness prompts *as though* they are being asked to assess hierarchical relations. The noun 'object,' for example, has imageability and concreteness ratings of 408 and 487, respectively, compared to its subordinate 'cat', which has corresponding ratings of 617 and 615 (Coltheart, 1981), despite the fact that cats are themselves objects and therefore cannot be more real, or more readily perceptible than objects. Likewise, 'object' receives only a 2.2 SER rating, while 'missile' has a 4.45 SER rating, even though missiles are themselves objects and thus cannot be easier to form a sensory image of than an object. (Similarly, on the Brysbaert et al. (2014) concreteness ratings, 'object' receives a score of 3.66, whereas 'missile' has a score of 4.83.) Thus, whether or not it is warranted by the nature of the prompts used to solicit such ratings, the kind of abstractness expressed by concreteness, imageability, and SER ratings aligns fairly well with the kind that is associated with hierarchical relations among categories, where concepts of superordinate categories are more abstract than concepts of subordinate categories, because they are stores of knowledge whose sensorimotor correlates abstract-away from past experience to a greater degree.[2]

As with imageability, concreteness, and SER scores, there are limitations to using the superordinate-to-subordinate relation as a measure of abstractness when assessing the relation of language to abstract thought. One limitation is that such relations are far clearer *within* particular hierarchies than across them, limiting the ability to compare a participant's performance on a wide range of concepts with different levels of abstractness. 'Labrador' is subordinate to 'dog,'

[2] Notably, the tendency to answer imageability and concreteness rating prompts as though hierarchical relations are being queried is inconsistent. The category 'dogs' is superordinate to 'Labrador', yet 'Labrador' has a lower concreteness score of 4.35; similarly, the superordinate category 'bird' has a rating of 5 while the subordinate 'robin' has a rating of 4.61 (Brysbaert et al. 2014). One way to make sense of this apparent instability in how such prompts are interpreted by subjects is to follow Rosch (1978) in proposing that there is a "basic" level of category representation that marks the most inclusive level at which there are attributes common to all or most members of the category. On Rosch's framework, a category such as *table* can be particularly salient for participants—and, accordingly, may seem easier to mentally image (and likewise generate higher imageability, SER, or concreteness ratings)—if it has a privileged place in one' conceptual scheme as a basic level category.

for instance, and therefore less abstract in the hierarchical sense. But which of 'Labrador' or 'screwdriver' is more abstract? We are not aware of any objective ratings that would enable such comparisons. A second limitation is that it is not always possible to place concepts within meaningful hierarchies, and thus not possible to assign hierarchy-based abstractness ratings. This is most obvious for verb and adjective concepts, but extends also to numerous putatively abstract noun concepts, such as CONCEPT, HOPE, PHILOSOPHY, and JUSTICE. A third limitation is that, in the context of a non-verbal stimulus, it can be difficult to discern which category concept is elicited by a stimulus, and thus which hierarchical level of category is relevant to rating the abstractness of the stimulus. The measure of abstract thought developed below overcomes these limitations, while again preserving the connection between the notion of abstract thought and the process of abstracting away from past perceptual experience.

1.4 Low versus high dimensionality and trial-relativity

A third approach to defining abstract thought—especially influential to the present study—can be found in Lupyan and Mirman (2013) and Perry and Lupyan (2017). In place of an abstract/concrete concept distinction, Lupyan and Mirman (2013) explore differences in what they term "low-dimensional" and "high-dimensional" categories, comparing performance between a population with aphasia and controls. Their distinction between high and low dimensional categories itself echoes the notion of *dense* versus *sparse* categories (Sloutsky, 2010) and the distinction between resemblance (or association)-based categories and rule-based categories (Couchman, Coutinho, & Smith, 2010; Minda, Desroches, & Church, 2008). On Sloutsky's account, statistically dense categories "have multiple intercorrelated (or covarying features) relevant for category membership," while sparse categories have members that share "very few relevant features" (2010, p. 1250). Similarly, a high-dimensional category, on Lupyan and Mirman's understanding, is one where the things united under the category share many salient features, while a low-dimensional category groups items that share only one or a few salient characteristics.

On its face, the proposal that lower dimensional categories are more abstract than high dimensional aligns well with the idea that superordinate categories are more abstract than subordinate ones. Superordinate categories compare to low dimensional categories in that their members tend to share fewer salient features than subordinate (and high dimensional) categories.

However, there is an important difference in these two ways of rating abstractness. The superordinate/subordinate distinction applies to categories. Whereas, properly understood, high and low dimensionality apply to individual *trials,* and not to specific categories.[3] Each of Lupyan and Mirman's trials presented participants with twenty images of familiar objects such as foods, vehicles, tools, and animals. Participants were then asked to select images that met a certain criterion. In an example of a low-dimensional trial, participants were asked to identify, from among the twenty images, all and only the *things that are blue.* Here the idea was that objects so grouped would have little or nothing in common other than their color, making the trial low-dimensional. To successfully group all of the blue items in Lupyan and Mirman's "things that are blue" trial, the blue objects' many differences must be ignored—they must be abstracted-*away* from—while only their color remains a point of focus. By contrast, in an example of a high-dimensional trial, participants were asked to identify all the pictured *fruit.* Unlike 'things that are blue,' members of the category 'fruit' share multiple salient properties, such as being edible, sweet, found in the produce department, and alive.

But consider now a hypothetical trial where participants are asked to group 'things that are yellow' (another of Lupyan and Mirman's low-dimensional trials), yet where all of the yellow items among the choices are bananas, and all of the distractor items are vehicles. The trial could in that case be considered high dimensional, insofar as many different properties (shape, flavor, use, common setting) could serve to anchor selection of the correct (yellow) items. It is only in a context where, relative to the distractor items, the correct choices share no salient similarities other than their color that 'things that are yellow' becomes a low-dimensional trial. Going in the other direction, suppose that, on the 'fruit' trial, all of the distractor items are vegetables that share rough visual similarities with fruits and that fruits with uncommon flavor characteristics (e.g., tomatoes and eggplants) are among the correct choices. What was formerly a high-dimensional trial is now low dimensional due to the similarity of the distractor items to the correct choices. The close perceptual and thematic similarities of the distractor items to the

[3] Lupyan & Mirman make no explicit mention of the task-relativity of low/high dimensionality. While we think they would agree with that dimensionality is trial-relative, they sway between speaking of *trials* being high or low dimensional and the *categories* themselves being high or low dimensional: "We call such *trials* low-dimensional. We reasoned that because such *categories* cohere on the basis of one or a small number of dimensions, they may require more on-line support from language" (2013, p. 1188, emphasis added).

correct choices forces participants to abstract away from almost all of the salient properties of the correct choices to focus on just one that unites them—*viz*., being a fruit—in order to arrive at the correct grouping.

Note, however, that while the degree of abstractness pertaining to high and low dimensionality may be trial-relative (i.e., relative to distractor items), there remains a close kinship between this notion of abstractness and that pertaining to hierarchical category relations. In both cases, the degree of abstractness increases as there are fewer salient features uniting a class of items. We have simply observed that, in some cases, features of a context—and not simply the uniting category itself—can play a role in determining how many salient features must be abstracted away from in order to appreciate a commonality between two or more things. In the next section, we introduce the term 'Trial Concreteness' as a measure of this sort of trial-relative abstract thought and relate it more definitively to traditional measures of (what we will call) 'Concept Concreteness,' which is the concreteness of the (trial-independent) concept that links the target and match.

1.5 Trial Concreteness compared to Concept Concreteness

The trial-relativity of low and high dimensionality is especially important to the present study, as it is likely to occur within many other non-verbal assessments of abstract thought, which are themselves important to investigating the relation of language to abstract thought. In particular, such relativity occurs within standard pictorial semantic memory tasks, which serve as a framework for the test of abstract thought developed here. On a pictorial semantic memory task—such as the Camel and Cactus Test (CCT) (Bozeat, Lambon Ralph, Patterson, Garrard, & Hodges, 2000)—a target image is shown with four choice images below it, and the participant is asked to indicate which choice image best goes with the target image. While existing semantic memory tasks of this sort are not rated for the level of abstract thought they require, it is natural to think that some such trials will require more abstraction from past perceptual regularities—and, correspondingly, use of concepts that are themselves more abstract—than others. The aim of our norming study was to generate abstractness ratings for multiple trials of that sort, spanning a wide range of abstractness levels. To do so properly, the relativity of abstractness level—as a function of the distractor items—must be taken into account.

To see this vividly, suppose that the target image on a pictorial semantic memory trial is of a pear and the image it is to be matched with is of an apple (both being fruit). In a situation where the three distractor images are all of tools—a hammer, wrench, and screwdriver, for example—there are many salient features the pear shares with the apple that can facilitate linking one with the other. For that reason, we could say that it is a high-dimensional and (to extend ordinary usage somewhat) relatively *concrete* trial. But now consider a situation where the three distractor images are of vegetables—an artichoke, carrot, and onion—that share many features with fruits. In that context there are fewer salient characteristics shared only by the pear and apple to facilitate linking one with the other. It has become a lower-dimensional, more abstract trial. Nevertheless, in both cases, the *concept* linking target and match is the same—namely, FRUIT. Thus, the degree of abstraction required to arrive at the correct choice in a trial can vary while the concept, label, or category linking the target and match remains the same.

It is helpful, then, to distinguish two senses of concreteness, each with its corresponding notion of abstractness. First, there is *Concept Concreteness*, which is concreteness linked to individual words or concepts. Concreteness ratings, imageability ratings, and SER ratings are all ways of measuring concept concreteness, so understood. In addition, while there are no standard numerical scores corresponding to the hierarchical notion of concreteness earlier discussed, a category's place in a hierarchy can be considered an additional measure of the concreteness of the concept corresponding to that category.

But we can also speak of *Trial Concreteness,* where a semantic memory trial (of the sort just described) has high concreteness if there are many salient features shared by the target and match that, in the context, can be used to distinguish them as falling under a single category, and is abstract to the extent that there are very few features that can be used to so distinguish them. The two imagined versions of a fruit-related semantic memory task just described have different levels of Trial Concreteness, due to the difference in the distractor items. Yet, in each case, the concept (FRUIT) linking the target and match is the same. In that sense, we can say the Concept Concreteness of each task is the same.

The ability of Trial Concreteness ratings to differ as a function of context meshes with the idea that concepts themselves are context-dependent, insofar as there is no static representational structure exploited across contexts that trigger (what otherwise might seem to

be) the same concept (Barsalou, 1987; Casasanto & Lupyan, 2015; Yee & Thompson-Schill, 2016). Instead, it is proposed, concepts are "constantly changing" and "inextricably linked to their context" (Yee & Thompson-Schill, 2016). The difference in Trial Concreteness in the two fruit trials described above tracks some of this context-relativity, insofar as it captures the way in which exercising (or triggering) a concept may require more or less abstraction from present sensory input in different contexts.

1.6 Two components of Trial Concreteness: Visual Similarity and Common Setting

In the norming study described below, we generated Trial Concreteness ratings for a variety of pictorial semantic memory trials, with lower Trial Concreteness ratings corresponding to more abstract trials. We identified two dimensions along which a trial might differ in Trial Concreteness. First, a trial could be more concrete to the extent that the target and match share many visually perceptible similarities, in comparison to the target and distractor items. This aligns well with the sense in which subordinate categories are more concrete than superordinate ones (on the assumption that members of subordinate categories share more perceptible features than members of superordinate category). It also aligns with the sense in which categories with low imageability, concreteness, and SER scores are relatively abstract, at least insofar as categories are rated lower in concreteness or imageability when their members share fewer salient perceptible features.

A second dimension of abstractness relevant to semantic memory tasks consists in whether the two matching items are often found in a common setting and, for that reason, are strongly associated. A fork and a plate, for instance, do not share many visual similarities. However, they are thematically associated due to their typically appearing together at a commonly experienced type of event. Thematic connections of this sort are highly salient and strongly shape perceiver expectations. They are learned earlier than superordinate categorical relationships (Markman, 1981, 1990), and both children and adults default toward sorting items by thematic relationships (as opposed to categorical or taxonomic ones) when not given a word by which to sort (Lin & Murphy, 2001; Markman, 1990). Further, adults tend to sort items more quickly by thematic relation than by functional category (such as "footwear") (Kalénine, Mirman, Middleton, & Buxbaum, 2012; Kalénine et al., 2009). And, more generally, the

presentation of a word or image primes recognition of thematically related items (Estes & Jones, 2009; Mirman & Graziano, 2012).

In view of the high saliency of thematic associations, we propose that, just as a process of abstraction is required in order to sort perceptually dissimilar items together into a superordinate category (where one abstracts away from the salient perceptual features of the target to match it with another item), so, too, is abstraction involved in linking items that do not tend to occur together in a commonly experienced type of event. In the latter case, the participant must abstract away from the perceptual features not of things of the same categorical kind, but of things that, in one's experience, are commonly found together with the item. While this is not the very same notion of abstractness that is tracked by imageability and concreteness ratings, it is relatively well-aligned with the hierarchical notion abstractness. Plausibly, the further a category moves in the direction of being superordinate in relation to others, the less likely it is that its members will share salient perceptual *or* theme-related features.

Thus, while we recognize this as a somewhat novel approach, we propose that when considering the overall abstractness level of a pictorial semantic memory task, one should take account of *both* dimensions—visual similarity and common setting—simultaneously. One reason that "abstracting away from common settings" is sometimes overlooked in assessments of abstract cognition is that existing measures of Concept Concreteness are linked to single words (as opposed to phrases), and, as Markman (1990) observes, we typically lack single words to refer to thematic relationships, or event types. This leaves any rating system linked to individual words unable to measure the sort of abstraction involved in abstracting away from common thematic relationships.

(Notably, Barsalou's (1983) category of *ad hoc* concepts includes many thematic categories; however, there are no standard concreteness or imageability ratings available for *ad hoc* concepts. Note, also, that we should not expect all thematic categories to be equally concrete or abstract. For those who frequently camp, we can expect the *ad hoc* category 'Things to take on a camping trip' to be relatively concrete, in the sense that its members are very closely associated due to their typically appearing together in a frequently experienced kind of event. By contrast, the *ad hoc* category 'things to take from one's home during a fire,' will, for most, not unite items that are frequently experienced together in a common setting, simply because such

events are rarely experienced. However, for someone who never goes camping but has the misfortune of experiencing many house fires, the category 'things to take from one's home during a fire' will have more strongly associated members—and linking them will require less abstraction-away from regularities in past experience—than 'things to take on a camping trip.')

Accordingly, in the norming study described below, we assigned Trial Concreteness ratings to individual semantic memory trials by summing a visual similarity score (relating to how visually similar the target and stimulus are, relative to the target and distractor items) with a common setting score (relating to how frequently the target and match are found together in a common setting, relative to the target and distractor items). In our view, this allows for a more complete measure of the degree of abstraction-from-past-experience required for properly answering a trial than either taken alone. When a trial has low Trial Concreteness ratings—and is therefore highly abstract—there will be few salient features shared by the target and match that can serve to alert the participant to the fact that they go together. By contrast, were one only to take account of visual similarity in generating Trial Concreteness ratings, relatively simple trials where, for instance, a fork is matched with a plate, could be rated as "highly abstract" due to the lack of visual similarity between the two. Intuitively, this is the wrong result. The frequent cooccurrence of forks and plates within a common theme or setting suggests, to the contrary, that trials linking them should be viewed as relatively concrete, in the sense that answering them requires relatively little abstraction from past experience (*modulo* the distractor items).

Finally, it bears noting that, in the context of a pictorial semantic memory task, trials will tend to become more difficult as Trial Concreteness decreases. This is because, as trials decrease in Trial Concreteness, there are fewer salient perceptual or theme-related features shared only by the target and match to serve as cues for making the correct selection. (The same is likely true as the category linking target and match becomes more superordinate with respect to others.) While it may be possible to dissociate difficulty from Trial Concreteness (see Section 4.1), we did not attempt to keep difficulty constant while manipulating Trial Concreteness in our stimuli. Instead, we used a mediation analysis (Section 3.4.3) to explore whether Trial Concreteness affects performance over and above the contribution of difficulty.

1.7 Assessing the relation of abstract thought and metacognition to language: aims and predictions

The trials developed and normed for Trial Concreteness in the norming study do not incorporate words as stimuli or require them as responses. This makes them suitable for use with a language-impaired population in assessing whether their language deficits lead to corresponding deficits in abstract thought. The main experiment presented here compares a group of people with aphasia ("PWA") to age-, education-, and gender-matched controls on their performance in selecting the correct matching image on the normed trials. Should PWA, despite their language impairment, show facility with categorizations on trials that are very low in Trial Concreteness, this would provide some evidence for the language-independence of abstract thought. Likewise, if PWA show greater difficulties than controls with stimuli low in Trial Concreteness compared to those high in Trial Concreteness, this would give reason to think that language is an important resource for abstract thought.

Our main prediction was that PWA would indeed show pronounced difficulties, compared to controls, on trials with low Trial Concreteness. That is, while lower performance overall is to be expected in the PWA population, due to their having experienced a stroke, we predicted that their relative difficulties would be proportionately more pronounced for trials low in Trial Concreteness—resulting in proportionately lower accuracy, longer response times, and lower confidence as the abstractness of a trial increases. This is because we shared with Lupyan and Mirman (2013) and Lupyan and Lewis (2019) the hypothesis that, in cases where only one or very few salient features serve to link two items, being able to produce the linguistic label for that feature will promote recognition of the link. Evidence for linguistic labels serving as a cognitive support for linking items in the absence of other salient perceptual or thematic connections is reported by a number of others, including Davis & Yee (2019), Sloutsky & Deng (2019), Louwerse (2018), and Vigliocco et al. (2018). Accordingly, we predicted that, for instance, those who cannot produce the words 'forecast' or 'predict' may have difficulty on grouping two things as both involved in forecasting or predicting, *provided that* the target and its match lack other salient properties that could alert one to the correct grouping (such as being commonly found together, or being visually similar). Such results would cohere with Sloutsky's (2010) thesis that "language provides learners with an additional set of cues that allow them to form more abstract distinctions" (p. 1248).

If language supports abstract thought in roughly these ways, we could expect abilities with related acts of abstraction to vary as a function of one's linguistic capacity. Therefore, we also predicted that the severity of language production impairments of the PWA—as measured by sub-components of the Western Aphasia Battery-R (Kertesz, 2006)—would correlate with their accuracy, confidence, and response times on the main task.

We did not, however, expect that our various measures of Concept Concreteness would correlate with PWA or control success at categorization to the same degree as with Trial Concreteness ratings, because we see Trial Concreteness as a finer-grained and more comprehensive measure of the kind of abstract thought required by the trials (in ways discussed above). Nevertheless, we thought it worthwhile to explore the relative effect of each. Note, however, that in order to compare the effects of Trial Concreteness and Concept Concreteness on performance, it is necessary to associate a single word with each of our (non-linguistic) trials. That way, the ratings for that word can provide the relevant Concept Concreteness for the trial. We call this the 'Linking Word' for each trial, which was were determined through a norming process described below. In some cases, however, associating a single word with the trial may be somewhat artificial (such as for thematic connections, as discussed above)—a point we return to in our discussion of the norming study (Section 2.5).

In addition to facilitating abstract thought, it has been proposed that language—sometimes in the form of "inner speech"—also supports increased levels of self-awareness (Carruthers, 2018; Morin, 2009), more accurate self-monitoring (Alderson-Day & Fernyhough, 2015; Jones & Fernyhough, 2007), and better and more comprehensive knowledge of one's own mental states (Bermudez, 2018; Carruthers, 2011; Clark, 1998; Langland-Hassan, 2014), including abstract concepts in particular (Borghi, 2020). An earlier study found preliminary evidence that on-line language use (in the form of inner speech) is an important resource for metacognitive self-assessments with respect to abstract categorizations (Langland-Hassan et al., 2017). To investigate this possible link, we included a second prompt on the main experimental stimuli querying participants' confidence in their responses. This metacognitive question aimed at assessing whether language has a role in increasing metacognitive accuracy distinct from any it may play in facilitating categorization itself, and whether this role is especially acute in trials with low Trial Concreteness.

We predicted that a disproportionate effect of low Trial Concreteness on PWA would show itself in this metacognitive aspect of the trials as well, with PWA showing proportionately lower confidence, and longer response times in reporting confidence, compared to controls, as Trial Concreteness ratings decreased. This would be in keeping with the results of Langland-Hassan et al. (2017), who, using a similar paradigm, found PWA to be impaired, relative to controls, in the accuracy of their metacognitive judgements with respect to whether they had correctly categorized items by an abstract category.

2. *Norming Study*

2.2 Methods

2.2.1 Materials:

430 full-color, high resolution images were used to create 86 trials of five images each, selected from targeted internet searches and used in accordance with principles of fair use (Brewer, 2008). No image was used more than once. Each trial consisted in a target image at the top of the screen with four choice images below it (See Figure 1). The correct match was determined in advance by the experimenters. In addition, the experimenters provisionally assigned a *Linking Word* to each trial that describes the association between the target and match. For example, in the trial displayed in Figure 1, the correct matching image was the weather forecaster, and the experimenter-assigned Linking Word was 'predict'. Trials were presented by means of a JavaScript application stored on a secure external web server.

2.2.2 Participants:

The population for these trials was drawn from Amazon's mechanical turk (mTurk) via the TurkPrime interface (TurkPrime, 2019; Behrend, et al., 2011; Buhrmeister, et al., 2011; Litman, et al., 2017). The final sample of norming participants consisted of 1000 US-based users of Amazon mTurk workers with a 95% or higher approval rating on past efforts (39% female/57.6% male; average age = 41.0 years, ±12.1). These participants were divided into three sub-groups, with each group completing a different kind of task, described below. All participants were presented with an online consent form and explanation of the nature of the study. Participants were compensated for their time at a level commensurate with the median pay rate for work requiring similar time commitments.

2.2.3 Procedure:

Trials were sorted into four banks to ensure that no mTurk worker would work for more than 30 minutes at a time. Each bank contained of either 21 or 22 trials (with a total of 86 trials in all banks). All trials were viewed by mTurk workers on their own computer screens and entered responses by clicking or typing. There were three types of task designed to generate norms for the following: (1) correct choices and Linking Words for each trial; (2) the frequency with which target and correct choice are found together in a common setting, relative to the other items; and (3) the visual similarity of the target and correct choice, relative to the other choices. Each task is now described in turn. Each mTurk participant completed only one type of task.

2.2.4 Stimulus Norming: Correct choice and linking word

mTurk participants (n=585) were presented with two practice trials followed by either 21 or 22 experimental trials (i.e., one of the four banks of trials) and asked to select for each which of the four choice images "best goes with" the target image at the top. After selecting the image with a mouse click, participants were asked to type the "one word that best describes how the image you chose relates to the top image" (see Figure 1).

As earlier noted, prior to mTurk norming, a Linking Word was preliminarily assigned to each trial to capture the intended linking concept (e.g., 'predict' for Figure 1). However, it was planned that in cases where the mTurk participants who answered a trial correctly gave an alternative word more frequently than the experimenter pre-assigned word (or a stem-variant thereof), the word given by mTurk participants would be reassigned as the trial's Linking Word. (However, if a mere stem-variant on the preliminary word—e.g. 'seeking' for 'seek'—was given more frequently, the preliminary word would be preserved).

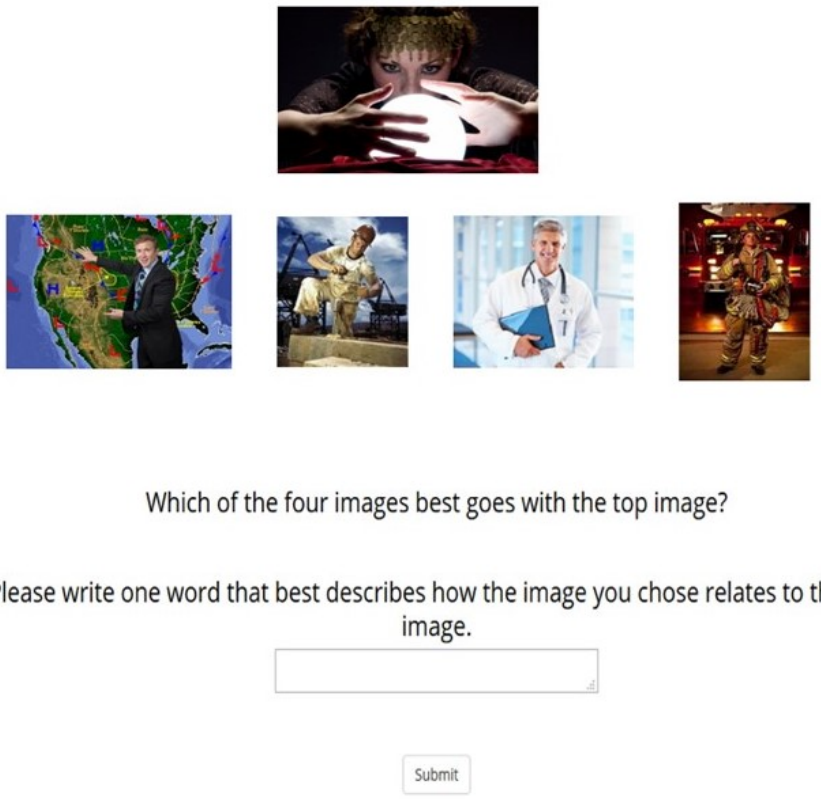

Figure 1 – Assessing the Linking Word during mTurk norming

2.2.5 Stimulus Norming: Common Setting (COM)

During the Common Setting (COM) norming procedure, n=197 mTurk participants were shown 86 trials corresponding to the 86 5-image sets used in the Concept and Linking Word norming procedure described above. Each participant saw only a subset of the trials (21 or 22 total), with

the trials split into four sets. For each trial, participants were asked to sequentially rate, on a seven-point scale, the frequency with which the target image and each of the four choice images is "found together in the same place or setting." The lowest rating (=1) was "almost never," while the highest (=7) was "almost always." Each trial required four distinct answers—one corresponding to each pair of images. A trial would begin with the target image and one of the choice images being outlined in green; the participant was asked to rate how often the two images marked in green are found together in a common setting (see Figure 2). Following a participant's response, the same target image and a different choice image would appear marked in green; the participant was again asked to rate how commonly the two images marked in green are found together in a common setting—and so on, for the remaining two images. The order in which specific choice images were highlighted with the target image, and ratings sought, was randomized by trial, as was the order in which the trials appeared.

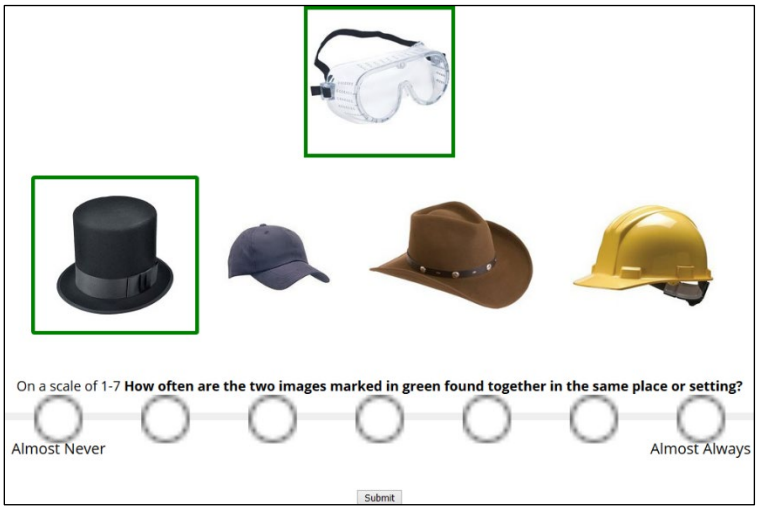

Figure 2 – Gathering Common Setting (COM) ratings during mTurk norming—at the stage where the target and top hat pair are queried.

2.2.6 Stimulus Norming: Visual Similarity (VIS)

The Visual Similarity (VIS) norming procedure was structurally identical to the Common Setting norming procedure with the following exceptions. Each of n=191 mTurk participants was shown

one of four subsets of the same 86 five-image trials (with each subset containing either 21 or 22 5-image trials). For each trial, participants were asked to sequentially rate, on a seven-point scale, "how similar looking" the target image was to each of the four choice images. The lowest rating (=1)was "not similar at all," while the highest (=7) was "almost exactly the same."

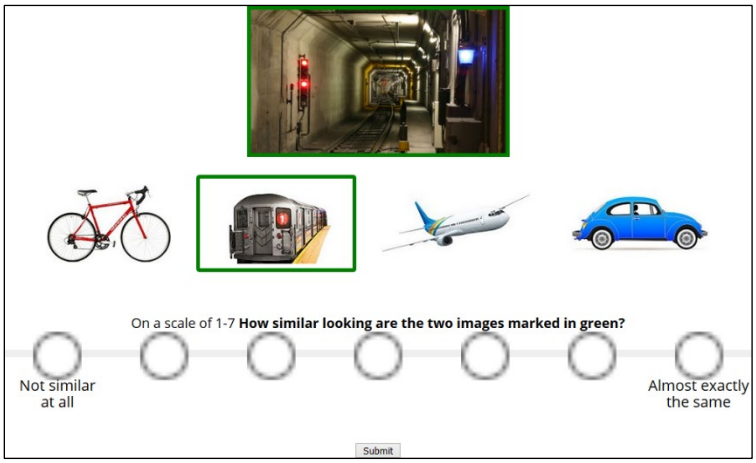

Figure 3 – Gathering Visual Similarity Ratings during mTurk norming

2.3 Data Analysis

2.3.1 Correct Choice and Linking Word:

Data from the Correct Choice and Linking Word trials were used to assess the rate at which participants selected the experimenter-intended "matching" image with the target image and to confirm that their reasons for selecting the match—as revealed by their word choice—cohered with experimenter intentions for the trial.

Based on data from Correct Choice and Linking Word norming, two trials were excluded from the final set to be used in the main experiment. On one of the trials ('even'), the average accuracy was only 26.5% and it was clear from written responses that participants rarely saw the intended semantic connection. On the other excluded trial ('police'), the correct response was given by only 51.0% of mTurk participants, with remaining responses nearly evenly distributed among the distractors. Further, on many "correct" answers for this trial, the words provided by mTurk participants to express the nature of the relationship between target and match did not fit the intended semantic association. It was therefore excluded from the set to be used in the main

experiment. Four additional trials ('dance', 'king', 'bath' and 'graduate') were also excluded from further analysis in the norming study, so that they could be used as practice trials in the main experiment. This left a total of 80 trials to be analyzed.

While mTurk participants sometimes provided alternative, semantically-related words from the experimenter-assigned words (e.g. 'forecast' instead of 'predict'), there were no trials where a particular word was given more frequently than the experimenter-assigned Linking Word. The resulting Linking Words for each trial are shown in Table 1.

2.3.2 Common Setting (COM):

The mean rating (from 1 to 7) given by mTurk participants to each pair of images during norming for Common Setting was calculated. Let *r* be the mean rating that participants gave to the target image and the correct choice image on a particular trial; and let *s* be the mean rating that participants gave to the highest-rated pair of target and choice images of the three remaining pairs for that trial. The formula for calculating the Common Setting rating for a trial was: $r - s$. In some cases, *s* was larger than *r,* resulting in a negative Common Setting rating for the trial. COM is intended as a measure of how many common-setting-related semantic associations there are between the target image and the correct choice image, *relative to the other available choices*.

2.3.3 Visual Similarity (VIS):

The mean rating (from 1 to 7) given by mTurk participants to each pair of images during the Visual Similarity norming procedure was calculated. Let *v* be the mean rating that participants gave to the target image and the correct choice image on a particular trial; and let *q* be the mean rating that participants gave to the highest-rated pair of target and choice images of the three remaining pairs for that trial. The formula for calculating the Visual Similarity rating for a trial was: $v - q$. In some cases, *q* was larger than *v,* resulting in a negative Visual Similarity rating for the trial. VIS is intended as a measure of how many visually perceptible similarities there are between the target image and the correct choice image, *relative to the other available choices*.

2.3.4 Trial Concreteness (Trial Concreteness):

Common Setting and Visual Setting scores were summed to arrive at a *Trial Concreteness* (TC) rating for each trial. Trials with low Trial Concreteness ratings were those where the target and correct choice items were (on average) judged neither to be found together commonly nor to be visually similar, relative to one or more of the other available choices for the trial. The 'cow' trial shown below (Figure 4a), where leather and milk are the target and correct choice, respectively, is an example of a low-Trial Concreteness (and thus highly abstract) trial; the 'predict' trial (Figure 1) is another. Trials with high Trial Concreteness ratings are those where the target and correct choice items were (on average) judged either to be found together commonly, or to be visually similar, or both, relative to all the other available choices for the trial. The 'subway' trial shown above (Figure 3), where a subway tunnel and a subway train are the target and correct choice, respectively, is an example of a high Trial Concreteness trial.

Images of each trial, together with their mTurk-normed Trial Concreteness ratings, Common Setting ratings, Visual Similarity ratings, and Linking Words are included, with the correct choice highlighted, in Appendix A.

2.4 Results

Data recorded from the three procedures resulted in a spectrum of COM, VIS and TC ratings for 80 total trials [dataset](Langland-Hassan, Faries, Gatyas, Dietz, & Richardson, 2021). COM scores ranged from 4.68 to -1.68 (mean=1.02, median=0.59, +/-1.46); VIS scores ranged from 2.9 to -2.65 (mean=0.67, median=0.72, +/-=0.82); Trial Concreteness scores ranged from 6.44 to -1.66 (mean=1.70, median=1.30, +/-=1.89). mTurk participant accuracy in providing the experimenter-intended answers ranged from 99% to 34% (mean=83%, median=89%, +/-=16%). See Table 1 for all scores per trial, with trials ordered from lowest Trial Concreteness rating to highest.

On the basis of the Linking Word established for each trial, concreteness (Brysbaert et al., 2014), imageability (Coltheart, 1981), Sensory Experience Ratings (Juhasz & Yap, 2013), and word frequency ratings (Log 10 transformed) (Brysbaert & New, 2009) were assigned to each trial, as shown in Table 1. In some cases, ratings from the respective databases were not available for a Linking Word. Such trials were excluded from subsequent correlation analyses.

Table 1: Trial Linking Words and Ratings, sorted by Trial Concreteness rating (low to high)

Trial Linking Word	Common Setting (COM)	Visual Similarity (VIS)	Trial Concreteness (TC)	mTurk Participant mean Accuracy	Imageability (Coltheart, 1981)	Concreteness (Brysbaert et al., 2014)	Sensory Experience Rating (Juhasz & Yap)	Word Frequency (Brysbaert & New, 2009)
memory	-1.68	0.02	-1.66	0.34	391	2.38	NA	3.39
identification	-0.64	-0.96	-1.6	0.57	345	3.4	NA	2.61
wine	1.38	-2.65	-1.27	0.94	624	4.79	3.82	3.49
hot	-0.96	-0.08	-1.04	0.42	551	4.31	3.82	3.99
scent	-0.76	0.05	-0.71	0.56	421	3.97	NA	2.48
sting	-0.04	-0.61	-0.65	0.67	553	4.41	3.45	2.56
potato	-1.38	0.92	-0.46	0.92	617	4.85	NA	2.76
jump	-0.15	-0.17	-0.32	0.65	506	4.52	NA	3.55
danger	-0.21	-0.04	-0.25	0.58	505	2.68	3.4	3.35
through	-0.21	-0.04	-0.25	0.52	320	2.9	1.5	4.45
jagged	-0.14	-0.04	-0.18	0.44	512	3.74	NA	1.48
leather	-0.92	0.78	-0.14	0.74	586	4.82	5.4	2.84
large	-0.18	0.07	-0.11	0.56	449	3.37	1.8	3.33
find	0	-0.1	-0.1	0.76	370	2.63	1.5	4.63
cow	-0.02	0.02	0	0.44	632	4.96	4.27	3.11
right	0.02	-0.02	0	0.44	372	3.47	2.91	5.31

rare	-0.16	0.24	0.08	0.68	439	1.96		1.82	3.04
kick	-0.04	0.32	0.28	0.58	551	4.33	NA		3.57
low	-0.1	0.4	0.3	0.75	378	3.34	NA		3.48
rough	-0.06	0.375	0.315	0.76	491	3.83		4.09	3.28
save	-0.32	0.72	0.4	0.81	365	2.42		2.1	3.92
turn	0.28	0.18	0.46	0.96	384	3.44		1.91	4.19
iron	0.2	0.34	0.54	0.89	561	4.59		3.55	2.96
small	0.32	0.25	0.57	0.82	447	3.22		1.8	3.80
blow	0.62	0	0.62	0.84	458	3.74		3.27	3.70
fight	0.54	0.12	0.66	0.61	543	4.2		3.3	4.01
alive	0.76	-0.04	0.72	0.77	426	3.14		3.82	3.90
whole	-0.23	0.96	0.73	0.86	377	3.25		2.64	4.29
crawl	0.04	0.72	0.76	0.77	488	4.27		2.7	2.79
kiss	0.46	0.36	0.82	0.95	633	4.48		4.5	3.79
security	0.66	0.16	0.82	0.93	391	2.82	NA		3.68
neck	-0.1	0.93	0.83	0.90	622	5		3.2	3.48
soft	0.44	0.41	0.85	0.92	476	3.88		4.6	3.21
predict	0.48	0.43	0.91	0.89	372	2		2.3	2.31
oil	0.22	0.7	0.92	0.75	573	4.93	NA		3.32
maple	0.66	0.35	1.01	0.65	511	4.46		4.1	2.22
hand	0.92	0.09	1.01	0.79	598	4.72	NA		4.15
wood	0.2	0.98	1.18	0.96	577	4.85		3	3.14

sand	0.52	0.71	1.23	0.93		603	5	3.4	3.02
alcohol	0.66	0.61	1.27	0.88		598	4.76	NA	2.93
slow	0.58	0.74	1.32	0.87		377	3.28	2	3.59
fast	0.6	0.72	1.32	0.96		411	3.32	2.6	3.85
wax	0.32	1.02	1.34	0.87		547	4.97	4.5	2.66
foot	0.04	1.32	1.36	0.95		597	4.9	2.9	3.52
sick	1.84	-0.36	1.48	0.87		456	2.97	3.82	3.93
race	0.22	1.29	1.51	0.97		457	3.59	3.1	3.50
plan	1.49	0.19	1.68	0.80		379	3.4	1.91	3.87
glass	0.15	1.56	1.71	0.93		585	4.82	2.73	3.49
think	0.91	0.94	1.85	0.98		384	2.41	2.64	5.14
bee	1.19	0.71	1.9	0.89		623	4.88	4.27	2.72
airport	2.32	-0.32	2	0.91	NA		4.87	4.7	3.29
myth	1.2	0.84	2.04	0.88		359	2.17	1.9	2.55
net	0.28	1.94	2.22	0.96		540	4.53	3.82	2.90
sports	1.52	0.8	2.32	0.95	NA		3.79	NA	3.15
cube	0.04	2.35	2.39	0.83		575	4.62	2.75	2.18
religion	1.52	0.89	2.41	0.93		434	1.71	NA	2.85
deliver	0.66	1.77	2.43	0.98		388	3.12	NA	3.16
up	1.76	0.8	2.56	0.96		391	3.83	2.91	5.27
ocean	2.58	0	2.58	0.86		623	4.86	6	3.19
luck	2	0.88	2.88	0.95		399	1.33	1.5	3.89

ninja	1.86	1.09	2.95	0.81	NA		4.28		5	2.12
wait	1.92	1.16	3.08	0.90		357	2.68		1.92	4.63
real	1.44	1.67	3.11	0.96		313	2.5	NA		4.35
dog	0.9	2.31	3.21	0.93		636	4.85		3.9	3.99
hat	1.18	2.14	3.32	0.96		562	4.88		2.91	3.52
pool	3.16	0.24	3.4	0.81		577	4.77	NA		3.38
baby	2.17	1.25	3.42	0.98		608	5		5.4	4.41
shoe	1.32	2.9	4.22	0.93		601	4.97		3.36	3.19
farm	3.48	0.78	4.26	0.94		560	4.59		4.25	3.19
safety	3.36	1.24	4.6	0.95		397	2.37	NA		3.22
rain	3.7	0.98	4.68	0.98		618	4.97	NA		3.40
paint	3.66	1.11	4.77	0.99		567	4.79		4.82	3.27
science	4.24	0.65	4.89	0.97		423	2.79		2.45	3.28
knight	3.44	1.53	4.97	0.95		608	4.79		3.64	3.14
justice	4.16	0.86	5.02	0.93		379	1.45		2.7	3.28
disco	3.38	1.76	5.14	0.97	NA		3.63		5.4	2.47
camp	3.66	1.59	5.25	0.98		588	4.35		2.5	3.42
pirate	4.1	1.18	5.28	0.95	NA		4.64		3	2.58
party	3.86	2.04	5.9	0.99		596	3.89		4.8	4.08
subway	4.68	1.76	6.44	0.98	NA		4.86		5.3	2.74

Correlations were assessed among multiple measures of abstractness and participant accuracy (Table 2). There were no significant correlations between Trial Concreteness (or its subcomponents of COM and VIS) and the Linking Word's imageability, concreteness, or word frequency; there were weak but significant correlations between Trial Concreteness and Sensory Experience Ratings ($r(59) = .26$, $p = 0.045$) and COM and Sensory Experience Ratings ($r(59) = .29$, $p = 0.026$). mTurk participant accuracy showed a strong correlation with Trial Concreteness ($r(78) = .67$, $p < 0.001$), Common Setting ($r(78) = .57$, $p < 0.001$), and Visual Similarity ($r(78) = .51$, $p < 0.001$). There were no significant correlations between mTurk participant accuracy and the Linking Word's imageability, concreteness, Sensory Experience Rating, or word frequency.

Table 2: Correlations among Abstractness Measures, Accuracy, and Linking Word Familiarity

Variable	n	M	SD	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1. Trial Concreteness (Trial Concreteness)	80	1.70	1.89	--							
2. Common Setting (COM)	80	1.02	1.46	.91**[4]	--						
3. Visual Similarity (VIS)	80	.67	.82	.68**[5]	.32**	--					
4. mTurk Participant Accuracy	80	.83	.16	.67**	.57**	.51**	--				
5. Imageability (Coltheart, 1981)	74	493.66	98.68	.17	.11	.17	.14	--			
6. Concreteness (Brysbaert et al, 2014)	80	3.85	1.01	.08	.03	.13	.08	.85**	--		

[4] COM (together with VIS) is a summed component of Trial Concreteness, rendering their correlation unsurprising.
[5] VIS (together with COM) is a summed component of Trial Concreteness, rendering their correlation unsurprising.

7. Sensory Experience Rating (Juhasz & Yap, 2013)	61	3.33	1.14	.26*	.29*	.08	.13	.70**	.62**	--	
8. Word Frequency (Brysbaert & New, 2009)	*80*	*3.40*	*.71*	*-.04*	*-.01*	*-.06*	*.06*	*-.30*	*-.23**	*-.32**	--

*$p < .05$. **$p < .01$.

2.5 Discussion

The aim of generating a set of non-verbal stimuli normed for Trial Concreteness was achieved, with the full range of trials covering a wide spectrum of Trial Concreteness values. Taking mean mTurk participant accuracy on each trial as a measure of its difficulty, the strong correlation between Trial Concreteness and accuracy suggests that Trial Concreteness is a key contributor to the difficulty of a trial. By contrast, the Linking Word's imageability, concreteness, word familiarity, and Sensory Experience Ratings appeared to have little effect on accuracy. This supports the general thesis that a trial with a low Trial Concreteness rating requires a sophisticated form of reasoning, whereby one must abstract-away from the most salient perceptual and theme-related features of a stimulus to appreciate a broader similarity.

The lack of strong (=$p<.01$) correlations among Trial Concreteness ratings and lexical measures of Concept Concreteness—including imageability scores, concreteness scores, and SER ratings—confirms our expectation that Trial Concreteness would provides a measure abstract thought distinct from that captured by Concept Concreteness. Of note, however, is a weak but a significant ($p<.05$) correlation between Sensory Experience Rating and Trial Concreteness. The correlation between SER ratings and Trial Concreteness may result from the fact that words with low SER ratings tend to mark categories whose members are not highly visually similar (increasing the likelihood of a low VIS score), while words with high SER ratings tend to mark categories whose members are visually similar (increasing the likelihood of a high VIS score). This trend will be defeated, however, to the extent that a visually dissimilar target and match are nevertheless judged to be frequently found together in a common setting (thereby raising TC rating), or a visually similar target and match occur in a setting where the distractor items are equally visually similar to the target (thereby lowering the TC rating).

To appreciate the distinctive cognitive challenge posed by trials with low Trial Concreteness scores, it is instructive to view the ways in which Trial Concreteness and Concept Concreteness ratings varied independently—at times cohering (either being both high or both low on a trial) and at times diverging (being high on in one rating and low on the other, on a trial). For instance, the 'predict' trial received a low Trial Concreteness rating (=0.91) and also relatively low Concept Concreteness ratings (viz., imageability, concreteness, Sensory

Experience ratings). This trial, seen in Figure 1, presented a target image of a fortune-teller looking over a crystal ball, with the four choice images being of a firefighter, doctor, weather forecaster, and construction worker. The correct choice was the weather forecaster, with the fortune-teller and weather forecaster being united under the concept of *predict*, or of *forecast*. The target and correct choice are not found together in a common setting, nor are they especially visually similar, compared to the other choices, accounting for its low Trial Concreteness rating. Of course, to recognize that it is predicting and forecasting that the two have in common, participants must exploit a concept such as *predict*, which is also low in Concept Concreteness (see Table 1).

However, some trials with low trial concreteness scores nevertheless have high concept concreteness scores. For example, on the 'cow' trial (Figure 4a), the target image was of a piece of leather, while the choice images were four beverages: orange juice, cola, wine, and milk. The correct choice was milk, with milk and leather both deriving from cows. In this case, the concept linking the target and choice images—*cow*—is paradigmatically concrete and easy to generate an image of; it has correspondingly high Concept Concreteness ratings (see Table 1). Yet the trial received a low Trial Concreteness rating (= 0.0), as participants did not judge the leather and milk to be commonly found together, or to be visually similar, relative to the other pictured items. Conversely, the 'justice' trial featured brass scales as the target image and four kinds of hammer as the choice images: a rubber mallet, a claw hammer, a gavel, and a sledgehammer (Figure 4b). The correct choice was the gavel, with the concepts of *justice* and *law* linking the scales to the gavel. In this case, the corresponding Linking Word for the trial ('justice') had relatively low Concept Concreteness ratings (see Table 1). However, participants assigned the trial a high Trial Concreteness rating (=5.02), likely due the fact that scales and gavels are commonly found together in representations of law and justice, and, in addition, bear some visual similarities (due to their part brass construction).

A reviewer raises the possibility that the concept linking leather and milk, in the 'cow' trial, is not in fact *cow* but rather the *ad hoc* concept THINGS THAT DERIVE FROM COWS, which would likely receive far lower concept concreteness ratings than cow itself. Similarly, the concept linking the gavel and scales in the 'justice' trial may not be JUSTICE but rather SYMBOLS OF JUSTICE, which would presumably receive higher concept concreteness ratings than JUSTICE,

as such symbols are easier to perceive than justice itself. We think these are legitimate and important possibilities to bear in mind. They provide a way of seeing how Trial Concreteness ratings could correlate more highly with (properly chosen) Concept Concreteness ratings than they did with the ratings we judged appropriate. Unfortunately, we do not know of a principled means by which to assess whether one or the other concept in contrasting pairs of this sort is the concept participants typically rely upon to arrive at correct answers. This is a difficulty inherent in non-verbal sematic memory tasks more generally. Nevertheless, we see it as a benefit of the Trial Concreteness measure that it provides an abstractness rating for each trial without requiring an answer to the vexed question of which concept or category—among many closely related ones—is required for recognizing the correct answer.

A limitation of Trial Concreteness ratings, however, is that they do not distinguish among potentially different types of abstract trial. It is plausible that abstract concepts constitute a heterogenous group, with different processing pathways relating to numbers, emotions, evaluative categories, and social categories, respectively (Borghi, Barca, Binkofski, & Tummolini, 2018). It is worth bearing in mind that such differences as may exist among trials are not tracked by our measure of Trial Concreteness.

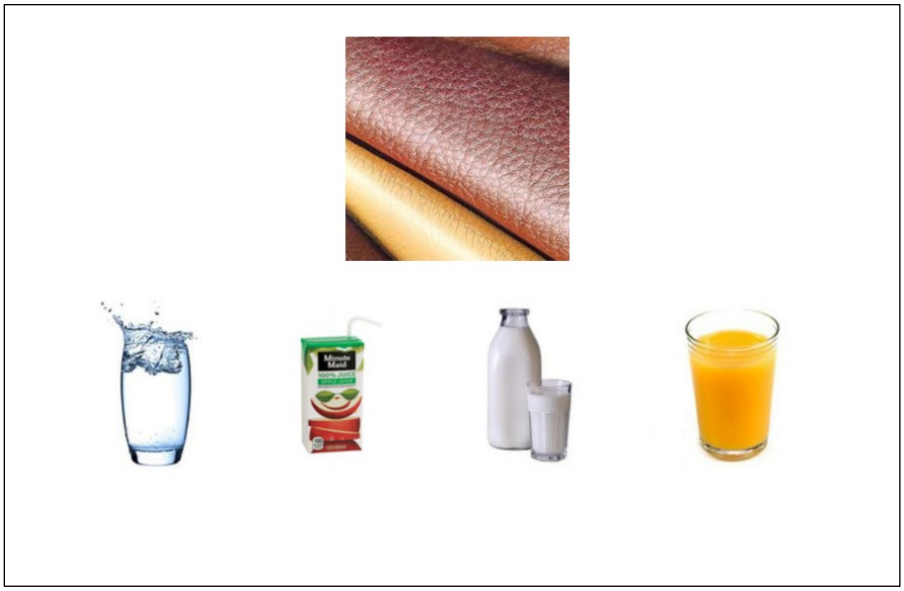

Figure 4a – The 'cow' trial

Figure 4b – The 'justice' trial

3. *Main Experiment*

Having created and normed a set of non-linguistic semantic memory trails for their degree of Trial Concreteness, we were then prepared to use them with a population of people with aphasia (and matched controls) to assess the effect of language deficits on abstract thought.

3.2 Methods

3.2.1 People with Aphasia (PWA) Group:

Twenty-three individuals with chronic post-stroke aphasia were recruited from two sources. The first was a database held at the University of Cincinnati Language Recovery & Communication Technology Lab. The second was a list of attendees of the Group Rehabilitation for Aphasia and Communication Effectiveness (GRACE) meetings, held at the Eardley Family Clinic for Speech, Language, and Hearing at Fontbonne University (St. Louis, MO). While almost all PWA have some impairment in both language production and language comprehension, some variants of aphasia (such as Broca's, transcortical motor, conduction, and anomic aphasia) pertain primarily to language production. PWA fitting that profile were selected, due to their ability to understand verbal task instructions, despite moderate to severe difficulties producing language.

The Western Aphasia Battery-Revised (WAB-R) (Kertesz, 2006) was used to confirm aphasia severity and type (Table 3). To ensure adequate comprehension of experimental tasks, participants were included only if their auditory-verbal comprehension score was ≥4. This yielded a group that included the following aphasia types: Broca's, transcortical motor, conduction, and anomic. To provide a general overview of cognitive functioning, the PWA also completed the Cognitive Linguistic Quick Task (CLQT) (Helm, 2003). Although the CLQT rates participants on five different cognitive measures, only the non-linguistic measures of *attention*, *executive function*, *visuospatial skills*, and *memory* were administered. (The linguistic measures were already assessed in the WAB-R.) For each of these measures, the participants were classified as *within normal limits*, *mildly impaired*, *moderately impaired*, or *severely impaired* (Table 3).

Three PWA from the original 23 were excluded from analysis for the following reasons. Participants 1010 and 1020 did not understand the main experimental task and were unable to follow task instructions. Participant 1012 was excluded due to a low attention score on the CLQT and being unable to follow instructions. These exclusions resulted in N = 20 participants with aphasia. (9M/11F, mean age 56 ± 9.1, age range 35-72, mean years of education 16 ± 1.7). Table 3 includes demographics for the PWA population.

Table 3: PWA participants

Participant Number	Age	Gender	MPO[1]	Education	Aphasia type[2]	Aphasia Quotient (AQ)	Attention[3]	Memory	Executive Function	Visuospatial Skill
1001	53	F	216	14	Conduction	79.4	161	63	27	98
1002	64	M	132	16	Broca's	62.9	158	43	28	91
1003	66	F	187	14	Conduction	67.6	172	54	26	88
1004	63	M	48	15	Broca's	53.4	131	41	21	74
1005	47	F	36	16	Anomic	87.0	163	63	29	100
1006	48	M	108	16	Conduction	77.4	190	64	30	94
1007	62	M	120	18	Trans. Motor	67.2	189	64	31	100
1008	72	F	228	16	Broca's	47.8	159	50	15	51
1009	67	M	48	18	Anomic	94.0	166	56	24	79
1011	44	M	48	14	Broca's	60.2	190	61	27	100
1013	56	F	372	12	Anomic	89.8	170	65	26	89
1014	53	F	24	16	Conduction	69.4	150	52	24	82
1015	56	F	60	16	Anomic	85.9	187	63	26	97
1016	52	F	18	16	Conduction	52.4	185	42	27	91
1017	35	F	24	14	Broca's	64.0	184	61	21	94
1018	55	M	240	18	Conduction	72.3	187	62	25	97
1019	69	M	18	18	Broca's	68.9	157	41	19	74

1021	53	F	72	18	Anomic	87.3	191	65	32	101
1022	58	F	48	14	Anomic	90.5	47	63	17	52
1023	55	M	60	18	Anomic	85.5	182	63	24	93

Note. [1]Months post-onset. [2]*Western Aphasia Battery-Revised* used to determine type and severity (Aphasia Quotient), total possible points = 100 (≤ 93 standard cutoff for formal diagnosis of aphasia). [3]*Cognitive-Linguistic Quick Test*: Ranges for participants up to 69 years old: Attention: Within Normal Limits=203-168, Mild=167-113, Moderate=112-38, Severe=37-0 (range reflects 12 point adjustment, due to deleting Story Retelling task from sum); Executive Functions: WNL: 35-19, Mild: 18-15, Moderate: 14-11, Severe: 10-0 (range reflects 5 point adjustment, due to deleting Generative Naming task from sum); Visuospatial Skills: WNL: 105-82, Mild: 81-52, Moderate: 51-42, Severe: 41-0. For participants over 70 years and older, the adjusted ranges are: Attention: WNL: 203-148, Mild: 147-88, Moderate: 87-28, Severe: 27-0; Executive Functions: WNL 35-14, Mild: 13-9, Moderate: 8-3, Severe 2-0; Visuospatial Skills: WNL 105-62, Mild: 61-37, Moderate: 36-22, Severe 21-0.

3.2.2 Control Participants:

Twenty adults with no reported history of brain injury, aphasia, mental illness, or substance abuse were recruited to participate as part of the control group (7M/13F, mean age 55 ± 9.4, age range 41-70, mean years of education 17 ± 1.8) (Table 4). In terms of age [$t(38) = -0.421$, $p = 0.676$], gender [$t(38) = 0.632$, $p = .531$](7M/13F), and education [$t(38) = 1.140$, $p = .261$], there were no significant differences between the PWA and the control participants.

Table 4 – Control Participants

Participant Number	Age	Gender	Education
C101	69	M	18
C102	66	F	16
C103	60	F	14
C104	70	F	14
C105	54	M	14
C106	55	F	18
C107	41	F	14
C108	50	F	14
C109	50	F	18
C110	55	F	16
C111	43	F	18
C112	41	F	18
C113	48	F	18
C114	56	M	14
C115	56	F	18
C116	58	F	18
C117	69	M	18
C118	50	M	18
C119	43	M	16

C120 69 M 18

3.2.3 Materials:

Trials were displayed on a 21" touchscreen computer using a proprietary JavaScript program. (The program runs in a standard web browser and is available from the corresponding author on request.) The 84 trials (80 experimental + 4 sample) were those generated during the norming study described above.

3.2.4 Procedure:

Participants sat at the computer and completed 84 trials. The first four trials were always the same and were used only for explaining the task; the data from these was not factored into the results. Thereafter, the 80 experimental trials generated and normed in the norming procedure were presented, with the order of trials randomized across participants. The left-to-right location of the choice images on each trial was also randomized across participants. In the first stage of each trial, participants were shown a target image with four choice images below it and were asked to select the choice image that best goes with the target image (as in Figures 4a-4b). Participants could answer either by touching the choice image via the touchscreen or by using a mouse, at their preference, but could not change their answers after selecting an image. The response time ($RT_{response,}$ also referred to hereafter as "accuracy response time") was recorded beginning with when the images appeared on the screen and ending with when a choice was made. After a choice image was selected, the following metacognitive question appeared below the target and choice images: "How confident are you that you selected the correct image?" (See Figures 5a-b). The following four responses were available, from left to right: 'I am guessing,' 'I am not confident,' 'I am quite confident', and 'I am very confident.'

For data analysis, responses were converted to numerical scores, with '1' corresponding to 'I am guessing' and '4' corresponding to 'I am very confident.' After selecting a confidence rating, participants could move directly to the next trial by selecting a box labelled 'Submit.' The response time in selecting the confidence rating ($RT_{confidence,}$ also referred to hereafter as "confidence response time") was recorded beginning with when the four response choices

appeared on the screen and ending with when the 'Submit' button was pressed. There were no time limits imposed on any of the selections.

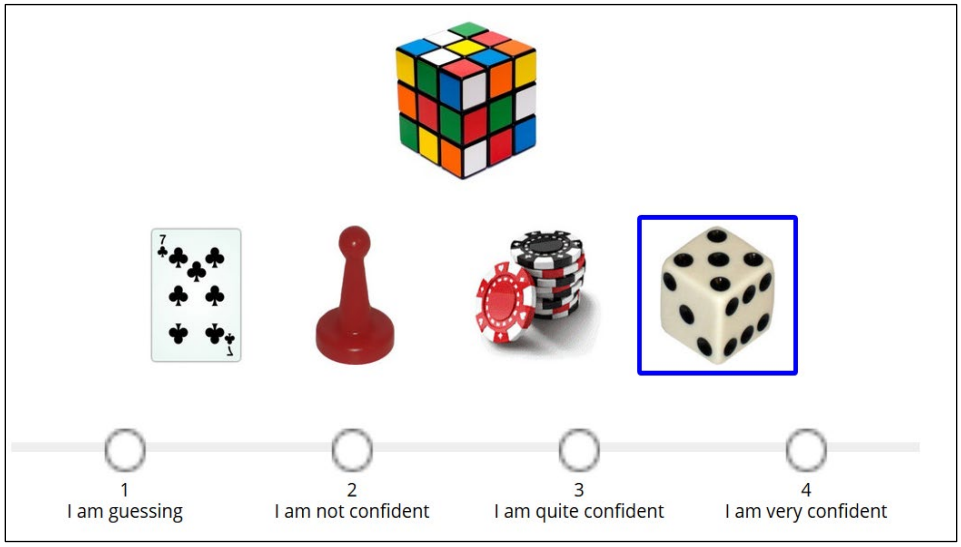

Figure 5a – The *'cube' trial, with correct answer highlighted, at the metacognitive stage*

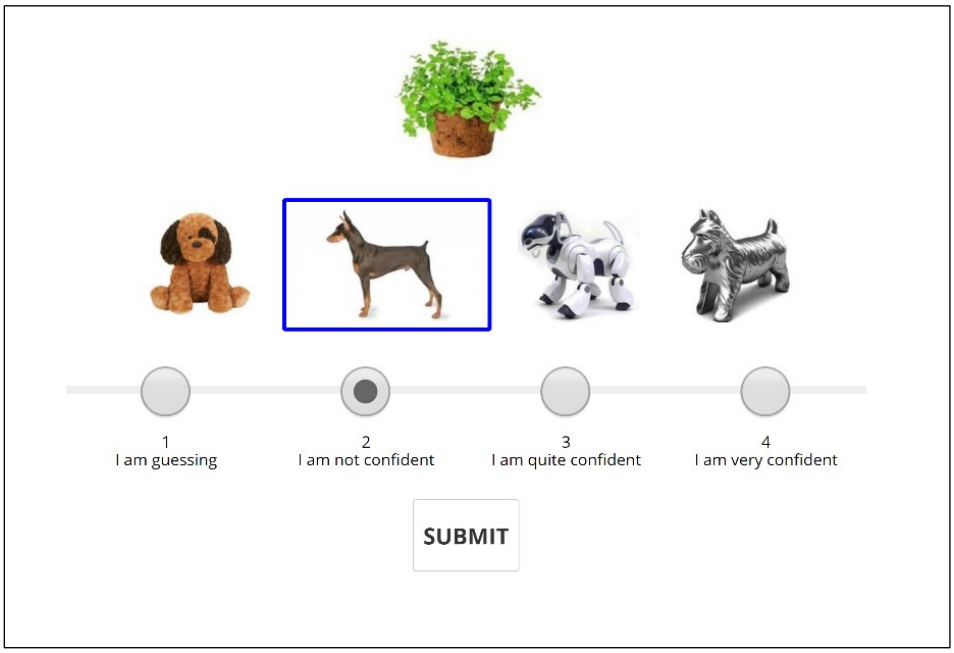

Figure 5b – The 'alive' trial, with correct answer highlighted, at the metacognitive stage

3.3 Data Analysis

Of particular interest was the effect of stimulus Trial Concreteness ("TC") rating on the response accuracy of PWA participants compared to control participants, as well as corresponding effects of Trial Concreteness rating with regard to response time ($RT_{response}$), response confidence, and confidence response time ($RT_{confidence}$). For the sake of simplicity, prior to analysis the Trial Concreteness score for the eighty stimulus sets was centered around zero, such that Trial Concreteness ranged from -3.36 to 4.74, with Trial Concreteness ratings < 0 corresponding to stimuli with a less than average Trial Concreteness rating and Trial Concreteness ratings > 0 corresponding to stimuli with a greater than average Trial Concreteness rating.

For illustrative purposes, we also group stimuli into four 20 stimulus Trial Concreteness groups. These groups corresponded to stimuli that had either *Low*, *Moderate-Low*, *Moderate-High* and *High* Trial Concreteness ratings (where e.g., Low = stimulus sets with the lowest 20 Trial Concreteness scores; and High = stimulus sets with the 20 highest Trial Concreteness ratings). As can be seen from an inspection of Figures 6 and 7, this provided a gross, yet discernable picture of the relationship between Trial Concreteness and the performance of PWA and control participants. For statistical analysis, however, rather than employing these Trial Concreteness categories or traditional ANOVA or regression techniques, we employed logistic and linear mixed effects models to test for effects of Trial Concreteness (as a continuous variable) and participant Group (i.e., PWA vs. control).

For each dependent variable, mixed effects modeling was conducted using STATA 16.0 (StataCorp, LP), with the fixed effects of Trial Concreteness (continuous), Group (PWA=1, Control=0), and TC×Group included in each model. Each analysis included 3,200 observations (80 trials × 20 Participants × 2 Groups). Linear models were tested using both (i) random intercepts for participant and (ii) random (by-subject) coefficients for participant (random intercepts) and TC (random slope), with unstructured covariance assumed for the latter. Logistic models were tested using structures (i) and (ii), as well as (iii) random intercepts for participant (by-subject) and stimulus (by-item). For all of the dependent measures assessed, however, the results for the fixed effects of TC, Group, and TC×Group for (i), (ii) and (iii) were the same. Thus, for the sake of brevity, only the result for models including random intercepts for participant are detailed here.

3.4 Results

3.4.1 PWA and Control Performance: Response Accuracy and Times.

Given that response accuracy was binary (i.e., 1 = correct response, 0= incorrect response), we analyzed response accuracy using a logistic mixed effects model. As expected, the resulting model revealed a significant effect of Group (β = -1.022, *SE* = 0.216, z = -4.73, $p < .001$, *CI* = [-1.446, -0.599]), with PWA participants exhibiting overall lower accuracy (M = 0.67 SD = 0.14) compared to controls (M = 0.82, SD=0.10). There was a significant effect of Trial Concreteness (β = 0.625, *SE* = 0.055, z = 11.41, $p < .001$, *CI* = [-0.517, -0.732]), with participant accuracy decreasing as the Trial Concreteness rating of a stimulus set decreased. There was no interaction between Group and Trial Concreteness (β = -0.049, *SE* = 0.069, z = -0.71, p = 0.479, *CI* = [-0.183, 0.086]), however, indicating that although the PWA exhibited less accuracy overall compared to controls, the relative reduction in accuracy between the PWA and controls remained relatively stable across Trial Concreteness ratings. Taking into account the logistic nature of accuracy (i.e., max performance = 1), the latter result can be discerned from an inspection of Figure 6 (*top*), where the plot on the right represents the linear mixed models marginal predicted mean accuracy scores for PWA and control participants for Trial Concreteness ratings between -4 and 5.

Consistent with the results observed for response accuracy, the linear mixed effects analysis of $RT_{response}$ also revealed significant effect of Group (β = 7.541, *SE* = 1.451, z = 5.20, $p < .001$, *CI* = [4.697, 10.386]), with PWA taking longer to respond than controls, as well as a significant effect of Trial Concreteness (β = -1.276, *SE* = 0.156, z = -8.16, $p < .001$, *CI* = [-1.583, -0.970]), with both PWA and controls taking longer to respond as Trial Concreteness decreased [dataset](Langland-Hassan et al., 2021). However, there was also a significant interaction between Group and Trial Concreteness (β = -1.256, *SE* = 0.221, z = -5.68, $p < .001$, *CI* = [-1.689, -0.832]), with decreases in the Trial Concreteness rating of stimuli having a greater relative effect on the response time of PWA compared to controls.

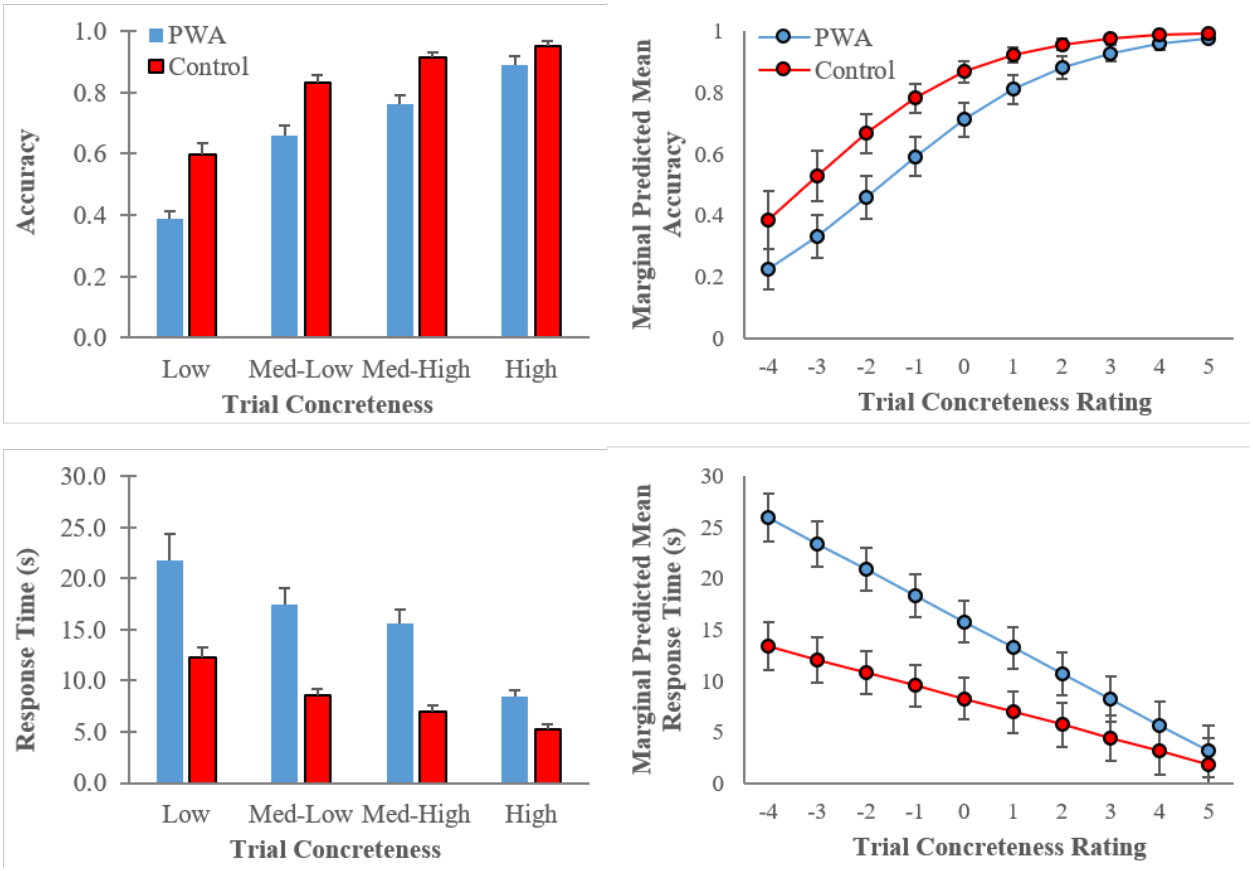

Figure 6. (*left*) Response accuracy and responses time for PWA and Controls as a function of stimulus sets with low, moderated-low, moderate-high and high Trial Concreteness ratings; error bars correspond to standard errors of the mean. (*right*) Marginal predicted means of the linear mixed effect model (random intercepts only) employed to determine main and interaction effects of Group (PWA vs Control) and Trial Concreteness; error bars representing 95% confidence intervals. See text for more details.

3.4.2 PWA and Control Performance: Response Confidence and Confidence Response Time.

Recall that response confidence was an ordinal response from 1 (not confident) to 4 (confident). Accordingly, we analyzed response confidence using a standard linear mixed model (with response confidence as a continuous variable), as well as an ordered logistics (logit) mixed effects model. The same fixed effects results were observed for both analyses; therefore, we only present the results of the linear mixed model here. As expected, there was a significant effect of Trial Concreteness (β = -0.202, SE = 0.013, z = 15.92, p < .001, CI = [-0.177, 0.227]), with

participants responding with less confidence as Trial Concreteness decreased (see Figure 7 top). However, there was no effect for Group (β = -0.218, *SE* = 0.173, *z* = -1.26, *p* = .206, *CI* = [-0.556, 0.120]), nor an interaction between Group and Trial Concreteness (β = -0.012, *SE* = 0.018, *z* = -0.237, *p* = .237, *CI* = [-0.014, 0.056]), with comparable confidence ratings for PWA and Controls across changes in Trial Concreteness.

The linear mixed model analysis of confidence response time only revealed a significant effect of Group (β = 0.914, *SE* = 0.303, *z* = 3.01, *p* = .003, *CI* = [0.319, 1.508]), with PWA taking longer to make confidence responses than controls. That is, neither the effect of Trial Concreteness (β = -0.027, *SE* = 0.032, *z* = -0.85, *p* = .394, *CI* = [-0.090, 0.036]), nor the interaction between Group and Trial Concreteness (β = -0.048, *SE* = 0.045, *z* = -1.05, *p* = .295, *CI* = [-.137, 0.041]) were significant, with the difference in confidence response time between PWA and Controls remaining constant across changes in Trial Concreteness (see Figure 7 bottom).

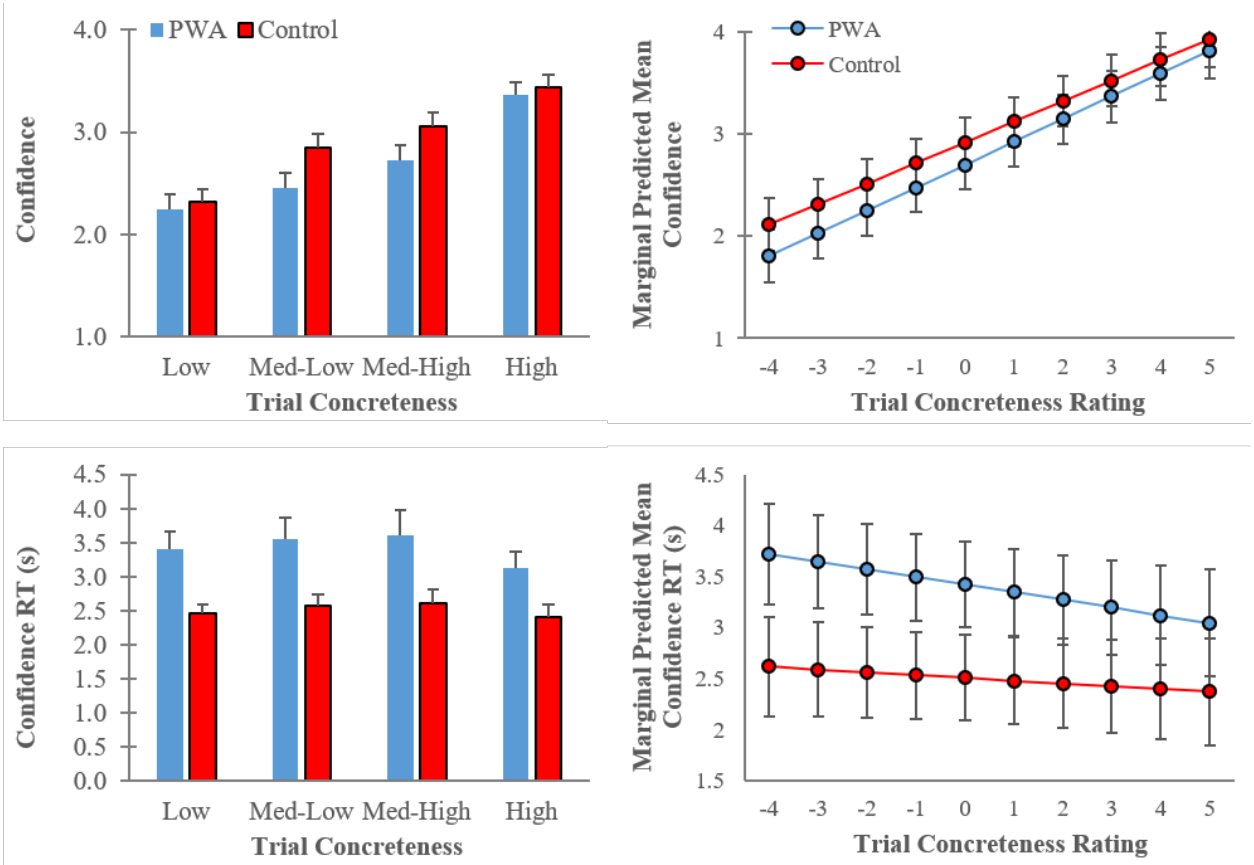

Figure 7. (*left*) Confidence response and confidence responses time (RT) for PWA and Controls as a function of stimulus sets with low, moderated-low, moderate-high and high Trial Concreteness ratings; error bars correspond to standard errors of the mean. (*right*) Marginal predicted means of the linear mixed effect model (random intercepts only) employed to determine main and interaction effects of Group (PWA vs Control) and Trial Concreteness; error bars representing 95% confidence intervals. The difference in slope (*lower right)* is not statistically significant. See text for more details.

3.4.3 Trial Concreteness and Stimulus Difficulty

Although it was assumed that Trial Concreteness was the key factor influencing the performance of PWA, it was also possible that stimuli that had lower Trial Concreteness ratings were simply more difficult and, thus, that the effects of Trial Concreteness on PWA performance was really an effect of stimulus difficulty. Indeed, if one assumes that the accuracy of controls

provides a pseudo measure of stimulus difficulty, then the significant positive correlation between the mean PWA and Control accuracy ($r = 0.773$, $p < 0.001$) for the 80 stimulus sets, as well as the significant positive correlations between stimulus Trial Concreteness rating and both mean PWA ($r = 0.727$, $p < 0.001$) and mean Control accuracy ($r = 0.597$, $p < 0.001$), suggested that this may have been the case. Accordingly, we employed a standard linear regression analysis to determine if the relationship between the stimulus Trial Concreteness rating and average PWA accuracy was mediated (Baron & Kenny, 1986) by the accuracy of controls (i.e., stimulus difficulty).[6]

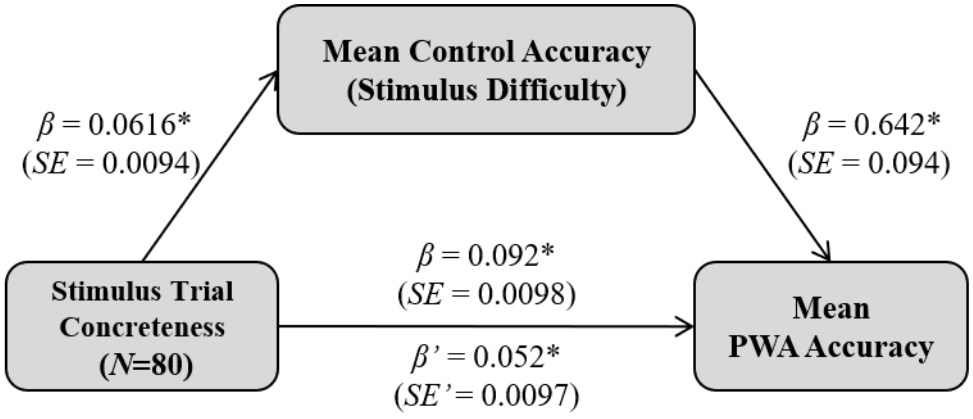

Figure 8. Stimulus difficulty (mean control accuracy) partially mediate the relationship between PWA accuracy and the Trial Concreteness score for the 80 stimulus sets.

As illustrated in Figure 8, we performed the mediation analysis by first conducting a linear regression between the Trial Concreteness rating of the 80 stimuli and mean PWA accuracy for each stimulus, with Trial Concreteness rating as the predictor variable. As noted above, this revealed a significant positive relationship between Trial Concreteness rating and PWA accuracy for the 80 stimuli employed in the study ($\beta = 0.092$, $SE = 0.0098$, t=9.35, $p < 0.001$, CI = [0.072, 0.111], overall $R^2 = 0.528$, $F(1,78) = 87.41$, $p < 0.001$). We then determined the relationship between Trial Concreteness rating and mean Control accuracy, which was also significant ($\beta = 0.0616$, $SE = 0.0094$, t=6.57, $p < 0.001$, CI = [0.043, 0.080], overall $R^2 = 0.356$, $F(1,78) = 43.18$, $p < 0.001$). We then added control accuracy as a second predictor with regard to PWA accuracy.

[6] A bootstrapping analysis (Preacher & Hayes, 2004) was also conducted to validate the mediation analysis presented here. This also demonstrated that the effects of Trial Concreteness on PWA performance were only partially mediated by control accuracy.

This analysis resulted in an overall R^2 = 0.707 (F(1,78) = 92.86, p < 0.001), with both Control accuracy (β = 0.642, SE = 0.094, t=6.85, p < 0.001, CI = [.456, 0.829]) and Trial Concreteness (β = 0.052, SE = 0.0097, t=5.36, p < 0.001, CI = 0.033, 0.071] operating as significant predictors. Thus, although the Sobel test for mediation was significant (S= 4.738, SE=0.0083, p < .001), the effects of Trial Concreteness on PWA performance were only partially mediated by control accuracy. In other words, Trial Concreteness did in fact influence the accuracy performance of PWA beyond the difficulty induced effects that Trial Concreteness appeared have on controls.

We also examined whether the other measures of stimulus Concept Concreteness described above—including concreteness (Brysbaert et al., 2014), imageability (Coltheart, 1981), Sensory Experience Rating (Juhasz & Yap, 2013) and word frequency (Brysbaert & New, 2009)—could have accounted for the apparent effects of Trial Concreteness on PWA and Control performance. However, as can be discerned from Table 5, none of these measures was significantly correlated (without correcting for multiple comparisons) with mean PWA or Control performance on any measure, with the exception of Sensory Experience Ratings, which were weakly correlated with PWA confidence (r(59) = .27, p=0.034) and Control confidence response time (r(59) = -.26, p = 0.045). By contrast, Trial Concreteness ratings were strongly correlated with PWA accuracy (r(78) = .73, p < 0.001), PWA accuracy response time (r(78) = -.70, p < 0.001), PWA confidence (r(78)= .74, p < 0.001), and with control accuracy(r(78) = .60, p < 0.001), control accuracy response time (r(78) = -.62, p < 0.001), and control confidence (r(78)= .66, p < 0.001).

Table 5. Correlations between different stimulus abstractness ratings and mean PWA and Control performance measures.

Stimulus Rating System	PWA				Control			
	Accuracy	Accuracy RT	Confidence	Confidence RT	Accuracy	Accuracy RT	Confidence	Confidence RT
Trial Concreteness (Trial Concreteness)	.73**	-.70**	.74**	-.22	.60**	-.62**	.66 **	-.16
Common Setting (COM)	.71**	-.71**	.74**	-.23*	.53**	-.56*	.58**	-.22*
Visual Similarity (VIS)	.41**	-.35**	.37**	-.09	.43**	-.43	.49**	.03
Imageability (Coltheart, 1981)	.12	-.17	.16	-.10	.2	-.15	.19	-.05
Concreteness (Brysbaert et al., 204)	.00	-.06	.06	-.01	.14	-.12	.16	.06
Sensory Experience	.25	-.24	.27*	-.13	.18	-.20	.22	-.26*

Rating (Juhasz & Yap, 2013) Word Frequency (Brysbaert & New, 2009)	.10	.00	.04	-.19	.05	-.02	-.03	-.02

* = $p < .05$, ** = $p < .01$

3.4.4 PWA Performance and WAB/CLQT measures

Finally, we examined whether PWA performance on the main task was related to their performance on various sub-components of the WAB-R and the CLQT. As detailed in Table 6, there were no significant correlations observed between the WAB-R measures and experimental accuracy, accuracy RT, or confidence. However, (and without correcting for multiple comparisons) significant relationships were observed between PWA confidence RT and the WAB measures of Spontaneous Speech ($r(18) = -.520$, $p = .019$), Naming ($r(18) = -.584$, $p = .007$) and overall Aphasia Quotient (AQ) ($r(18) = -.443$, $p = .050$), and also between PWA confidence RT and the CLQT measure of Memory ($r(18) = -.581$, $p = .007$), and between PWA confidence RT and Age ($r(18) = .484$, $p = .031$).

Table 6. *Correlations between PWA Performance and WAB/CLQT measures.*

			WAB					CLQT			
	Age	Years with Aphasia	Spontaneous Speech	AV Comprehension	Repetition	Naming	Aphasia Quotient (AQ)	Attention	Memory	Executive Function	Language
Accuracy	-.271	.362	.159	.214	-.104	.161	.109	.283	.396	.359	.232
Accuracy RT	.068	.171	.124	.004	.069	.016	.085	-.130	.113	.017	-.131
Confidence	-.098	-.044	-.116	-.026	-.045	-.123	-.103	-262	-.091	-.159	-.183
Confidence RT	.484*	-.251	-.520*	-.356	-.026	-.584*	-.443*	-.132	-.581*	-.088	-.207

* = $p \leq .05$, ** = $p < .01$

4. Discussion

4.1 The Effect of the Trial Concreteness Rating on Accuracy and Response Time:

As expected, the effect of Trial Concreteness was significant ($p < 0.01$) for both PWA and control populations on three key performance measures: accuracy, response time, and confidence (see Table 5), with lower Trial Concreteness ratings resulting in lower accuracy, lower confidence, and longer response times. PWA also showed significantly lower accuracy and longer response times than controls across the spectrum of Trial Concreteness scores. This was to be expected, due to the PWA group's history as stroke patients. Nevertheless, it bears noting that most of the PWA scored within normal limits on the non-verbal sections of the Cognitive Linguistic Quick Task (Table 3).

Echoing the results of the norming study, and again using $p < .01$ as a significance threshold, none of the three measures of Concept Concreteness (imageability, concreteness, and Sensory Experience Rating) showed correlations with PWA or control performance. This is no indictment of those constructs as they are applied in other contexts and in other kinds of tasks. However, it bolsters the motivation for countenancing Trial Concreteness as a distinct measure of abstract thought—one that takes into account the relativity of abstractness induced by variations in distractor items, and which factors in the need to abstract away from both thematic associations and perceptual similarities.

Our main predictions were that PWA would show proportionately lower accuracy, longer response times (for both accuracy and confidence responses), and lower metacognitive confidence than controls as trials became more abstract (in the Trial Concreteness sense). This would be the case if language serves as a crucial support for such abstract thought. These predictions were only partly confirmed. On the one hand, the response times of PWA were not simply longer, on average, than those of controls; they were also disproportionately longer as Trial Concreteness levels decreased, suggesting that the effect of low Trial Concreteness levels was especially pronounced on the language-impaired population. However, this conclusion must be tempered by the fact that a similar disproportionate impact of low Trial Concreteness rating was not observed with respect to PWA accuracy, PWA confidence, or PWA confidence response times. Nevertheless, response time on the main matching task is arguably a finer-grained measure of subjective difficulty than accuracy, reported confidence, or confidence response time.

Thus, the disproportionately longer response times of PWA, as Trial Concreteness ratings decreased, remains an important confirmation of the main predictions.

One might ask, however, whether it was not lower Trial Concreteness rating that accounted for the longer response times in PWA, but, rather, the overall difficulty of the trial. For it may not be surprising that a population recovering from stroke would have proportionately more difficulties with trials that are themselves more difficult. If it is indeed difficulty in general—and not Trial Concreteness rating, as such—that accounts for the result, the results would not warrant any conclusions concerning the relationship between language and abstract thought (as measured by Trial Concreteness ratings).

In response, and as remarked in the Introduction, Trial Concreteness cannot be completely disentangled from difficulty in general. However, Trial Concreteness level was not the *only* factor relevant to a trial's difficulty, as shown by the above mediation analysis. When control accuracy is taken as a proxy for trial difficulty, we see that Trial Concreteness has an effect on PWA performance independent of the effect imposed by difficulty in general. Put otherwise: while it is unsurprising and undisputable that lower Trial Concreteness trials are, in general, more difficult than high Trial Concreteness trials, this does not stand in the way of assessing the influence of Trial Concreteness on participant performance, separate from other forms of difficulty. The mediation analysis warrants treating Trial Concreteness as a distinct variable of interest—one that influences response time for both participant groups in the main experiment and that predicts proportionately longer response time in the PWA, compared to Controls.

It would nevertheless bolster future work to vary Trial Concreteness within the experimental stimuli to a greater degree while holding difficulty fixed. While doing so presents challenges, this should be possible. The main difficulties are, first, to generate relatively easy trials with low Trial Concreteness—where the correct choice jumps out at participants, despite the target image and its match not being judged visually similar, or to commonly occur together, relative to the target and distractor images. The second challenge is to create relatively difficult trials with high Trial Concreteness—where participants have difficulty identifying the match, despite the target and match being judged either highly visually similar, or to commonly occur together, relative to the target and distractor items. Increasing the number of stimuli meeting

these two conditions is an important project for future research. Further, while we maintain that summing visual similarity and common setting scores into a single Trial Concreteness score provides the most accurate measure of the overall abstractness of a stimuli, it is well worth systematically exploring the separate contributions of common setting and visual similarity ratings to stimulus difficulty in future work.

4.2 The relationship of language impairments to performance in PWA

Even if one were confident that the response times of PWA on lower Trial Concreteness trials were due to their greater difficulties with abstract thought, it is a further question whether those difficulties in turn derived specifically from the language impairments of the PWA. We predicted that the severity of language production impairments of the PWA—as measured by the WAB-R—would correlate with their accuracy, confidence, and response times, and thereby support the broader hypothesis of a link between language and abstract thought. For the most part, these predictions were not fulfilled, with the exception that confidence response times were significantly correlated with three linguistic scores from the WAB-R (Spontaneous Speech, Naming, and Aphasia Quotient) (see Table 6). Taken at face value, these correlations suggest that impaired language production influences one's metacognitive ability to assess one's own confidence level for a prior judgment, even if it does not similarly influence accuracy or confidence scores themselves.

Before considering why this would be the case, it is worth noting that the failure to establish broader correlations between PWA task performance and WAB-R language scores is not entirely surprising. Even if language serves as an important support for navigating low Trial Concreteness trials, it is not likely to be the *only* such support. Performance is likely to be influenced by a number of cognitive capacities in combination, including language abilities, executive function, attention, and so on. Untangling the effect of each of these on the other is difficult, though may have been possible with more in-depth cognitive testing. The low power of our analysis, with a sample of n=20 per group, makes it especially difficult to unearth such relationships through multiple regression. Nevertheless, it remains an indirect direct form of support for language's mediation of PWA performance that their response times were disproportionately affected by low Trial Concreteness, as impaired language was the most salient

cognitive difference between the two participant groups (with most PWA scoring within normal limits on all non-verbal sections of the CLQT).

Despite the PWA population's WAB-R scores not correlating with individual accuracy, accuracy response time, or confidence levels, several correlations—noted above—were found with respect to confidence response times (i.e., the time taken to indicate confidence level) and linguistic abilities. While inferences concerning the reasons for these correlations can only be tentative, it is interesting to note that they align with some previous results. Comparing a similar population of people with aphasia to controls, Langland-Hassan *et al.* (2017) found the reliability of metacognitive self-assessments to be lower for PWA than controls for abstract categorizations. (The notion of trial-abstractness used in that study was made via stipulation and not through the kind of norming process used here.) In the present study, confidence response time can be taken as a measure of difficulty in arriving at a settled judgment about one's own level of confidence. A number of philosophers and psychologists have proposed that covert online language—in the form of "inner speech"—provides an important resource for different forms of self-reflective, metacognitive thought (Bermudez, 2018; Carruthers, 2018; Clark, 1998; Jackendoff, 1996). Such proposals cohere with the fact that the linguistic impairments of individual PWA predicted greater difficulty in arriving at metacognitive determinations of confidence. Moreover, time taken to settle on a confidence response is arguably a more reliable index of actual confidence level than the explicit confidence score reported; response time is indeed often taken as proxy for confidence level (Ratcliff & Starns, 2009; Volkmann, 1934). The lack of similar correlations with respect to actual confidence scores thus need not be seen as undercutting this conclusion. In terms of the specific mechanism responsible for delayed assessments of confidence, it may be that the ability to generate a plausible linguistic label that describes the match—either out loud or through the use of inner speech—serves as a valuable cue that one has correctly identified the match. Thus, even if a participant has in fact selected the correct image as a match for the target, the participant may feel less confident that they have done so if their language production deficits prevent them from generating a word for the match, which might serve as a relevant cue for success. Such a cue, in the form of inner speech, would be especially valuable if, as some have proposed, the relevant abstract thought processes only occur consciously when in linguistic form (Bermudez, 2003; Carruthers, 2011; Jackendoff, 1996).

5. General Discussion

The measurement and comparison of abstract thought capacities can only be as precise as our understanding of what it is for thought to be "abstract." The predominant understanding of abstract thought has been in terms of concepts that are about things that are in some sense difficult to perceive, or that represent categories that are superordinate with respect to many others (Borghi et al., 2017; Boroditsky, 2001; Yee, 2019). Corresponding rating systems—such as for concreteness, imageability, and Sensory Experience Ratings—have been devised for rating concepts by their level of abstractness, so understood. However, these ratings are inevitably keyed to specific words associated with the concepts, making it difficult to assess abstract thought without necessarily, in the process, taxing linguistic capacities. Further, an important function of abstract thought is in abstracting away from the object associations that arise out of commonly experienced event types. Typically, there is no single word for such events, and thus no corresponding concreteness ratings. The present norming study and experiment developed and tested a new means for assessing abstract thought non-verbally—where the nature of abstract thought was understood in a way that relativizes the degree of abstract thought required to the nature of the matching task at hand, and where ability to abstract away from common setting associations was factored in an instance. It was shown that even tasks that involve linking two pictures by appeal to their shared connection to a concrete concept—such as *cow*—can require high degree of abstract thought. While much the same understanding of abstractness has predecessors in the notion of category "sparseness" (Sloutsky, 2010) and low versus high dimensionality (Lupyan & Mirman, 2013), the present study is the first to generate and test a spectrum of stimuli normed for different degrees of abstractness in this sense. It and other paradigms like it may serve as useful tools for the continued study of abstract thought.

The relationship of abstract thought—in both the Concept and Trial Concreteness senses—to language remains unsettled, both within psychology and in consideration of the present results. Results from people with aphasia and other language-impaired populations have long suggested that abstract thought—and complex reasoning in general—is not strictly dependent upon a concurrent ability to generate language (Bloom, 2000; Langland-Hassan et al., 2017; Siegal & Varley, 2006; Varley & Siegal, 2000). The traditional thesis that abstract thought is a distinct cognitive module from language production and comprehension (Fodor, 1975, 1983)

finds some support from the main experiment in the fact that the PWA, many of whom had severe productive aphasias, accurately selected the correct matching response at levels well above chance (where chance =25% correct) even for trials with low, and mid-low levels of Trial Concreteness (see Figure 6, top). Further, the fact that the WAB-R language scores of PWA did not correlate with accuracy in the selection task is also suggestive of abstract thought's not being entirely constrained by linguistic abilities.

On the other hand, there is also support within the present results for the idea that language plays a significant scaffolding or facilitating role in abstract thought, in keeping with the Words as Social Tools (WAT) thesis (Borghi et al., 2019; Borghi et al., 2017), the Language is an Embodied Neuroenhancement Scaffold (LENS) theory (Dove, 2019; Dove, Barca, Tummolini, & Borghi, 2020), and other related proposals (Boroditsky, 2001; Davis & Yee, 2019; Lupyan, 2009). This includes the fact that PWA response times were proportionately slower as a function of Trial Concreteness ratings and that PWA confidence response times correlated with their linguistic abilities. Further, it should be acknowledged that language could still have played a role in supporting PWA performance insofar as all PWA had normal language abilities at some point in their development, and may thus have acquired neural structures that are developmentally dependent upon language. Such structures could conceivably remain intact, and facilitate abstract thought, even if, after stroke, they are not sufficient for active language production.

While people with aphasia are intriguing participants for studies of the relation of thought to language, they also present special challenges. It is often difficult to recruit a large number of participants with the needed combination of production deficits and comprehension abilities, resulting in relatively low-powered analyses—as in the present case. In addition, each stroke affects somewhat distinct areas of the brain, leading to distinct patterns in patient deficits, and making it difficult to generalize across participants. The present study would benefit from corroboration in other populations, such as healthy participants in a dual task paradigm. In such a paradigm, a participant's linguistic capacities are taxed by a task such as verbal shadowing, while they concurrently complete a task thought to require abstract thought (Hermer-Vazquez, Spelke, & Katsnelson, 1999; Ratliff & Newcombe, 2008). Cross-cultural comparisons in performance on the stimuli developed in the norming study could also be used to assess whether

structural differences in languages map on to differences in abstract categorizations (as in, e.g., Boroditsky (2001, 2011, 2018)). Assessing changes in how abstract stimuli are matched over the course of child development—and whether such changes correlate with relevant changes in vocabulary—could offer another route to establishing corroborating evidence for language's involvement in abstract thought.

While it is natural to wish for a univocal verdict on the relation between abstract thought and language, the complexity of the results in the main experiment may simply reflect a complex underlying reality. It seems unlikely that the various abilities we are inclined to recognize as "abstract thought" will depend upon language in an all-or-nothing manner. Determining the precise form of support that language provides to abstract thought will require continued refinement of our understanding both of what it means for thought to be "abstract," and of the paradigms aimed at measuring it.

Supplementary Material: Data reported here and experimental stimuli used have been archived via the Open Science Framework and can be accessed using at the following web address: osf.io/gs2xn

Acknowledgments: We thank the John Templeton Foundation and the Cambridge New Directions in the Study of the Mind Project for funding support. We are also grateful to Amanda Eaton of Fontbonne University for assistance in recruiting participants with aphasia, and to Gary Lupyan for valuable advice concerning experimental design and data analysis.

References

Alderson-Day, B., & Fernyhough, C. (2015). Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychol Bull, 141*(5), 931-965. doi:10.1037/bul0000021

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts *Concepts and conceptual development: Ecological and intellectual factors in categorization.* (pp. 101-140). New York, NY, US: Cambridge University Press.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*(1), 617-645. doi:10.1146/annurev.psych.59.103006.093639

Begg, I., & Paivio, A. (1969). Concreteness and imagery in sentence meaning. *Journal of Verbal Learning and Verbal Behavior, 8*(6), 821-827. doi:https://doi.org/10.1016/S0022-5371(69)80049-6

Bermudez, J. L. (2003). *Thinking without Words*. Oxford: Oxford University Press.

Bermudez, J. L. (2018). Inner Speech, Determinacy, and Thinking Consciously About Thoughts. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices*. Oxford: Oxford University Press.

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *J Cogn Neurosci, 17*(6), 905-917.

Bloom, P. (2000). Language and thought: Does grammar makes us smart? *Current Biology, 10*(14), R516-R517. doi:https://doi.org/10.1016/S0960-9822(00)00582-0

Bolognesi, M., & Steen, G. (2018). Editors' Introduction: Abstract Concepts: Structure, Processing, and Modeling. *Topics in cognitive science, 10*(3), 490-500. doi:10.1111/tops.12354

Borghi, A. M. (2020). A Future of Words: Language and the Challenge of Abstract Concepts. *Journal of Cognition, 3*(1), 42-42. doi:10.5334/joc.134

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of life reviews, 29*, 120-153.

Borghi, A. M., Barca, L., Binkofski, F., & Tummolini, L. (2018). Varieties of abstract concepts: development, use and representation in the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1752), 20170121. doi:doi:10.1098/rstb.2017.0121

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychol Bull, 143*(3), 263-292. doi:10.1037/bul0000089

Boroditsky, L. (2001). Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time. *Cognitive Psychology, 43*(1), 1-22. doi:https://doi.org/10.1006/cogp.2001.0748

Boroditsky, L. (2011). How language shapes thought. *Scientific American, 304*(2), 62-65.

Boroditsky, L. (2018). Language and the Construction of Time through Space. *Trends in Neurosciences, 41*(10), 651-653. doi:https://doi.org/10.1016/j.tins.2018.08.004

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia, 38*(9), 1207-1215. doi:https://doi.org/10.1016/S0028-3932(00)00034-8

Brewer, M. (2008). Fair Use Evaluator. Retrieved from https://librarycopyright.net/resources/fairuse/index.php

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990. doi:10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Carruthers, P. (2018). The Causes and Contents of Inner Speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: Nature and Functions*. Oxford: Oxford University Press.

Casasanto, D., & Lupyan, G. (2015). All Concepts are Ad Hoc Concepts. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New directions in the study of concepts* (pp. 543-566). Cambridge: MIT Press.

Clark, A. (1998). Magic words: how language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and Thought: Interdisciplinary Themes* (pp. 162-183). Cambridge: Cambridge University Press.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General, 137*(1), 22-38. doi:10.1037/0096-3445.137.1.22

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers, 36*(3), 384-387.

Couchman, J. J., Coutinho, M. V. C., & Smith, J. D. (2010). Rules and resemblance: Their changing balance in the category learning of humans (Homo sapiens) and monkeys (Macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes, 36*(2), 172-183. doi:10.1037/a0016748

Davis, C. P., & Yee, E. (2019). Features, labels, space, and time: factors supporting taxonomic relationships in the anterior temporal lobe and thematic relationships in the angular gyrus. *Language, Cognition and Neuroscience, 34*(10), 1347-1357. doi:10.1080/23273798.2018.1479530

Dove, G. (2014). Thinking in words: language as an embodied medium of thought. *Topics in cognitive science, 6*(3), 371-389.

Dove, G. (2018). Language as a disruptive technology: abstract concepts, embodiment and the flexible mind. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1752), 20170135.

Dove, G. (2019). More than a scaffold: Language is a neuroenhancement. *Cognitive Neuropsychology*, 1-24.

Dove, G., Barca, L., Tummolini, L., & Borghi, A. M. (2020). Words have a weight: language as a source of inner grounding and flexibility in abstract concepts. *Psychological Research*. doi:10.1007/s00426-020-01438-6

Estes, Z., & Jones, L. L. (2009). Integrative priming occurs rapidly and uncontrollably during lexical processing. *Journal of Experimental Psychology: General, 138*(1), 112-130. doi:10.1037/a0014677

Fodor, J. A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay of Faculty Psychology*. Cambridge, MA: MIT Press.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. *Language acquisition and conceptual development, 3*, 215-256.

Gilead, M., Trope, Y., & Liberman, N. (forthcoming). Above and Beyond the Concrete: The Diverse Representational Substrates of the Predictive Brain. *Behavioral and Brain Sciences*.

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of Flexibility in Human Cognition: Dual-Task Studies of Space and Language. *Cognitive Psychology, 39*(1), 3-36. doi:https://doi.org/10.1006/cogp.1998.0713

Jackendoff, R. (1996). How language helps us think. *Pragmatics and Cognition, 4*(1), 1-34.

Jones, S., & Fernyhough, C. (2007). Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and Cognition, 16*(2), 391-399.

Juhasz, B. J., & Yap, M. J. (2013). Sensory experience ratings for over 5,000 mono- and disyllabic words. *Behavior Research Methods, 45*(1), 160-168. doi:10.3758/s13428-012-0242-9

Kalénine, S., Mirman, D., Middleton, E. L., & Buxbaum, L. J. (2012). Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts. *J Exp Psychol Learn Mem Cogn, 38*(5), 1274-1295. doi:10.1037/a0027626

Kalénine, S., Peyrin, C., Pichat, C., Segebarth, C., Bonthoux, F., & Baciu, M. (2009). The sensory-motor specificity of taxonomic and thematic conceptual relations: a behavioral and fMRI study. *NeuroImage, 44*(3), 1152-1162. doi:10.1016/j.neuroimage.2008.09.043

Kertesz, A. (2006). Western Aphasia Battery-Revised (WAB-R). Austin, TX.

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 804-823. doi:10.1037/0278-7393.20.4.804

Langland-Hassan, P. (2014). Inner Speech and Metacognition: In Search of a Connection. *Mind and Language, 29*(5), 511-533.

Langland-Hassan, P., Faries, F. R., Gatyas, M., Dietz, A., & Richardson, M. J. (2021). *Data Set: Assessing Abstract Thought and its Relation to Language with a New Noverbal Paradigm: Evidence from Aphasia*. Retrieved from: osf.io/gs2xn

Langland-Hassan, P., Gauker, C., Richardson, M. J., Dietz, A., & Faries, F. R. (2017). Metacognitive deficits in categorization tasks in a population with impaired inner speech. *Acta Psychologica, 181*, 62-74. doi:https://doi.org/10.1016/j.actpsy.2017.10.004

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *J Exp Psychol Gen, 130*(1), 3-28. doi:10.1037/0096-3445.130.1.3

Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. *Topics in cognitive science, 10*(3), 573-589. doi:https://doi.org/10.1111/tops.12349

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review, 16*(4), 711-718.

Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in cognitive science, 8*(2), 408-424.
Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience, 34*(10), 1319-1337. doi:10.1080/23273798.2017.1404114
Lupyan, G., & Mirman, D. (2013). Linking language and categorization: Evidence from aphasia. *Cortex, 49*, 1187-1194.
Markman, E. M. (1981). Two different principles of conceptual organization. In M. E. Lamb & A. L. Brown (Eds.), *Advances in Developmental Psychology* (pp. 199-236). Hillsdale, NJ: Erlsbaum.
Markman, E. M. (1990). Constraints Children Place on Word Meanings. *Cognitive Science, 14*(1), 57-77. doi:https://doi.org/10.1207/s15516709cog1401_4
Medler, D. A., & Binder, J. R. (2005). MCWord: An On-Line Orthographic Database of the English Language. Retrieved from http://www.neuro.mcw.edu/mcword/
Minda, J. P., Desroches, A. S., & Church, B. A. (2008). *Learning rule-described and non-rule-described categories: A comparison of children and adults.* (34 doi:10.1037/a0013355), American Psychological Association, US.
Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General, 141*(4), 601-609. doi:10.1037/a0026451
Morin, A. (2009). Self-awareness deficits following loss of inner speech: Dr. Jill Bolte Taylor's case study☆. *Consciousness and Cognition, 18*(2), 524-529.
Paivio, A. (1971). *Imagery and verbal processes*: Psychology Press.
Perry, L. K., & Lupyan, G. (2017). Recognising a zebra from its stripes and the stripes from "zebra": the role of verbal labels in selecting category relevant information. *Language, Cognition and Neuroscience, 32*(8), 925-943. doi:10.1080/23273798.2016.1154974
Pinker, S. (1994). *The Language Instinct*. New York: William Morrow and Company.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717-731. doi:10.3758/BF03206553
Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review, 116*(1), 59-83. doi:10.1037/a0014086
Ratliff, K. R., & Newcombe, N. S. (2008). Is language necessary for human spatial reorientation? Reconsidering evidence from dual task paradigms. *Cognitive Psychology, 56*(2), 142-163. doi:https://doi.org/10.1016/j.cogpsych.2007.06.002
Reilly, J., Westbury, C., Kean, J., & Peelle, J. E. (2012). Arbitrary Symbolism in Natural Language Revisited: When Word Forms Carry Meaning. *PLoS ONE, 7*(8), e42286. doi:10.1371/journal.pone.0042286
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.
Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(1), 82-102. doi:10.1037/0278-7393.9.1.82
Siegal, M., & Varley, R. (2006). Aphasia, language, and theory of mind. *Social Neuroscience, 1*(3-4), 167-174. doi:10.1080/17470910600985597

Sloutsky, V. M. (2010). From Perceptual Categories to Concepts: What Develops? *Cognitive Science, 34*(7), 1244-1286. doi:10.1111/j.1551-6709.2010.01129.x
Sloutsky, V. M., & Deng, W. (2019). Categories, concepts, and conceptual development. *Language, Cognition and Neuroscience, 34*(10), 1284-1297. doi:10.1080/23273798.2017.1391398
Thibodeau, P. H., Hendricks, R. K., & Boroditsky, L. (2017). How Linguistic Metaphor Scaffolds Reasoning. *Trends in cognitive sciences, 21*(11), 852-863. doi:https://doi.org/10.1016/j.tics.2017.07.001
Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Curr Biol, 10*(12), 723-726. doi:10.1016/s0960-9822(00)00538-8
Vigliocco, G., Ponari, M., & Norbury, C. (2018). Learning and Processing Abstract Words and Concepts: Insights From Typical and Atypical Development. *Topics in cognitive science, 10*(3), 533-549. doi:https://doi.org/10.1111/tops.12347
Volkmann, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychol Bull, 31*(9), 672-673.
Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *31*(10), 1459-1468. doi:10.1002/hbm.20950
Wilson-Mendenhall, C. D., Simmons, W. K., Martin, A., & Barsalou, L. W. (2013). Contextual Processing of Abstract Concepts Reveals Neural Representations of Nonlinguistic Semantic Content. *Journal of Cognitive Neuroscience, 25*(6), 920-935. doi:10.1162/jocn_a_00361 %M 23363408
Yee, E. (2019). Abstraction and concepts: when, how, where, what and why? *Language, Cognition and Neuroscience, 34*(10), 1257-1265. doi:10.1080/23273798.2019.1660797
Yee, E., Jones, M. N., & McRae, K. (2018). Semantic memory. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, 3*, 1-38.
Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review, 23*(4), 1015-1027. doi:10.3758/s13423-015-0948-7