WHY ARE YOU TALKING TO YOURSELF? THE EPISTEMIC ROLE OF INNER SPEECH IN REASONING[i]

Wade Munroe

Department of Philosophy and the Weinberg Institute for Cognitive Science, University of Michigan

Abstract

People frequently report that, at times, their thought has a vocal character. Thinking commonly appears to be accompanied or constituted by silently 'talking' to oneself in inner speech. In this paper, we explore the specifically epistemic role of inner speech in conscious reasoning. A plausible position—but one I argue is ultimately wrong—is that inner speech plays a *solely* facilitative role that is exhausted by (i) serving as the vehicle of representation for conscious reasoning, and/or (ii) allowing one to focus on certain types of objects or relations, e.g., causal relations, abstracta, counterfactuals, etc., or to consciously entertain structured propositional contents that it would be hard (or impossible) to focus on or entertain with representations in other (e.g., imagistic) formats. According to this position, inner speech doesn't figure as a justificatory element in our reasoning or as the partial epistemic basis of our conclusions—it merely facilitates reasoning through (i) and/or (ii). In contrast to the view that inner speech is a *mere* facilitator, I establish that (outside of potentially playing roles (i) and/or (ii)) the language we use itself serves as a crucial source of information in reasoning. In other words, we reason from propositions about the language we use in inner speech as opposed to exclusively reasoning from the semantic contents of the speech. My conclusion follows from how we use language as a cognitive tool to keep track of information, e.g., the contents of premises, lemmas, previous reasoning results, etc., in reasoning.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/nous.12385.

> "It seems to me that the soul when it thinks is simply carrying on a discussion in which it asks itself questions and answers them itself, affirms and denies."
>
> —Plato (Theatetus 190a, Cooper & Hutchinson, 1997)

People frequently report that, at times, their thought has a vocal character. Thinking commonly appears to be accompanied or constituted by silently 'talking' to oneself in inner speech (Heavey & Hurlburt, 2008).[ii] Although there is debate about how best to define inner speech, I join Alderson-Day and Fernyhough in, roughly, fixing the analysandum as "the subjective experience of language in the absence of overt and audible articulation" (2015, p. 1).[iii] It's been argued that inner speech plays an important role in (inter alia) problem solving, task-switching, reading, and language comprehension and acquisition (Alderson-Day & Fernyhough, 2015; Perrone-Bertolotti, Rapin, Lachaux, Baciu, & Lœvenbruck, 2014). In this paper, we explore the specifically epistemic role of inner speech in conscious reasoning.[iv] A plausible position—but one I argue is ultimately wrong—is that inner speech plays a *solely* facilitative role that is exhausted by (i) serving as the vehicle of representation for conscious reasoning, and/or (ii) allowing one to focus on certain types of objects or relations, e.g., causal relations, abstracta, counterfactuals, etc., or to consciously entertain structured propositional contents that it would be hard (or impossible) to focus on or entertain with representations in other (e.g., imagistic) formats (cf. Bermúdez, 2003; Jackendoff, 1996).[v] According to this position, inner speech doesn't figure as a justificatory element in our reasoning or as the partial epistemic basis of our conclusions—it merely facilitates reasoning through (i) and/or (ii).

In contrast to the view that inner speech is a *mere* facilitator, I establish that (outside of potentially playing roles (i) and/or (ii)) the language we use itself serves as a crucial source of information in reasoning. In other words, we reason from propositions about the language we use in inner speech (or respond to quasi-perceptual seemings regarding the language, as explained in subsection 2.1) as opposed to exclusively reasoning from the semantic contents of the speech.[vi] My conclusion follows from how we use language as a cognitive tool to keep track of information, e.g., the contents of premises, lemmas, previous reasoning results, etc., in reasoning. As Tyler Burge rightly notes, "[a]ny reasoning in time must rely on memory…In a deduction [for example], reasoning processes' working properly depends on memory's preserving the results of previous reasoning" (Burge, 1993, p. 463). As I establish, keeping information in mind (e.g., previous reasoning results) with the aid of inner speech in working memory requires reasoning from the language itself in virtue of the fact that the inner speech consists of bare phonological representations that need to be semantically interpreted. In addition, I argue that understanding the epistemic role of inner speech as a justificatory element in reasoning requires that we distinguish first-order and metacognitive/meta-reasoning processes—more colloquially, 'thinking about thinking'(Ackerman & Thompson, 2017)—that serve to guide the execution of decision procedures at the first-order level. As I demonstrate, at least one epistemic function of inner speech is to serve as a source of information about our reasoning itself.[vii]

Before proceeding to the argument, a note about the empirical literature I cite is in order. I draw extensively from literature in cognitive psychology on the role of language in arithmetic calculation. As I establish in section 1, we have to use a learned generative number representation system (hereafter, GRS)—for example, a language with compositional rules for combining lexical primitives to generate expressions of successive integers—to engage in exact calculation. In other words, we do not think purely in number unmediated by inner speech or some other form of numeric representation. By focusing on arithmetic calculation I avoid discussions that are beyond the scope of this paper regarding whether

conscious thought must occur through inner speech or some other form of mental imagery (Carruthers, 2014; Prinz, 2012) or whether it's possible that we engage in (what's commonly known as) unsymbolized thought, unmediated by any mental imagery (Hurlburt & Akhter, 2008). However, what I say about reasoning in the arithmetic domain will apply, mutatis mutandis, to other areas as well. There is nothing special about our use of inner speech to keep track of intermediate reasoning results or other information in arithmetic calculation that would make it illegitimate to generalize my conclusion to functionally identical uses of inner speech in reasoning in other domains.

In addition, my arguments rely on a model of working memory advanced by Alan Baddeley and colleagues (2007, 2017a). Working memory is "a mental workspace that is involved in controlling, regulating, and actively maintaining relevant information" (Raghubar, Barnes, & Hecht, 2010, p. 110) in conscious reasoning. Baddeley's model is assumed in much of the cognitive psychological literature I cite and has ample empirical support. However, it's beyond the scope of this paper to defend the model against contemporary alternatives. If you, the reader, are familiar with the literature on working memory and dubious of Baddeley's model, I encourage you to read my arguments as an exploration of some of the model's epistemic implications (on the assumption the model is accurate).

The paper is structured as follows. In section 1, I argue that, given our cognitive architecture, we have to use a GRS to engage in exact calculation. In section 2, I establish that keeping lemmas or intermediate reasoning results in mind in an extended process of reasoning with the aid of inner speech and working memory requires reasoning from the language used. In section 3, I discuss how we ought to think of our use of language in reasoning in relation to the justificatory status of the conclusions we draw. In addition, I argue that understanding the epistemic role of language requires that we distinguish first-order and metacognitive processes in reasoning.

1. GRS AND ARITHMETIC CALCULATION

There is a growing consensus that humans and certain non-human animals share an innate number sense that includes (i) the ability to represent exact quantities up to about three and (ii) an analogue representation system for dealing with approximate quantity (Feigenson, Dehaene, & Spelke, 2004). However, the innate number sense isn't sufficient to engage in exact calculation (at least when dealing with quantities larger than three) without the use of a learned GRS.

Several studies of relatively isolated groups of individuals in the Amazon, namely, the Pirahã and Munduruku, provide strong evidence for the claim that we are unable to perform exact calculation without a GRS. The Pirahã and Munduruku use languages with very limited number vocabularies—if the languages have words for exact quantities at all—and cannot reliably perform elementary calculations or remember the cardinality of sets larger than three, despite possessing similar abilities to reason with approximate quantity as aged matched individuals who utilize a GRS (Michael C Frank, Everett, Fedorenko, & Gibson, 2008; Gordon, 2004; Pica, Lemer, Izard, & Dehaene, 2004).[viii]

Similar studies have been performed with congenitally deaf homesigners raised in numerate communities. Homesigners have not acquired a conventional signed language, typically, because of a lack of access to educational resources, and thus have to develop a gestural idiolect (called a home sign) as a means of communicating. The homesigners studied,

> appreciate that a set of objects has an exact cardinal value, and that a unique gesture communicates that value…However, despite the fact that homesigners use their fingers to communicate about number, they do not consistently produce gestures that accurately represent the cardinal values of sets containing more than three items. (Spaepen, Coppola, Spelke, Carey, & Goldin-Meadow, 2011, p. 3167)

If we are not taught to use a GRS—even if we live in a numerate community—we will lack the means to reliably engage in exact calculation.

Verbal shadowing—that is, repeating meaningful speech directly after hearing it, thus loading psycholinguistic processes responsible for speech planning, production, and comprehension—severely impairs exact calculation in numerate individuals. During verbal shadowing, numerate individuals perform just like the Pirahã on the same mathematical tasks used to test the Pirahã's numeracy (Micheal C Frank, Fedorenko, & Gibson, 2008). So, preventing an individual from using inner speech and a GRS in conscious reasoning by burdening the relevant psycholinguistic processes responsible (which we discuss in the next section) results in the individual being unable to reliably engage in exact calculation.

In sum, without a learned GRS we lack the cognitive means to engage in exact calculation. In addition, if we are prevented from using a GRS in conscious reasoning with the aid of, e.g., inner speech, we will not be able to reliably keep track of quantities larger than three—a skill crucial for, say, carrying values and storing partial products in determining the product of two integers through commonly used calculation algorithms. Learning a GRS does not merely facilitate the acquisition of mathematical concepts that can be subsequently deployed without the use of the GRS. We have to consciously use the GRS as a representational resource to engage in exact calculation by, e.g., using inner speech to token linguistic representations of numbers.

Although using a GRS in conscious reasoning through, e.g., inner speech, appears to be psychologically necessary in order for creatures like us to engage in exact calculation, we have yet to establish my conclusion that inner speech itself serves as a source of information from which we reason. It may be that using inner speech and a language with a generative number vocabulary plays a purely facilitative role in conscious mathematical reasoning by allowing us to focus attention on exact quantities through tokening linguistic representations of those quantities.

2. Working Memory and Inner Speech

We can distinguish reasoning *through* an utterance in inner speech from thinking *about* the language used (or imaged).[ix] For instance, say I hear someone assert, "está lloviendo," (in English, "it's raining"). I know nothing of Spanish and am not paying attention to contextual cues to decipher the utterance meaning. I may parrot the utterance in auditory imagery, but I am not thinking *through* the language as a means of occurrently judging that it is raining. This bit of auditory imagery cannot enter into a conscious practical inference that—along with my desire to refrain from getting wet and a host of background beliefs about how to do so—terminates in my intention to wear a raincoat. At best, I am attending to or thinking *about* the language with the aid of inner speech. A fluent Spanish speaker, on the other hand, could use an inner speech 'utterance' of, "está lloviendo," as an expression of a belief about the weather that could enter in further deliberation about how to appropriately dress for the outdoors.[x,xi]

In subsection 2.1, I use research on the role of inner speech in working memory to argue that we have to reason *about* features of the language used in order to keep track of the information with which we are reasoning. Employing inner speech as a tool to maintain access to our premises, lemmas, etc., does not involve thinking through the inner speech to its semantic content but thinking about (represented) phonological features of the language in which the inner speech is encoded. I also contextualize my position in the larger literature on inner speech. I then discuss two case studies in subsection 2.2 that provide compelling evidence for my claim that keeping information in mind with the aid of inner speech involves reasoning *about* language. I close the section (in 2.3) by responding to the objection that my conclusions are an artefact of a myopic focus on our use of inner speech in arithmetic calculation. I proceed to discuss the epistemic implications of my arguments in section 3.

2.1 *Working memory and inner speech*

As mentioned in the introduction, I adopt Baddeley's model of working memory. The model consists of a central executive that utilizes and governs two slave systems that store information in different modal formats, namely, the phonological loop (hereafter, PL) and the visuospatial sketchpad (hereafter, VSSP).[xii] If we keep track of, say, a carried value in determining the product of two integers by using the aid of a language (as opposed to, e.g., Arabic numerals or an abacus system, which lack an acoustic expression), then we would employ PL.[xiii] There are two routes to maintaining our access to some bit of information, Γ, with PL. The first route—the Utterance Loop—involves generating an overt utterance, U, with the content of Γ. The Utterance Loop should be a familiar means of keeping information in mind. For instance, imagine you attempt to keep track of a phone number while you search for pen and paper to write the number down. Until pen and paper are secured, you will likely rehearse the telephone number by repeatedly saying the number out loud. If you don't engage in an active process of rehearsal, you have about a two second buffer (in virtue of the functioning of the phonological store, as discussed below) before your conscious representation of the number degrades (Baddeley, 2007). The Utterance Loop involves generating and deciphering an extramental representation—an overt utterance—in the process of maintaining access to Γ. For the sake of this paper, we focus on the second route—the Inner Speech Loop—as we are interested in the epistemic role of inner speech in reasoning.[xiv]

As Oppenheim and Dell (2008, p. 529) note, "we produce inner speech the same way that we speak, except that articulation is not present." There is near universal agreement within psycholinguistics that speech planning is subdivided into levels or tiers in which different aspects of a to-be-spoken message are constructed and then sent to lower tiers for further processing (Levelt, 1993). Roughly,

speech planning occurs through the following: The first tier—the semantic level—involves a preverbal representation of the content that is to be encoded into language. The semantic level is followed by a syntactic/lemma level at which lemmas are placed in a grammatico-functional structure, where a lemma for a word is a set of semantic and syntactic information stored in the mental lexicon, dealing with syntactic category, grammatical function, and diacritical features like tense, mood, etc.[xv] At the morphophonological level the structure chosen at the syntactic/lemma level is realized in morphemes and phonemes while the prosodic character of the overall utterance is also determined. The message is then translated by the motor control areas of the brain into instructions executable by the vocal tract.

Inner speech is produced by halting the process of speech planning and production at a level that precedes overt articulation, thus resulting in a representation in a quasi-motoric code (Wilson, 2001). A sensory forward model is then generated—that is, an internal prediction of the sound of the planned (yet aborted) utterance—and sent to sensory areas (Scott, 2013; Tian & Poeppel, 2012).[xvi] The resulting mental state—the inner speech—is a representation of the phonological structure of an utterance. In turn, this phonological representation is kept in a buffer—the phonological store—for up to around two seconds before the representation needs to be 'refreshed'. So, we can split PL into two subcomponents:

> **The Phonological Store**: Commonly known as the 'inner ear,' the phonological store is a buffer that stores a phonological representation of an utterance (inner speech) for up to around two seconds.
>
> **The Articulatory Rehearsal Process**: Commonly known as the 'inner voice,' the articulatory rehearsal process involves the generation of a representation of an utterance in a phonological-sensory code (inner speech) that loads into or refreshes the representation in the phonological store.

Although it's beyond the scope of this paper to examine the full range of evidence for the inner voice and inner ear (cf. Wilson, 2001), we presently discuss the word length and phonological similarity effects, which provide strong support for the existence of the two subcomponents of PL. The word length effect is the following,

> **Word Length Effect**: It is more taxing on working memory to keep an object/concept, O_1, in mind than an object/concept, O_2, with the aid of a language, L, and PL if the name of O_1 takes longer to speak than the name of O_2 in L (Alan Baddeley, Thomson, & Buchanan, 1975).

For example, say you are trying to keep your grocery list in mind while engaging in practical deliberation regarding the most efficient route to take through the market. It would be easier for you to keep track of and reason with the following four items, assuming you are using the English language to code the list,

> (1) apple, plum, grapes, dates
>
> than the following,
>
> (2) pineapple, banana, watermelon, cantaloupe

The discrepancy in stress placed on working memory in keeping (1) as opposed to (2) in mind has nothing to do with the types of fruits in the lists and everything to do with the amount of time required to speak the names of the types in English. Pineapples are no more conceptually complex or harder to

comprehend than plums, but it takes longer to say, "pineapple," than it does, "plum." If there was no articulatory planning involved in refreshing the phonological representation kept in working memory, then we wouldn't expect there to be any discrepancy in stress placed on working memory in keeping (1) or (2) in mind. The word length effect is the result of the operations of the articulatory rehearsal process (the inner voice) in refreshing the inner speech in the phonological store.

Returning to arithmetic cognition, certain languages provide a relative cognitive advantage in calculation in virtue of how long it takes an average speaker to pronounce number words. For example, Stigler, Lee, and Stevenson (1986) demonstrate that Chinese students (first and fifth graders) tend to possess a larger digit span than their English-speaking American counterparts, that is, Chinese students demonstrate an ability to keep a larger sequence of digits in working memory. The larger an individual's digit span the easier it is for her to keep track of partial products, carried or borrowed values, etc., in the execution of common calculation algorithms. The researchers suggest that the discrepancy in digit span is due, in part, to differences in pronunciation times for number words in the respective languages used.[xvii] In addition, Geary, Bow-Thomas, Fan, and Siegler (1993) demonstrate that Chinese students, even prior to receiving formal training in school, tend to be better than their American counterparts at mental addition. Preceding formal training, children frequently use counting algorithms to determine the sums of integers (Lemaire, Barrett, Fayol, & Abdi, 1994). Geary et al. suggest that the relative advantage exhibited by the Chinese students is, in part, due to discrepancies in ease of counting—it is faster and, thereby, less taxing on working memory to count in Chinese (more specifically, Wu Chinese, as the students tested were from Hangzhou) than it is in English in virtue of discrepancies in pronunciation time.

The phonological similarity effect is the following,

> **Phonological Similarity Effect**: If the names of the respective members of a group of objects/concepts, Θ_1, in a language, L, are phonologically more similar than the names of the respective members of a group of objects/concepts, Θ_2, in L, it will be more taxing on working memory to keep Θ_1 in mind with the aid of L and PL than Θ_2.

For example, people have a harder time keeping the following objects/concepts in mind, assuming they are using the English language to code the list,

(3) man, cat, cap, map, can

than the following,

(4) huge, big, long, wide, tall (Alan Baddeley, 2017b)

in virtue of the fact that the words used to express the objects/concepts in (3) in English are phonologically more similar than those used to express the objects/concepts in (4). Phonological similarity causes interference in the proper functioning of working memory in virtue of the format—the phonological-sensory code—of the representations in the phonological store (the inner ear).

Note: the objects/concepts in (4) are more closely related (all being size concepts) than those in (3). The relatedness of the members of a set of objects/concepts only has a marginal (if any) interference effect on working memory in comparison to the phonological similarity of the words used to refer to those objects/concepts. However, the reverse holds for long-term memory—lists of objects/concepts that are closer in kind (in some relevant sense), like (4), are significantly harder to store in long-term memory than lists of unrelated objects/concepts, like (3) (Alan Baddeley, 1966). This discrepancy is indicative of the difference in how information is coded in working memory as opposed to long-term memory.

According to prominent models of mathematical cognition, mental calculation—the retrieval and utilization of mathematical facts and calculation algorithms—occurs through,

> (5) amodal numeric concepts, e.g., number 'words' in the language of thought (hereafter, LoT) (McCloskey, 1992; Sokol, McCloskey, Cohen, & Aliminosa, 1991),
>
> or
>
> (6) some combination of representational formats, including (i) a verbal word frame in which numbers are represented through syntactically sequenced word lemmas, which, as discussed, are word representations that don't encode phonological information, (ii) a visuospatial code used in processing GRSs like Arabic numerals, and (iii) an analog, spatial magnitude code used in representing the relative magnitudes of numeric quantities (Dehaene, 1992; Dehaene & Cohen, 1995; Skagenholt, Träff, Västfjäll, & Skagerlund, 2018).[xviii]

Theorists offer various reasons for accepting either (5) or (6) as the representational medium(s) of mental arithmetic.[xix] However, what matters for our concerns is that there is a consensus that the representational substrate of mental arithmetic does *not* consist in representations of the phonological structure of number words. We do not store memorized mathematical facts in a phonological format, nor does mathematical reasoning consist in operations over bare phonological representations of number words. In addition, neither amodal numeric concepts nor word lemmas are kept active in parallel with the corresponding phonological representations in the phonological store.[xx] So,

> (7) When using inner speech and PL, intermediate calculation results are stored in a phonological-sensory code in the phonological store and refreshed through speech planning and production processes.
>
> (8) Mathematical cognition consists in operations over amodal concepts or syntactically sequenced word lemmas (depending on your theoretical persuasions and assuming we are using language as a representational resource), as opposed to bare phonological representations.
>
> (9) Neither amodal numeric concepts nor word lemmas are kept active in parallel with the corresponding phonological representations in the phonological store.

The inner speech used to keep information in mind with PL consists of bare phonological representations of linguistic expressions. Therefore, in order to retrieve the numeric content we are trying to keep in mind

with the aid of inner speech and PL in a format that we can use in mathematical cognition, we have to reason from the phonological representations that constitute the inner speech, given our (tacit) background knowledge of the phonology of the language used.[xxi] In other words, in order to re-access and use the semantic content of the inner speech in PL, we have to interpret the speech based on the (represented) phonological information, just as we have to interpret the overt speech of another in order to access the semantic content of the person's utterances.

In the next subsection, I examine two case studies that provide compelling evidence for (9) and demonstrate that we engage in self-interpretation to re-access the information kept in mind with the aid of inner speech and PL in conscious reasoning. However, before discussing the cases, several general remarks on inner speech are in order. First, it's quite plausible that inner speech is not a homologous mental phenomenon. Depending on the cognitive task in which inner speech figures, inner speech may vary in terms of the phonological/phonemic/articulatory information encoded in the representation.[xxii] In some instances, inner speech may take an expanded form in which it exhibits (or, more accurately, represents) the phonological features of articulated speech and may, in other instances, take a condensed form in which the phonological features are greatly attenuated (Alderson-Day & Fernyhough, 2015; Fernyhough, 2004). Again, speech planning and production occurs in a series of stages in which different aspects of an utterance are generated. The level at which the speech planning and production process is aborted to generate inner speech may differ depending on the cognitive operation of which the inner speech is a part. It may be that, for certain tasks, inner speech is produced by halting the speech planning and production system at the syntactic/lemma level such that the inner speech consists of a series of lemmas placed in a grammatico-functional structure (Vicente & Jorba, 2019). At this stage of processing, no phonological information is encoded, and, therefore, we shouldn't expect this inner speech to be subject to, say, the phonological similarity effect, unlike the inner speech involved in PL. However, we are focused on our use of inner speech as a means of maintaining our access to information in working memory, and in this and (especially) the following subsections I give ample reason to think that the inner speech involved in PL is the result of halting speech planning and production at the phonological level. Information is not kept in mind in PL through a rehearsal of word lemmas or a direct reactivation of some amodal conceptual representation in LoT.

Second, it's beyond the scope of this paper to fully situate my arguments in the context of other philosophical work on inner speech. However, it will be instructive to discuss how my position comports with Peter Carruthers's extensive writings on the topic (e.g., 1998, 2002, 2011, 2018).[xxiii] According to Carruthers, inner speech crucially allows us to gain (indirect) access to the content of non-conscious conceptual thought (in LoT) through the following process: Non-conscious conceptual thought is fed into the speech planning and production system. A sensory forward model is generated from the quasi-motoric output of speech planning and production and sent to sensory areas. Speech comprehension processes then interpret the sensory forward model resulting in an episode of inner speech that consists of a phonological representation of a linguistic expression and an amodal conceptual representation of the semantic content of the expression 'bound into' the phonological representation (to use Carruthers's terminology).[xxiv] The interpretation of the sensory forward model happens 'upstream' of the conscious inner speech episode thus resulting in our,

> having immediate access to a particular phonological representation, together with its interpretation. The latter point is worth stressing: when I form an image of a natural-

> language sentence in a language that I understand…what figures in consciousness is not just a phonological object standing in need of interpretation. Rather, what figures there is **already interpreted—I hear meaning in the words**. (2000, p. 121, emphasis mine)

I am arguing that the inner speech involved in maintaining access to information in PL consists of uninterpreted 'linguistic images' (Jackendoff, 1996), that is, representations of phonological features of expressions (or utterances of those expressions) unbound from any amodal concepts or word lemmas. By generating and attending to a phonological representation of a linguistic expression with the semantic content, Γ, we are able to 'anchor' or 'stabilize' in working memory (as Jackendoff puts it, ibid.) a representation that can be used by speech comprehension processes to re-token the thought that Γ in an extended line of reasoning. Pace Carruthers, images of expressions in languages we understand needn't be bound to amodal conceptual representations of the content of the expressions.

Although my account of inner speech in PL (and the case studies discussed in the following subsection) certainly puts pressure on Carruthers's characterization of inner speech—being more conservative—there may be a way to make our respective accounts compatible. Working memory is a limited capacity mental workspace involved in both (i) processing and manipulating information, e.g., in conscious reasoning and (ii) buffering and storing the results of (i). As Daneman and Carpenter aptly put it, "[w]orking memory is assumed to have **processing as well as storage functions**; it serves as the site for executing processes and for storing the products of these processes' (1980, p. 450 emphasis mine).[xxv] While Carruthers is primarily interested in offering an account of our use of inner speech in (i), I am concerned with our use of inner speech in (ii). It may be that inner speech takes different forms when used as a cognitive tool in (i) or (ii). So, when inner speech is at the forefront of attention and functioning as, say, an occurrent judgment (or the expression thereof) in deliberation, the inner speech consists of a representation of the phonological structure of an expression bound to an amodal conceptual representation of the semantic content of the expression. On the other hand, the inner speech used in PL to keep track of information through an extended deliberative process does not have an amodal conceptual representation (or word lemma) bound into the speech. Admittedly, the distinction between (i) and (ii) is rather fine-grained, and more needs to be said about why and how inner speech would take different forms in its uses in (i) and (ii); but, for sake of space, I leave the ultimate compatibility of Carruthers's and my respective positions as an open question for further research.[xxvi]

In addition, although I cite Jackendoff's view of inner speech favorably, Jackendoff argues (more radically) that 'thought per say' is never conscious, where 'thought per say' for Jackendoff includes (minimally) propositional representations realized by discrete, amodal concepts as well as spatial representations realized in an analogue, geometric mode of representation. However, as my discussion of the potential compatibility of Carruthers's and my respective positions should make clear, I am not committed to the claim that amodal concepts are never conscious. Again, our focus in this paper is on our use of language as an expedient cognitive tool to keep track of information in reasoning. I'm merely committed to it being the case that amodal concepts are not active in the phonological store of PL. In order to establish the more radical claim that amodal concepts are *never* conscious, we would need to examine and explain away phenomena that provide at least prima facie evidence for the contrary position. For example, we would need to discuss and explain the phenomena of unsymbolized thought (Vicente & Martínez-Manrique, 2016), briefly mentioned in the introduction, which is beyond the scope of this paper.

Third, throughout this section I talk as if we *reason* from propositions about the phonological representations that constitute our inner speech in PL to retrieve the information we are attempting to

store with the speech. One might object that I've illicitly assumed an inferentialist picture of speech comprehension (or at least an inferentialist picture of *inner* speech comprehension). In turn, one might argue that speech comprehension for utterances that occur in languages with which we are sufficiently fluent is perceptual. We *perceive* the utterance meaning of speech; we don't infer the utterance meaning from lower-level (e.g., phonological) features of the speech (cf. Bayne, 2009; Brogaard, 2018, 2019; Fricker, 2003; Siegel, 2005).[xxvii] Therefore, the process of re-accessing the content stored through inner speech in PL is not inferential; it is perceptual (or so the argument goes). However, what ultimately matters for my conclusion and for the discussion of the epistemic implications in section 3 is the following,

> **Fundamental Conclusion**: Given that the inner speech in PL consists of bare phonological representations of linguistic expressions, keeping information in mind with the aid of inner speech requires generating and subsequently *interpreting* the speech. Interpreting the inner speech in PL could either consist of reasoning from (represented) features of the inner speech (i.e., the represented phonological structure) or relying on a quasi-perceptual speech comprehension faculty to carry out the interpretation.

2.2 *Case studies*

In this subsection we discuss two case studies that provide compelling reason to believe that neither amodal concepts nor word lemmas are kept active in parallel with the corresponding phonological representations in the phonological store. So, if we use inner speech as a tool to keep track of information in an extended process of conscious reasoning, we will have to engage in self-interpretation to re-access the information kept in mind.

The first case study is reported by Benson and Denckla (1969) and involves a retired army sergeant who presented with numeric paraphasia and paragraphia in virtue of brain damage, meaning, the sergeant frequently misspoke number words and miswrote numerals.[xxviii] For example, the sergeant was orally asked to write the Arabic numeral corresponding to, "two hundred twenty-one," and he produced, "215."[xxix] Although the sergeant's numeric production mechanisms and ability to translate between representational codes (e.g., English number words and Arabic numerals) were clearly compromised, his numeric comprehension and background knowledge of calculation procedures and arithmetic facts appeared to be intact (therefore, the sergeant's background mathematical knowledge wasn't stored in an orthographic or phonological format). For example, the sergeant could reliably circle the larger of two numbers presented visually. In addition, the sergeant could near flawlessly choose the correct solution to simple arithmetic problems when presented with a list of possible solutions, although he couldn't reliably utter the number word or write the Arabic numeral corresponding to the solution. For example, when orally asked the sum of four and five the sergeant uttered, "eight," wrote, "5," but pointed to, "9," on a list of possible solutions. The sergeant knew at the time of uttering, "eight," that he intended to refer to the number nine—as evidenced by his pointing to, "9," on a list of possible solutions—however, his paraphasia prevented him from producing a correct linguistic utterance. The sergeant's case demonstrates that linguistic production and comprehension are dissociable. The sergeant was able to access background

(tacit) knowledge regarding the phonology and orthography of English to map spoken and written words onto numeric contents without being able to do the reverse in his own speech and writing.

The sergeant had trouble using calculation algorithms, like long multiplication, the implementation of which requires storing intermediate results (e.g., carried values and partial products) in working memory. As Benson and Denckla report, the sergeant would at times mutter "incorrect numbers for the intermediate steps and often used the paraphasic response in the following steps" (ibid. p. 98), which resulted in the sergeant being unable to indicate the correct value on a list of possible solutions. Given what we know about the sergeant, his overt linguistic reports didn't reliably reflect his intermediate calculation results. The sergeant could correctly determine, say, the partial products of two integers by rote recall. In other words, the sergeant's mathematical woes were not caused by an inability to access and utilize background mathematical knowledge in conscious reasoning. Instead, because of his paraphasia, the sergeant couldn't consistently produce phonological representations of number words corresponding to his intermediate calculation results. Therefore, the sergeant couldn't reliably use the English language as a cognitive resource to code and store his intermediate calculation results because he couldn't reliably generate and decipher phonological representations of the number words corresponding to the numeric values he was attempting to keep in mind. If the sergeant was able to store his intermediate calculation results in working memory as a series of amodal number concepts or word lemmas—as opposed to, or in addition to, bare phonological representations of number words—then his paraphasia would be no impediment to his implementation of long multiplication. (This is my interpretation of the sergeant's errors. Benson and Denckla do not provide a diagnosis of the sergeant's working memory deficit. However, Benson's and Denckla's case study occurred well before Baddeley first published a developed account of his model of working memory in 1974. Therefore, Benson and Denckla couldn't use Baddeley's model to interpret their work.)

The second study is also a case in which linguistic production and comprehension are dissociated. Locke and Kutz (1975) examined a group of children who frequently substituted the phoneme /w/ for /r/ in speech, e.g., the children pronounced the English word, "ring," as, "wing." These phoneme substitution errors were restricted to speech production and didn't affect comprehension. When presented with a series of pictures—including one picture of a ring and one of a wing—the children could consistently point to the ring and to the wing upon hearing the respective name of the pictured object as uttered by one of the experimenters. So, the children could reliably understand and differentiate utterances of, "ring," and, "wing," (and differentiate the objects being referred to) while being unable to reliably produce these distinct utterances. The children had no trouble thinking differentially about rings and wings, either in light of exogenous or endogenous factors, and could tell what they were intending to talk about at the time of uttering, "wing," (they knew whether they meant to indicate a ring or wing). However, importantly, when audio recordings of their utterances of, "wing," were played back at a later time, outside of the context in which they were occurrently thinking about either rings or wings, the children could not reliably tell whether they had intended to refer to rings or wings in the utterances.

Over a series of varied trials, the children were presented with a recording of one of the experimenters saying three words chosen from the following list: "ring," "wing," "shack," and, "sack," where the experimenter might repeat one of the words on a given trial. After an eight second pause the children had to point to the pictures of the objects named by the experimenter, where the children would point twice at an object if it was named twice in the recording. The experimenters state, "[t]he central finding of this experiment was that children who correctly perceived /w/ and /r/ but consistently spoke one in place of the other experienced significantly more /w - r/ confusions in recall than children who characteristically produced a correct /w/ and /r/" (ibid. pp. 184-185). In other words, the children's inability to differentially pronounce, "ring," and, "wing," affected their ability to tell whether they were trying to keep rings or wings in mind with the aid of inner speech and PL.

Again, if there was something like an amodal concept or word lemma tagged to or processed in parallel with the inner speech used in PL, then neither the sergeant nor the children would exhibit problems keeping certain information in mind with the aid of inner speech and PL. The sergeant and children would just use the concept or word lemma that accompanied whatever phonological representation they generated to access the information being stored. What best explains the pattern of errors we see in the cases of the sergeant and children is the fact that inner speech in PL consists of bare phonological representations, unbound from amodal concepts or word lemmas.

As already stated, and as Locke and Kutz demonstrate, there is nothing special about the inner speech used in mathematical cognition to keep track of intermediate conclusions in conscious reasoning. What I say about arithmetic calculation applies, mutatis mutandis, to other domains as well. Regardless of whether we are thinking about numbers or rings and wings (or other non-numeric entities) with the aid of inner speech, the inner speech used to keep track of information in PL consists of bare phonological representations that need to be interpreted in order for us to use the semantic content of the represented expressions in further deliberation.

Our use of inner speech and PL is functionally similar to writing our premises and lemmas on a chalkboard to keep track of relevant information in an extended process of reasoning. When we look back at the orthographic markings we've generated as a means of re-accessing, say, a lemma or interim conclusion, we have to interpret the markings. The chalkboard serves as a tool to *offload* the storage of our premises, interim lemmas, etc., to an expedient (external) linguistic medium. In offloading storage to the chalkboard, we clearly don't *also* maintain, say, amodal conceptual representations of our premises and lemmas in long-term or working memory such that the orthographic markings merely serve to cue these stored representations that we subsequently employ in deliberation (this would be a complete waste of cognitive resources). Instead, the orthographic markings on the chalkboard *replace our need to store the information in another representational format*. In order to retrieve the semantic content of a given lemma written on the chalkboard so as to use the lemma in a further step in reasoning, we have to reason from information about the structure of the markings and our (tacit) background knowledge of the orthography of the language used to encode the lemma. Likewise, the inner speech in PL serves to offload the storage of premises, lemmas etc., to an expedient (internal) linguistic medium. In order to extract the semantic content of the inner speech, we have to reason from the phonological information represented using our (tacit) background understanding of the phonology of the language used to encode the information being stored.

It is certainly not my claim that using a chalkboard to store premises, lemmas, etc., in an extended line of reasoning is functionally identical to using inner speech and PL. For example, storing information with PL requires executive resources to 'refresh' the representation in the phonological store, while writing on a chalkboard creates a stable representation that requires no executive resources to maintain. However, the functional differences between using a chalkboard and using PL are irrelevant for our epistemic understanding and evaluation of an extended process of reasoning. In epistemically evaluating an extended process of (theoretical) reasoning, we are concerned with one's evidence/reasons, the inferential transitions one performs, and whether one's resultant attitudes are properly supported and grounded in one's evidence/reasons. Using a chalkboard and using PL both require (i) encoding the information with which one is reasoning into a linguistic format (either orthographic or phonological) and (ii) decoding the information from the orthographic or (represented) phonological properties of the language using one's background orthographic/phonological understanding of the language in which one encodes the information.

2.3 *The banality of arithmetic calculation*

In the previous subsection, I used Locke's and Kutz's work on children who substitute the phoneme /w/ for /r/ in speech to justify the claim that there is nothing special about the inner speech used in arithmetic cognition to keep track of intermediate conclusions in conscious reasoning. One might grant that keeping a series of partial products or a list of words (e.g., the words, "ring," and, "wing") in mind with the aid of inner speech and PL involves bare phonological representations that must be interpreted; however, one might object that keeping interim conclusions in mind (outside of the context of arithmetic reasoning) isn't like keeping track of a series of numbers or a list of words. So, one might be suspicious that my conclusion is an artefact of my focus on arithmetic reasoning. In response, it will be informative to look at how language deficits can affect performance on reasoning tasks outside of the arithmetic domain. To this end, we will briefly examine how patients with aphasia (hereafter, PWA) perform on the Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 1993)—a commonly used test of multistep problem solving.[xxx]

In the WCST, examinees are presented with four 'target' cards on which a series of colored shapes are printed. Each card has tokens of only one shape type and one color printed on it. The shapes are easily named (e.g., "triangles"), and only a small number of shapes are printed on a given card. Therefore, our innate number sense is sufficient to determine, without explicitly counting, whether any two cards have the same number of printed shapes. The examinee is tasked with matching a fifth card to one of the four target cards. The cards are to be matched on the basis of either color, number of shape tokens, or shape type. The examinee must determine the correct matching rule (e.g., match on the basis of shape type) by trying a match and receiving feedback from the experimenter. After the examinee tries a match, the experimenter simply responds by saying, "correct/incorrect." Periodically, the experimenter will switch the matching rule without alerting the examinee, and the examinee must learn the new rule solely on the basis of experimenter feedback. So, successfully deducing and using the matching rule requires that examinees keep track of (i) experimenter feedback and (ii) *their own interim conclusions about the matching rule*—e.g., the rule can't be to match on the basis of color, given the experimenter said, "incorrect," when I attempted to do so.

Although the relevant features of the cards are presented visually, as Baddeley notes, "adult subjects typically opt to name and [in inner speech] subvocally rehearse visually presented items, thereby transferring the information from a visual to an auditory code" (2000, p. 419), which allows individuals to take advantage of PL in storing relevant information in working memory. Language is an expedient representational resource we frequently use to code and track information with which we are reasoning. In neurotypical examinees, when inner speech is disrupted through articulatory suppression (e.g., forcing examinees to repeatedly say, "na na na," thus loading the speech planning and production system), the examinees typically perform worse on the WCST because of their inability to keep track of information in a convenient linguistic format (Baldo et al., 2005). Performance on the WCST also correlates with a number of language measures, and PWA struggle on the WCST in proportion to the severity of their language deficit (ibid.). In contrast, PWA demonstrate normal performance on purely visuospatial reasoning tasks, like maze learning (Archibald, Wepman, & Jones, 1967) or visual pattern completion (Baldo, Paulraj, Curran, & Dronkers, 2015), for which people don't typically use the strategy of coding and storing relevant information with the aid of inner speech. In the experiments reported in (Baldo et al., 2005), PWA were able to match cards on the basis of color, number of shape tokens, or shape type when *explicitly given the matching rule*. However, when PWA had to deduce and keep track of changes to the matching rule on the basis of experimenter feedback, PWA didn't perform as well as their non-aphasic counterparts. Because PWA—like neurotypical examinees—attempt to use inner speech as a resource to code information and intermediate conclusions regarding what the matching rule could be (e.g., the rule

isn't to match on the basis of color), their language deficits prove a hinderance to their problem-solving abilities. Similarly, the sergeant's mathematical woes were due to his inability to use inner speech as a representational medium to code and store his interim mathematical conclusions. If PWA were coding their interim conclusions about the matching rule in, say, an amodal conceptual medium that could be kept active in PL without having to generate and interpret phonological representations of linguistic expressions, then PWA's language deficits should be no impediment to their success on the reasoning task.

Although we've only discussed PWA's performance on the WCST, PWA show deficits on a range of reasoning and problem-solving tasks for which people commonly use language to code and track relevant information (see Baldo et al., 2015 for further citations).[xxxi] Again, there is nothing special about our use of inner speech in arithmetic calculation. Coding and storing information with inner speech and PL involves the generation and interpretation of bare phonological representations of linguistic expressions. For these generation and interpretation processes to be successful, speech planning, production, and comprehension systems need to be functioning well.

3. EPISTEMIC IMPLICATIONS

Given Fundamental Conclusion—keeping information in mind with the aid of inner speech requires generating and subsequently *interpreting* the bare phonological representations in the phonological store—it appears to follow that a belief that results from an extended process of reasoning in which one uses inner speech and PL to keep track of information will be partially epistemically based in either,

(10) attitudes about the phonological features of the language used in inner speech and PL (and tacit background attitudes about the phonological structure of the language),

or

(11) quasi-perceptual seemings regarding the utterance meaning of the inner speech,

where one's choice between (10) and (11) will depend on whether one adopts an inferentialist or perceptual account of speech comprehension, respectively. For example, a rational implementation of the long multiplication algorithm requires that we keep track of carried values and partial products that have to be summed. Insofar as we use inner speech to aid in calculation, we will have to reason from phonological features of the language used or rely on our quasi-perceptual speech comprehension faculty to decipher the meaning of the language in order to bring the relevant partial products back to the focus

of attention. Therefore, our conclusion will be partially based on (10) or (11), as either (10) or (11) will be actively used as an informational source in our reasoning.

If our resultant mathematical attitude is partially based on (10) or (11) then the attitude will be justified *a posteriori*. This is a surprising result given that mathematical knowledge is traditionally taken to be a paradigm of a priori knowledge (or knowledge that is justified a priori), at least when the knowledge is the result of an internal reasoning process, as opposed to, say, the use of a calculator.[xxxii] In addition, because there is nothing special about the mathematical domain and the nature of the inner speech used to keep track of information, any extended process of reasoning in which we rely on language as a means of coding our intermediate reasoning results will terminate in an attitude that is partially based in (10) or (11). Either (10) or (11) appears to be a near ubiquitous justificatory element in reasoning in which we use language as a representational resource.

In response, one could argue that working memory functions as a means of *preserving* our access to and justification for certain propositions, e.g., the content of our lemmas, as opposed to functioning as an evidential source regarding, e.g., the event of our proving some proposition at an earlier point in reasoning. As Burge argues,

> Purely preservative memory introduces no subject matter, constitutes no element in a justification, and adds no force to a justification or entitlement. It simply maintains in justificational space a cognitive content with its judgmental force. Like inference, it makes transitions of reason possible, but contributes no propositional content. Unlike inference, it is not a transition or move - so it is not an element in a justification. (1993, p. 465)

We aren't reasoning from or operating over phonological representations of number words—or so the objection goes—these phonological representations serve only to preserve our access to and justification for the information with which we are reasoning.

Pace Burge, keeping information in mind with inner speech and PL *does not* involve maintaining a 'cognitive content' (as Burge puts it)—as the cases of the sergeant and the children who substitute /w/ and /r/ in speech demonstrate—and it *does* involve a series of transitions and moves, e.g., a series of transitions between a representation of a number (or number word lemma) and a phonological representation of words that express that number in a given language (and vice versa). Again, if one uses inner speech and PL, the information being kept in mind is coded into a phonological representation of words that express the information as opposed to being represented and maintained in something like LoT. To use the information again, one must reason from, or utilize a quasi-perceptual faculty to determine the meaning of, the inner speech with the aid of a host of (tacit) background information regarding the phonological structure of the relevant language.

As a further response, one might rightly note that not all information used in reasoning needs to serve as part of the epistemic basis of the resultant attitude. For example, when one performs a conditional proof from a supposition that P to a conclusion of the form, 'if P then Q', one's epistemic basis for the conclusion clearly doesn't involve the initial assumption that P (cf. Valaris, 2016; Wright, 2014). Similarly, one might argue that although either (10) or (11) is used in reasoning, neither serve as part of the epistemic basis of the resultant attitude.

However, suppositions are different in kind from the inner speech phonological representations buffered and refreshed in PL. These phonological representations are not constitutive of occurrent suppositions or assumptions that are discharged in reasoning. Instead, the phonological representations

are best characterized as cognitive tools used to keep track of relevant information (just as orthographic markings on a chalkboard are tools to keep track of relevant information). We have to generate and then reason from features of the language used or rely on quasi-perceptual speech comprehension capacities to retrieve the information being stored with the aid of inner speech. Unlike conditional proof, there is no step of discharging (10) or (11) in our reasoning such that our resultant attitudes aren't based on (10) or (11).

Despite the failure of appeals to the role of supposition in reasoning or the ostensible preservative role of (working) memory, neither (10) nor (11) appears to serve as a reason for or a partial epistemic basis of our conclusions. For instance, say we come to believe that 37 x 43 = 1,591 through an internal execution of long multiplication with the aid of inner speech. It doesn't seem that we need to cite as a reason for our conclusion that, say, it quasi-perceptually appears to us that the inner speech we are using to store information about the first calculated partial product refers to the number 21. As I presently argue, in order to understand the epistemic role of language in maintaining our access to some bit of information, we have to draw a distinction between first-order procedures and metacognitive/meta-reasoning processes that serve to guide execution at the first-order level.

As Elena Rosca claims,

> It is now well established that the ability to solve a complex arithmetical operation is a multi-componential process that includes different types of knowledge: arithmetical facts (e.g. multiplication tables), procedural knowledge (e.g. the use of carrying and borrowing) and the understanding of arithmetic operations and the laws concerning these operations, a process known as "conceptual" knowledge. (2009a, p. 148)

There are numerous case studies reported in the cognitive psychological literature of patients suffering from brain damage that affects (i) the individual's procedural knowledge of certain calculation algorithms (Rosca, 2009b) or (ii) the individual's ability to metacognitively monitor the implementation of these algorithms (Semenza, Miceli, & Girelli, 1997), while leaving background conceptual knowledge and knowledge of arithmetic facts intact. The knowledge types that Rosca mentions are dissociable information stores that need to be used in tandem in successful calculation.

Procedural knowledge and error monitoring play a crucial role in mathematical cognition, but they do not serve as reasons for our resultant beliefs, unlike our knowledge of mathematical facts and background conceptual knowledge, which do play the role of reasons. For example, we wouldn't (nor should we!) cite as a reason for a mathematical belief that it didn't appear as if we made any errors in the implementation of the relevant calculation algorithm, despite the fact that—as evidenced by cases where metacognitive monitoring is damaged—our metacognitive monitoring was a crucial source of information used to accurately implement the relevant algorithm. More broadly, there is a significant body of cognitive psychological literature on the importance of metacognitive monitoring and control in reasoning. For instance, various metacognitive cues can result in a 'feeling of rightness' or a sense of processing fluency that we use to guide the allocation of cognitive effort, the time put towards reasoning, and the types of decision procedures we employ (Thompson et al., 2013). These metacognitive cues serve as a crucial

source of information about our reasoning itself that guides first-order reasoning procedures; however, these cues do not serve as part of the epistemic basis of or as reasons for our resultant attitudes.

The inner speech used with PL to keep information in mind plays a crucial metacognitive role in reasoning. In using inner speech in the phonological store to re-access previous reasoning results we are 'thinking about thinking'—we are thinking about previous reasoning results and using this information to guide further steps in reasoning. The distinction between,

> (12) information used at a first-order level in reasoning that serves as the epistemic basis of or reason for our conclusion,
>
> and
>
> (13) information used at a metacognitive level that serves to guide our implementation of the decision procedures used at the first-order level

hasn't been appreciated in contemporary epistemology, but it is central to understanding a crucial epistemic role played by inner speech.

It is beyond the scope of this paper to give a full account of the distinction between (12) and (13) and how drawing the distinction can further our understanding of the epistemic roles of various representations used in reasoning. However, in closing, I discuss at least one way in which (12) and (13) can epistemically interact and how drawing the distinction is crucial for understanding the interaction. For instance, say you come to believe a mathematical proposition, P, through your implementation of a calculation algorithm in which you use language as a representational resource to aid in storing and recalling intermediate results. You then gain evidence, E, that the inner speech you use to keep track of your previous reasoning results isn't a reliable guide because you are paraphasic like the sergeant. E appears to serve as a defeater for your belief that P—it would be (doxastically) irrational for you to maintain your belief in P given that you ineluctably rely on inner speech to serve as an indicator of, say, the partial products you previously calculated. But how does E operate as a defeater? It's clear that the traditional categories of rebutting and undercutting defeat are insufficient.[xxxiii] E isn't a rebutting defeater, as it doesn't provide you with direct evidence for the negation of your conclusion. E also isn't an undercutting defeater because it's not the case that E "attacks the connection between the evidence and the conclusion" (Pollock & Cruz 1999, p. 196), in that it doesn't directly provide you with evidence that your background mathematical attitudes don't justify the conclusion that P. Instead, E serves as an undercutter at a metacognitive level, as it 'attacks' the connection between the represented phonological features of the inner speech and your attitudes about, say, the value of previously calculated partial products.[xxxiv]

In sum, either (10) or (11) plays a crucial epistemic role in any extended process of reasoning in which we use language as a means of coding and storing information, like our lemmas or intermediate reasoning results. However, neither (10) nor (11) serves as a reason for or partial epistemic ground of our conclusions. In order to understand the epistemic role of (10) and (11) we need to distinguish first-order reasoning from metacognition.

4. CONCLUSION

I've established that using language and inner speech as a representational resource to keep track of information, like lemmas and intermediate conclusions, either involves reasoning from features of the language itself or using a quasi-perceptual speech comprehension faculty to interpret the speech. We don't merely reason through inner speech, we also use inner speech as a crucial source of information regarding past reasoning results. In turn, I argued that adequately understanding the epistemic role of language in keeping track of information requires that we distinguish first-order reasoning and metacognition/meta-reasoning processes.

Future research should explore how the distinction between first-order reasoning and metacognition can aid in an analysis of the epistemic role of our use of non-linguistic systems of representation as well, for instance, mental abacus (hereafter, MA). MA is an extremely efficient and effective method of mental calculation, which, as its name suggests, consists in the generation of and operation over a visuospatial representation of an abacus. The execution of MA involves extensive use of brain regions that subserve visuospatial and visuomotor processing as opposed to regions that subserve linguistic processing (Hanakawa, Honda, Okada, Fukuyama, & Shibasaki, 2003), and it's generally accepted that, "MA relies on visual resources…to create visual representations of exact number that differ fundamentally from those constructed using natural language"(Michael C Frank & Barner, 2012, p. 134). Outside of mathematical cognition, the distinction between first-order reasoning and metacognition may help explain the epistemic role of diagrams (Shin, 1994) and models (Nersessian, 1999, 2010) in reasoning, e.g., how we ought to epistemically conceive of our use of diagrams or models as a means of extracting information about the states of affairs diagrammed or modeled. In addition, future research should further examine how metacognitive and first-order reasoning processes epistemically interact in terms of the evaluation of one's reasoning processes and resultant attitudes.

Works Cited

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Shedding meta-cognitive light on reasoning research. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge international handbook of thinking and reasoning* (pp. 1-15). New York: Routledge.

Alderson-Day, B., & Fernyhough, C. (2015). Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychological bulletin, 141*(5), 931.

Archibald, Y., Wepman, J., & Jones, L. (1967). Nonverbal cognitive performance in aphasic and nonaphasic brain-damaged patients. *Cortex, 3*(3), 275-294.

Baddeley, A. (1966). The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly journal of experimental psychology, 18*(4), 302-309.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A, 49*(1), 5-28.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423.

Baddeley, A. (2007). *Working Memory, Thought, and Action*. Oxford: Oxford University Press.

Baddeley, A. (2017a). *Exploring Working Memory: Selected Works of Alan Baddeley*. London: Routledge.

Baddeley, A. (2017b). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. In *Exploring Working Memory* (pp. 9-14): Routledge.

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior, 14*(6), 575-589.

Baddeley, A. D., & Hitch, G. J. (2019). The phonological loop as a buffer store: An update. *Cortex, 112*, 91-106.

Baldo, J. V., Dronkers, N. F., Wilkins, D., Ludy, C., Raskin, P., & Kim, J. (2005). Is problem solving dependent on language? *Brain and language, 92*(3), 240-250.

Baldo, J. V., Paulraj, S. R., Curran, B. C., & Dronkers, N. F. (2015). Impaired reasoning and problem-solving in individuals with language impairment due to aphasia or language delay. *Frontiers in psychology, 6*, 1523.

Bayne, T. (2009). Perception and the Reach of Phenomenal Content. *Philosophical Quarterly, 59*(236), 385-404.

Benson, D. F., & Denckla, M. B. (1969). Verbal Paraphasia as a Source of Calculation Distrubance *Archives of Neurology, 21*(1), 96-102.

Bermúdez, J. L. (2003). *Thinking Without Words*: Oxford University Press.

Boghossian, P. (2014). What is Inference? *Philosophical Studies, 169*(1), 1-18.

Brogaard, B. (2018). In defense of hearing meanings. *Synthese, 195*(7), 2967-2983.

Brogaard, B. (2019). Seeing and hearing meanings: A non-inferential approach to speech comprehension. In *Inference and Consciousness* (pp. 99-124): Routledge.

Broome, J. (2009). The unity of reasoning? In S. Robertson (Ed.), *Spheres of Reason*: Oxford University Press.

Burge, T. (1993). Content Preservation. *Philosophical Review, 102*(4), 457-488.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and cognition, 5*(1), 41-66.

Caramazza, A., & McCloskey, M. (1988). The case for single-patient studies. *Cognitive neuropsychology, 5*(5), 517-527.

Carruthers, P. (1998). Thinking in language? Evolution and a modularist possibility. *Language and thought: Interdisciplinary themes*, 94-119.

Carruthers, P. (2002). The Cognitive Functions of Language. *Behavioral and Brain Sciences, 25*(6), 657-674.

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

Carruthers, P. (2014). On central cognition. *Philosophical Studies, 170*(1), 143-162.

Carruthers, P. (2018). The causes and contents of inner speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices* (pp. 31-52): Oxford University Press.

Christensen, D. (2010). Higher-Order Evidence. *Philosophy and Phenomenological Research, 81*(1), 185-215.

Cooper, J. M., & Hutchinson, D. S. (1997). *Plato: complete works*: Hackett Publishing.

Daneman, M., & Carpenter, P. A. (1980). Individual Differences in Working Memory and Reading. *Journal of verbal learning and verbal behavior, 19*(4), 450-466.

De Guerrero, M. C. (2006). *Inner speech-L2: Thinking words in a second language* (Vol. 6). New York: Springer Science & Business Media.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*(1-2), 1-42.

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical cognition, 1*(1), 83-120.

DiPaolo, J. (2018). Higher-Order Defeat is Object-Independent. *Pacific Philosophical Quarterly, 99*, 248-269.

Drożdżowicz, A. (2019). Do we hear meanings?–between perception and cognition. *Inquiry*, 1-33.

Ellis, N. (1992). Linguistic relativity revisited: The bilingual word-length effect in working memory during counting, remembering numbers, and mental calculation. In R. J. Harris (Ed.), *Cognitive Processing in Bilinguals* (Vol. 83, pp. 137-155): Elsevier.

Everett, D., Berlin, B., Gonalves, M., Kay, P., Levinson, S., Pawley, A., . . . Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current anthropology, 46*(4), 621-646.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307-314.

Fernyhough, C. (2004). Alien voices and inner dialogue: towards a developmental account of auditory verbal hallucinations. *New ideas in Psychology, 22*(1), 49-68.

Frank, M. C., & Barner, D. (2012). Representing Exact Number Visually Using Mental Abacus. *Journal of Experimental Psychology: General, 141*(1), 134-139.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition, 108*(3), 819-824.

Frank, M. C., Fedorenko, E., & Gibson, E. (2008). *Language as a cognitive technology: English-speakers match like pirahã when you don't let them count.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Frankish, K. (2004). *Mind and supermind*: Cambridge University Press.

Frankish, K. (2012). Dual systems and dual attitudes. *Mind & Society, 11*(1), 41-51.

Frankish, K. (2018). Inner Speech and Outer Thought. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices* (pp. 221-243): Oxford University Press.

Fricker, E. (2003). Understanding and knowledge of what is said. In A. Barber (Ed.), *Epistemology of Language* (pp. 325--366): Oxford University Press.

Gasparri, L., & Murez, M. (forthcoming). Hearing meanings: the revenge of context. *Synthese*.

Gauker, C. (2018). Inner Speech as the Internalization of Outer Speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices* (pp. 53-77): Oxford University Press.

Geary, D. C., Bow-Thomas, C. C., Fan, L., & Siegler, R. S. (1993). Even before formal instruction, Chinese children outperform American children in mental addition. *Cognitive development, 8*(4), 517-529.

Geurts, B. (2018). Making sense of self talk. *Review of Philosophy and Psychology, 9*(2), 271-285.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *science, 306*(5695), 496-499.

Gregory, D. (2016). Inner speech, imagined speech, and auditory verbal hallucinations. *Review of Philosophy and Psychology, 7*(3), 653-673.

Hanakawa, T., Honda, M., Okada, T., Fukuyama, H., & Shibasaki, H. (2003). Neural Correlates Underlying Mental Calculation in Abacus Experts: A Functional Magnetic Resonance Imaging Study. *Neuroimage, 19*(2), 296-307.

Hatzigeorgiadis, A., Zourbanos, N., Galanis, E., & Theodorakis, Y. (2011). Self-Talk and Sports Performance: A Meta-Analysis. *Perspect Psychol Sci, 6*(4), 348-356. doi:10.1177/1745691611413136

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test (WCST): manual: revised and expanded*: Psychological Assessment Resources (PAR).

Heavey, C. L., & Hurlburt, R. T. (2008). The Phenomena of Inner Experience. *Consciousness and Cognition, 17*(3), 798-810.

Hula, W. D., & McNeil, M. R. (2008). Models of attention and dual-task performance as explanatory constructs in aphasia. *Semin Speech Lang, 29*(3), 169-187; quiz C 163-164. doi:10.1055/s-0028-1082882

Hurlburt, R. T., & Akhter, S. A. (2008). Unsymbolized Thinking. *Consciousness and Cognition, 17*(4), 1364-1374.

Hurlburt, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition, 22*(4), 1477-1494.

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

Langland-Hassan, P. (2021). Inner speech. *WIREs Cognitive Science, 12*(2), e1544. doi:https://doi.org/10.1002/wcs.1544

Langland-Hassan, P. (2014). Inner speech and metacognition: in search of a connection. *Mind & Language, 29*(5), 511-533.

Lasonen-Aarnio, M. (2014). Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research, 88*(2), 314-345.

Laures-Gore, J., Marshall, R. S., & Verner, E. (2011). Performance of Individuals with Left-Hemisphere Stroke and Aphasia and Individuals with Right Brain Damage on Forward and Backward Digit Span Tasks. *Aphasiology, 25*(1), 43-56. doi:10.1080/02687031003714426

Lee, K.-M., & Kang, S.-Y. (2002). Arithmetic operation and working memory: Differential suppression in dual tasks. *Cognition, 83*(3), B63-B68.

Lemaire, P., Barrett, S. E., Fayol, M., & Abdi, H. (1994). Automatic activation of addition and multiplication facts in elementary school children. *Journal of Experimental Child Psychology, 57*(2), 224-258.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1): MIT press.

Locke, J. L., & Kutz, K. J. (1975). Memory for speech and speech for memory. *Journal of Speech and Hearing Research, 18*(1), 176-191.

Lœvenbruck, H., Grandchamp, R., Rapin, L., Nalborczyk, L., & Dohen, M. (2018). A Cognitive Neuroscience View of Inner Language. In P. Langland-Hassan & A. n. Vicente (Eds.), *Inner Speech: New Voices* (pp. 131-167).

Machery, E. (2005). You Don't Know How You Think: Introspection and Language of Thought. *British Journal for the Philosophy of Science, 56*(3), 469-485

Malmgren, A.-S. (2006). Is there a priori knowledge by testimony? *The Philosophical Review, 115*(2), 199-241.

Malmgren, A.-S. (forthcoming). A Priori Testimony Revisited. In A. Casullo & J. Thurow (Eds.), *The A Priori in Philosophy*: Oxford University Press.

Martin, R. C., & Allen, C. M. (2008). *A disorder of executive function and its role in language processing.* Paper presented at the Seminars in speech and language.

Martínez-Manrique, F., & Vicente, A. (2015). The activity view of inner speech. *Frontiers in psychology, 6*, 232.

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition, 44*(1-2), 107-157.

Miller, J. T. M. (2020). The ontology of words: Realism, nominalism, and eliminativism. *Philosophy Compass, 15*(7), e12691. doi:https://doi.org/10.1111/phc3.12691

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 5-22): Springer Science & Business Media.

Nersessian, N. J. (2010). *Creating scientific concepts*: MIT press.

O'Callaghan, C. (2011). Against hearing meanings. *The Philosophical Quarterly, 61*(245), 783-807.

Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition, 106*(1), 528-537.

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M., & Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research, 261*, 220-239.

Pettit, D. (2009). On the Epistemology and Psychology of Speech Comprehension. *The Baltic International Yearbook of Cognition, Logic and Communication, 5*.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *science, 306*(5695), 499-503.

Pollock, J. L., & Cruz, J. (1999). *Contemporary Theories of Knowledge* (2 ed. Vol. 35). Lanham, Maryland: Rowman & Littlefield.

Prinz, J. (2012). *The Conscious Brain*. New York: Oxford University Press.

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and individual differences, 20*(2), 110-122.

Rönnberg, J., Rudner, M., & Ingvar, M. (2004). Neural correlates of working memory for sign language. *Cognitive Brain Research, 20*(2), 165-182.

Rosca, E. C. (2009a). Arithmetic procedural knowledge: a cortico-subcortical circuit. *Brain research, 1302*, 148-156.

Rosca, E. C. (2009b). A case of acalculia due to impaired procedural knowledge. *Neurological sciences, 30*(2), 163.

Scott, M. (2013). Corollary discharge provides the sensory content of inner speech. *Psychological science, 24*(9), 1824-1830.

Semenza, C., Miceli, L., & Girelli, L. (1997). A Deficit for Arithmetical Procedures: Lack of Knowledge or Lack of Monitoring? *Cortex, 33*(3), 483-498.

Shanon, B. (1984). The polyglot mismatch and the monolingual tie: Observations regarding the meaning of numbers and the meaning of words. *New ideas in Psychology*.

Shapiro, S. (2000). *Thinking about mathematics: The philosophy of mathematics*: OUP Oxford.

Shin, S.-J. (1994). *The Logical Status of Diagrams*. Cambridge: Cambridge University Press.

Siegel, S. (2005). Which properties are represented in perception. In T. S. Gendler & J. Hawthorne (Eds.), *Perceptual Experience* (pp. 481--503): Oxford University Press.

Skagenholt, M., Träff, U., Västfjäll, D., & Skagerlund, K. (2018). Examining the Triple Code Model in numerical cognition: An fMRI study. *PloS one, 13*(6), e0199247.

Sokol, S. M., McCloskey, M., Cohen, N. J., & Aliminosa, D. (1991). Cognitive representations and processes in arithmetic: Inferences from the performance of brain-damaged subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 355.

Spaepen, E., Coppola, M., Spelke, E. S., Carey, S. E., & Goldin-Meadow, S. (2011). Number without a language model. *Proceedings of the National Academy of Sciences, 108*(8), 3163-3168.

Stigler, J. W., Lee, S.-Y., & Stevenson, H. W. (1986). Digit memory in Chinese and English: Evidence for a temporally limited store. *Cognition, 23*(1), 1-20.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The Role of Answer Fluency and Perceptual Fluency as Metacognitive Cues for Initiating Analytic Thinking. *Cognition, 128*(2), 237-251.

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in human neuroscience, 6*, 314.

Valaris, M. (2016). Supposition and Blindness. *Mind, 125*(499), 895-901.

Vicente, A., & Jorba, M. (2019). The Linguistic Determination of Conscious Thought Contents. *Noûs, 53*(3), 737-759. doi:10.1111/nous.12239

Vicente, A., & Martínez-Manrique, F. (2016). The nature of unsymbolized thinking. *Philosophical Explorations, 19*(2), 173-187.

Wilson, M. (2001). The case for sensorimotor coding in working memory. *Psychonomic bulletin & review, 8*(1), 44-57.

Wilson, M., & Emmorey, K. (1997). A visuospatial "phonological loop" in working memory: Evidence from American Sign Language. *Memory & Cognition, 25*(3), 313-320.

Wright, C. (2014). Comment on Paul Boghossian, "What is inference". *Philosophical Studies, 1*(1), 1-11.

[i] I would like to thank Luis Oliveira and the Department of Philosophy at the University of Houston, Brian Weatherson, Peter Carruthers, Emmalon Davis, and an anonymous referee for this journal for helpful comments and discussion.

[ii] Some theorists, e.g., Keith Frankish (2004, 2012, 2018), argue that inner speech is constitutive of conscious thought as opposed to merely accompanying the thought.

[iii] See (Langland-Hassan, 2021) for more on how best to define inner speech.

[iv] Of course, we use inner speech for a litany of purposes outside of reasoning. As (Geurts, 2018; Martínez-Manrique & Vicente, 2015) argue, inner speech likely serves as many functions as overt speech. For example, inner speech is used as a form of self-encouragement in athletes (Hatzigeorgiadis, Zourbanos, Galanis, & Theodorakis, 2011). It is not my intention here to give an exhaustive account of our varied uses of inner speech.

[v] In recent philosophical work on reasoning and inference, it's generally accepted that reasoning is an operation over the *content*s of our attitudes, as opposed to, e.g., the natural language expressions used to represent those contents. For instance, John Broome writes, "in reasoning you say to yourself the propositions that constitute [the content of your attitudes], and you reason about these contents"(Broome, 2009, p. 71). Similarly, Paul Boghossian argues that, "reasoning is an operation on thought contents and not on symbols (that have content)"(Boghossian, 2014, p. 17).

[vi] As I note in subsection 2.1, whether we take ourselves to be reasoning about the language we use in inner speech or relying on what I call quasi-perceptual seemings regarding the language depends on whether we adopt an inferentialist or perceptual view of speech comprehension (cf. Pettit, 2009). I assume an inferentialist picture in formulating my argument for ease of discussion. However, as I explain, the assumption isn't crucial to the argument.

[vii] My comments about the metacognitive role of inner speech differ from other discussions of inner speech and metacognition; however, it's beyond the scope of this paper to explore the differences (cf. Langland-Hassan, 2014).

[viii] Daniel Everett (2005) argues that the Pirahã language exhibits certain 'inexplicable gaps', e.g., the syntax doesn't allow for recursive embedding. While Everett's claims are highly contentious, there is no controversy (of which I am aware) regarding the nature of the number vocabulary of the Pirahã and their numeracy. Nothing I say commits me to the claims of Everett.

[ix] I talk as if inner speech constitutes the actual use of language 'in the head' (Gregory, 2016) as opposed to the mere auditory imagery of language use (Machery, 2005). This is in line with recent work on the metaphysics of words (Miller, 2020). However, this assumption is not crucial to my conclusion. In addition, I talk as if inner speech is a phonological representation of which we are conscious. Christopher Gauker (2018), on the other hand, argues that inner speech is whatever mental state or event causes the phonological representation, in the same way that an utterance is to be identified with a series of sound waves as opposed to a listener's conscious perception of the utterance. Again, this assumption is not crucial to my conclusion.

[x] In the acquisition of a second language, L2, it's common to use inner speech and L2 to think *about* or internally rehearse what one is going to say in L2 before saying it. But, importantly, there is a level of language proficiency after which L2 begins to be used as a means of entertaining thoughts in inner speech in which one is thinking *through* the inner speech in L2 (De Guerrero, 2006). On a related note, people are able to distinguish (i) episodes in which inner speech is 'in one's own voice' such that one is attaching, say, assertoric force to the utterance and (ii) episodes in which one is imagining an utterance as if it were someone else's (Hurlburt, Heavey, & Kelsey, 2013).

[xi] Metaphysical discussions of the nature of mathematical objects are orthogonal to our concerns. However, it should be noted that the distinction I draw between thinking through and thinking about the language of an utterance can be made even given a term formalism in which, say, the number two is identified as a certain symbol type, e.g., the type "2" (cf. Shapiro, 2000).

[xii] Here, I give a simplified characterization of Baddeley's model that is sufficient for our concerns. Baddeley adds an additional component to his model of working memory—the episodic buffer—in his (2000). We briefly discuss the episodic buffer in note 24.

[xiii] In this paper, I focus on spoken as opposed to signed languages. However, what I say about spoken languages will apply, mutatis mutandis, for signed languages as well. For instance, a native user of American Sign Language can keep information in mind through inner signing, an analogue of the inner speech loop, as we discuss in the following. In addition, there are similar interference patterns for signers as there are for users of spoken languages. For example, signers appear to be subject to the word length and phonological similarity effect, again, as we discuss in the following (Wilson & Emmorey, 1997). See (Rönnberg, Rudner, & Ingvar, 2004; Wilson, 2001) for further discussion.

[xiv] Of course, we could also ask about the epistemic role of private speech—that is, speaking to oneself—in reasoning. Because private speech involves the generation and utilization of extramental representations (i.e., overt utterances), it may introduce additional epistemic complications. It's beyond the scope of this paper to discuss private speech.

[xv] The sense of the term, "lemma," meaning a certain unit of lexical storage in the mental lexicon is clearly distinct from the sense of the term meaning an intermediate theorem or reasoning result. I use both senses of the term in this paper. The context in which I use, "lemma," should be sufficient to determine what sense of the term I mean to employ.

[xvi] In overt speech, the sensory forward model plays a variety of roles, e.g., allowing us to (i) rapidly distinguish self-produced sound from external acoustic signals and (ii) monitor and correct our overt speech.

[xvii] Similar results regarding digit span have been found for a host of other languages (Ellis, 1992).

[xviii] The combination of codes used for a given task will depend on the reasoning procedures and background knowledge one needs to deploy. For example, linguistic representation is commonly used in multiplication (Lee & Kang, 2002). This is likely due to the fact that the execution of common calculation algorithms, e.g., long multiplication, grid method, lattice method, etc., involve the use of multiplication and addition tables, which we acquire through rote *verbal* memorization.

[xix] For example, bilinguals tend to be more reliable and quicker when performing mental arithmetic with the aid of the language in which they were taught basic arithmetic facts and decision procedures, which suggests that arithmetic facts are stored in a linguistic format as opposed to being stored as amodal numeric concepts or number 'words' in LoT (Shanon, 1984).

[xx] Note: According to (6)—what is known as the Triple Code Model of numerical cognition—when we use inner speech in, say, determining the product of two numbers through an implementation of the long multiplication algorithm, our reasoning is a formal operation over representations of number words as opposed to an operation over amodal concepts of the numbers themselves. So, imagine we attempt to determine the product of 48 and 36 in our head using the English language as a representational resource. According to the Triple Code Model, we would begin with the first partial product and utter, "six times eight," in inner speech. This utterance would associatively activate the lemma, "forty-eight," in the mental lexicon, which would eventually result in us uttering, "forty-eight," in inner speech (again, the lemma associated with a word consists of stored syntactic and morphologic information in the mental lexicon). The first partial product would then have to be stored as inner speech in PL as we continue to implement the algorithm through a series of inner speech 'utterances' and rote associative connections. Eventually, the stored partial products would be re-accessed and summed to arrive at the final conclusion. However, even if the Triple Code Model is right, arithmetic calculation isn't an operation over *bare phonological representations of number words*; lemmas in the mental lexicon also need to be activated and associated with the phonological representations. Given the existence of homophones, a bare phonological representation isn't sufficient to determine which word one is using in inner speech. For instance, in representing the phonological structure, /wun/, we could be using either, "one," or, "won," in inner speech. So, if a representation of /wun/ is kept active in PL as a means of storing a carried value in mental arithmetic, then bringing the carried value back to the forefront of attention would still require *interpreting* the phonological representation as indicating the word, "one." This interpretation process would require that we activate the lemma associated with the term, "one," in the mental lexicon. In sum, even if our arithmetic reasoning is a formal operation over representations of number words, these representations are not purely representations of the phonological structure of linguistic expressions. Reasoning doesn't consist of operations over bare phonological representations, regardless of whether we are reasoning about numbers, politics, or what to wear given the temperature. In addition, as I argue, neither amodal concepts nor word lemmas are kept active in parallel with the corresponding phonological representations in the phonological store. Therefore, if we use language as a representational resource to store our intermediate reasoning results in working memory in an extended process of reasoning, we will have to interpret the inner speech in PL in order to re-access the information stored. (Thank you to an anonymous referee for pressing me to discuss the possibility that bare phonological representations serve as the medium of arithmetic reasoning.)

[xxi] As noted, I assume an inferentialist picture of speech comprehension. I discuss the assumption in the next subsection.

[xxii] For example, there is evidence that inner speech, at least at times, involves somatosensory imagery similar to the experience of employing one's larynx, tongue, mouth, and face in speaking. In other words, inner speech can involve an embodied simulation of speaking (Lœvenbruck, Grandchamp, Rapin, Nalborczyk, & Dohen, 2018).

[xxiii] Thank you to an anonymous referee for pressing me to discuss my position in relation to that of Carruthers's.

[xxiv] See (Langland-Hassan, 2014) for objections to Carruthers's position.
[xxv] It's common in the literature on working memory to mention both its storage and processing functions.
[xxvi] To flesh out the proposal a bit further: Baddeley adds an additional element to his model of working memory in his (2000), which he calls the episodic buffer. The episodic buffer is a passive store for episodes or chunks of information in a multidimensional code. The representations in the buffer are referred to as 'multi-dimensional' or 'multimodal' as they are the result of (i) integrating phonological/sound/articulatory information in PL with the visuospatial/kinetic/tactile information in VSSP as well as (ii) interfacing with "semantic and episodic long-term memory" (A. D. Baddeley & Hitch, 2019, p. 100). This integration happens outside of the episodic buffer—only the results of the integrative process are present in the buffer. It could be that inner speech can enter the episodic buffer as a result of the binding process Carruthers describes in which phonological and amodal conceptual representations are bound by speech comprehension processes. However, if the inner speech is looped through PL as a means of using the inner speech to maintain access to some bit of information, the amodal conceptual component will be lost and must be re-tokened by speech comprehension processes. (Martínez-Manrique and Vicente [2015] refer to the inner speech involved in PL as 'meaning ignoring' in contrast to the 'meaningful' inner speech used as, say, expressions of propositional attitudes in reasoning.) So, the (fast and loose) proposal is that inner speech in the episodic buffer functions as, say, a conscious occurrent judgment, supposition, etc., in deliberation and consists of a phonological representation bound into an amodal conceptual representation. But, if the inner speech is kept active through an extended deliberative process, it must be looped through PL, in which case the amodal conceptual component will be lost and can only be retrieved through interpreting the bare phonological representation in the phonological store.

My comments here are speculative. Whereas PL and VSSP are well-researched components of working memory (PL more so than VSSP), the episodic buffer and central executive have received comparatively little attention. Baddeley, for example, refers to the central executive as a 'ragbag' (1996) that, along with the episodic buffer, will likely be functionally decomposed into further subcomponents in future research, just as PL was decomposed into the phonological store and articulatory rehearsal process. The nature and function of the representations in the episodic buffer (and the episodic buffer itself) are certainly not as well understood as the nature and function of the representations used in PL.

[xxvii] See (Drożdżowicz, 2019; Gasparri & Murez, forthcoming; O'Callaghan, 2011) for arguments for the inferentialist account of speech comprehension.
[xxviii] The study of neuroatypical individuals (NA) has proven an important source of evidence for theories of cognition in neurotypical individuals (NT) in numerous domains. Although it may seem suspect to study cognition in NA in an attempt to understand cognition in NT, if cognition in a given domain, D, is (i) subserved by a series of functionally distinct, modular components, and (ii) brain damage can selectively disrupt a component(s) of the cognition (iii) without altering the functional organization of the components, then the study of NA can provide insight for a functional decomposition of the cognition of NT in D. For a detailed defense of the study of NA in service of understanding cognition in NT and the legitimacy of drawing conclusions from single case studies, as I am doing here, see (Caramazza, 1986; Caramazza & McCloskey, 1988).
[xxix] The sergeant's substitution errors were not systematic. It's not as if the sergeant remapped, "215," in his idiolect onto the number two hundred twenty-one such that he would always write, "215," when orally prompted to produce the Arabic numeral corresponding to the word, "two hundred twenty-one." The sergeant may very well produce different written numerals at various times in response to the same auditory prompt.
[xxx] Thank you to an anonymous referee for urging me to discuss an example of non-arithmetic reasoning.
[xxxi] It should be noted that it is notoriously difficult to determine experimentally exactly what explains PWA's deficits on a given non-verbal problem-solving task, like the WCST, where a task is non-verbal in the sense that the task doesn't appear, in principle, to require linguistic processing in generating or reporting a solution (thank you to an anonymous referee for raising this difficulty). The PWA's deficits could be due, specifically, to their linguistic impairment and the importance of language as a tool for problem solving, even on (what are traditionally categorized as) non-verbal tasks. Alternatively, the PWA's deficits could be explained by, e.g., a more general working memory deficit (Laures-Gore, Marshall, & Verner, 2011), impaired attention (Hula & McNeil, 2008), or impaired executive functions, like inhibition, cognitive flexibility, and planning (Martin & Allen, 2008), where these deficits result from the brain damage that also affects linguistic processing in the PWA in a manner unmediated by or independent of the PWA's specifically linguistic deficits. However, given the evidence regarding people's

tendency to linguistically label and subvocally rehearse visually presented items and the PWA's specific deficits in Baldo et al.'s study, the results provide good evidence that PWA's troubles on the WCST are due (in significant part) to their language deficits and, more specifically, their impaired ability to use inner speech and PL to buffer information in working memory.

[xxxii] One could argue along the lines of Burge that testimony that P provides one with (prima facie) a priori justification to believe P. In turn, one could argue that the inner speech used to keep information in mind is a form of self-testimony regarding past reasoning results. Therefore, our mathematical beliefs are still justified a priori. Although I don't have adequate space to respond to this argument, I am sympathetic to Anna-Sara Malmgren's arguments that testimony cannot provide a priori justification, as testimonial belief will be partially based on contingent facts about what the testifier actually said (Malmgren, 2006, forthcoming).

[xxxiii] The *locus classicus* on rebutting and undercutting defeat is (Pollock & Cruz, 1999).

[xxxiv] It might be suggested that E is higher-order evidence and is, thus, an instance of higher-order defeat (Christensen, 2010; DiPaolo, 2018; Lasonen-Aarnio, 2014). It is beyond the scope of this paper to discuss this possibility here.