***Acta Psychologica* (2017) 181: 62-74.**

Final version available here:
http://www.sciencedirect.com/science/article/pii/S0001691817300483

Metacognitive deficits in categorization tasks in a population with impaired inner speech

Peter Langland-Hassan[a*], Christopher Gauker[b], Michael J. Richardson[c], Aimee Dietz[d,] and Frank R. Faries[a]

[a] *Center for Cognition, Action & Perception and Department of Philosophy, University of Cincinnati, Cincinnati, OH, USA*

[b]*Department of Philosophy, Faculty of Cultural and Social Sciences, University of Salzburg, Salzburg, Austria*

[c]*Department of Psychology, Macquarie University, Syndney, Australia*

[d]*Department of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, OH, USA*

*Corresponding author. Email: Langland-Hassan@uc.edu

Metacognitive deficits in categorization tasks in a population with impaired inner speech

Peter Langland-Hassan[a*], Christopher Gauker[b], Michael J. Richardson[c], Aimee Dietz[d,] and Frank R. Faries[a]

[a]*Center for Cognition, Action & Perception and Department of Philosophy, University of Cincinnati, USA*

[b]*Department of Philosophy, Faculty of Cultural and Social Sciences, University of Salzburg, Austria*

[c]*Department of Psychology, Macquarie University, Syndney, Australia*

[d]*Department of Communication Sciences and Disorders, University of Cincinnati, USA*

Abstract: This study examines the relation of language use to a person's ability to perform categorization tasks and to assess their own abilities in those categorization tasks. A silent rhyming task was used to confirm that a group of people with post-stroke aphasia (PWA) had corresponding covert language production (or "inner speech") impairments. The performance of the PWA was then compared to that of age- and education-matched healthy controls on three kinds of categorization tasks and on metacognitive self-assessments of their performance on those tasks. The PWA showed no deficits in their ability to categorize objects for any of the three trial types (visual, thematic, and categorial). However, on the categorial trials, their metacognitive assessments of whether they had categorized correctly were less reliable than those of the control group. The categorial trials were distinguished from the others by the fact that the categorization could not be based on some immediately perceptible feature or on the objects' being found together in a type of scenario or setting. This result offers preliminary evidence for a link between covert language use and a specific form of metacognition.

Keywords: inner speech; metacognition; aphasia; language; categorization; concepts

*Corresponding author. Email: Langland-Hassan@uc.edu

1. Introduction

The most obvious function of language is that it facilitates communication. Yet there is increasing evidence that language has important *extra-communicative* cognitive functions as well, insofar as covert uses of language appear to influence performance on a diverse set of tasks unrelated to interpersonal communication. Some of these include categorization tasks (Plunkett, Hu, & Cohen, 2008; Lupyan & Mirman, 2013), memory tasks (Loewenstein & Genter, 2005; Papafragou, Hulburt, Trueswell, 2008), object individuation (Xu, 2002), relational judgments (Kotovsky & Gentner, 2005), event categorization (Papafragou & Selimis, 2010), task switching (Laurent *et al.*, 2016), and theory of mind judgments (Newton & de Villiers, 2007).

Some studies have investigated the cognitive functions of language by looking at the particular cognitive disabilities of people with aphasia (PWA), who have acquired language impairments due to stroke. For instance, some of these studies show that PWA have difficulties attending to specific dimensions of similarity when making taxonomic judgments during categorization (Noppeney & Wallesch, 2000; Cohen, Kelter, & Woll, 1980; Lupyan & Mirman, 2013); others reveal an influence of language on working memory capacity (Caspari et al, 1998) or attention (Murray, 2012). In some cases, these results from PWA have been corroborated in neurotypical populations under verbal interference (Lupyan, 2009). While some PWA have cognitive impairments that extend beyond their impaired linguistic systems (Glosser & Goodglass, 1990; Purdy, 2010; Murray, 1999), the above studies are of special interest in that they attempt to show that impaired task performance results specifically from damage to linguistic centers of the brain.

To date, however, there has been relatively little examination—in neurotypical or PWA populations—of the role that language may play in metacognition. A number of theorists have speculated that covert language production—in the form of "inner speech" (Alderson-Day &

Fernyhough, 2015)—may play an important role in bringing thoughts to consciousness (Carruthers, 1996; Jackendoff, 1996; Clark, 1998; Bermudez, 2003; Morin, 2009) and in allowing for critical reflection on one's own judgments and attitudes (Carruthers, 2011; Martinez-Manrique & Vicente, 2015). Yet there have been no quantitative empirical studies of this hypothesis as yet. In a similar vein, others have hypothesized that abnormalities in inner speech may lead to deficits in self-awareness; however, these reports pertain primarily to the auditory verbal hallucinations experienced by people with schizophrenia (Frith, 1992; Fernyhough, 2004; Langland-Hassan, 2008). As such, these speculative proposals do not directly speak to the role of language in normal cognition, including metacognition.

This study seeks to fill these gaps in the empirical literature by examining specifically the relation of language to a person's ability to perform categorization tasks and to assess their own abilities in those categorization tasks. In the study described here, participants performed a task in which they were called upon both to categorize objects and to judge whether they had done so correctly. The performance of control participants was compared to the performance of PWA who demonstrated an impairment with respect to inner speech, but who otherwise approached normal levels in other tests of cognitive abilities. The purpose of the study was twofold: first, to assess the extent to which language is needed for certain types of categorizations, and, second, to assess whether language plays a metacognitive role in enabling awareness of one's success in these types of categorizations.

1.1. Nonverbal tests of metacognition

Our means for testing metacognition in a language-impaired population was modelled in part on previous studies of metacognition with nonhuman animals. These studies have had the following structure: The subject is presented with tasks of varying difficulty, with known

rewards and penalties for answering correctly or incorrectly (Smith, Shields, Schull and Washburn, 1997; Hampton, 2001; Smith & Washburn, 2005; Kornell, Son, & Terrace, 2007; Smith, Beran, Couchman, & Coutinho, 2008). The actual abilities of the animal are typically assessed during an initial forced-choice condition. Subsequently, subjects are provided with a means for *opting out* of answering the prompt (typically by pressing a button designated as the opt-out choice). Opting out results in a lesser reward than answering the prompt correctly, but is preferable to the penalty (typically a time delay) received for answering incorrectly. Use of the opt-out key is said to be adaptive, and to indicate appropriate self-assessment, if it occurs on trials in which the subject would be expected to answer incorrectly, based on its prior performance.

In view of the similarity between the present study and these previous animal studies, we describe the present study as an investigation into metacognition. But in so describing it, we need to draw a distinction. Metacognition is often conceived of as the ability to have thoughts about thoughts, either one's own thoughts or the thoughts of others (Bermudez, 2003; Carruthers, 2011). Whether the tests of opt-out behavior are indeed proper tests of metacognition in this sense is a matter of some dispute. While most agree that the abilities possessed by species that can learn to use the opt-out key adaptively are of significant theoretical interest, some deny that they support the presence of metacognition on the grounds that they may not require the animal to form a representation of its own cognitive representations (Carruthers, 2008, 2011;Perner, 2012). Others reject this stringent standard for the use of the term "metacognition", on the grounds that cognition can have a metacognitive *function* even if it does not strictly involve forming a representation of a mental state (Proust, 2013). We will describe the experiment as a test of metacognition with the proviso that we do not assume that metacognition necessarily involves thoughts about other thoughts.

1.2. Covert language use and "inner speech"

It has been shown that the overt language deficits of PWA are not always mirrored by corresponding deficits in occurrent[1] covert language use (or "inner speech") (Geva et al., 2011; Stark, Geva & Warburton, 2017) and that the relationship between inner and outer speech capacities in PWA is complex and variable (Fama *et al.,* 2017). Therefore, to draw any conclusions about the role of occurrent covert language use from a population of PWA, it is important to establish that the population of PWA in fact have deficits in the covert use of language, or inner speech. We follow others in using silent rhyming abilities as a test for inner speech capacity (Levine, Calvano, & Popovics, 1982; Feinberg, Gonzalez Rothi, & Heilman, 1986; Geva, Bennett, Warburton, Patterson*,* 2011). These earlier studies have confirmed that people with aphasia often experience impaired inner speech. We incorporated a similar silent rhyming task into the present study in order to assess covert language use, while also confirming, via other cognitive screening tests (described below), that cognitive abilities of the PWA were normal or near normal in other respects. (Langland-Hassan *et al.*, 2015 contains a detailed examination of correlations among of the silent rhyming abilities of our population[2] and their abilities on non-linguistic cognitive tasks).

It might be questioned whether silent rhyming tasks are good tests for a lack of inner speech. In generating normal, overt speech, the mind must execute a number of theoretically separable tasks, including the selection of a grammatical structure and the selection of words to

[1] By 'occurrent' language use we mean what is sometimes called 'online' language use (Lupyan, 2009). These terms serve to distinguish the active exploitation of language production and processing capacities (which are 'occurrent' and 'online' uses of language), from the dispositional and structural changes to a cognitive system that may result from mastering a language. The latter may influence one's performance on a task—and thus be an effect of language—without requiring one to actively generate or comprehend new linguistic material.

[2] The population of PWA in Langland-Hassan *et al.* (2015) includes two participants who were excluded from the tests of categorization and metacognition reported here, due to their inability to follow instructions for those tasks.

populate the leaf nodes of the grammatical structure. (For one of many analyses, see Pickering & Garrod, 2013). A presumably late stage is the selection of a phonemic realization of the sentence or phrase to be spoken. An even later stage is the planning of muscle movements that generate the utterance, followed by the actual execution of the plan. During covert language use, the production does not get as far as actual movements of the vocal apparatus, but may include the generation of an iconic representation of the sound of a spoken utterance, which we will call the auditory imagery of inner speech. The phenomenon of inner speech might also include, as an additional step, an experience of "hearing", or processing, the inner speech that has been generated. Thus, a judgment of whether two words rhyme may presuppose yet a further level of articulation and comparison. How much of normal speech production is executed in inner speech is an open question, and the answer may vary from one occasion to another and between one person and another (Oppenheim & Dell, 2010; Perrone-Bertolotti *et al.*, 2014). For purposes of this study, we simply define inner speech operationally as *whatever capacity for occurrent covert language use is needed to pass the silent rhyming task*. This task is described in more detail below.

2. Methods

2.1 Participants

We recruited 13 participants with chronic post-stroke aphasia from a database held at the University of Cincinnati Augmentative and Alternative Communication and Aphasia Lab. Because we were primarily interested in the effect of inner speech deficits on metacognition, we selected individuals with conduction, anomic, or Broca's aphasia. In such patients, language comprehension is relatively strong, while overt language production is moderately-to-severely impaired. Eleven adults with no history of aphasia, mental illness, or substance abuse were

recruited to participate as part of the control group (3M/8F, mean age 58.5 ± 8.1, age range 47-78, mean years of education 14.6 ± 2.1) (Table 1a).

Four PWA from the original 13 were excluded from analysis for the following reasons. One was consented and underwent cognitive screening, but was excluded from the experimental sessions, including the silent rhyming test, on the grounds of receiving a WAB-R diagnosis of global aphasia and showing significant language-comprehension difficulties. Two were excluded from analysis because they did not understand the main experimental task and were unable to follow task instructions. Finally, a fourth participant was excluded because his high Aphasia Quotient score of 96 on the WAB-R no longer qualified him for a diagnosis of aphasia. (Any score above 94 is within normal limits). These exclusions resulted in N = 9 participants with aphasia (4M/5F, mean age 60.2 ± 8.2, age range 44-76, mean years of education 15.0 ± 1.7) (Table 1b). The control participants were roughly matched to the PWA in age [$t(18) = 0.434$, $p = 0.669$], gender, and education [$t(18) = 0.398$, $p = 0.696$]

2.2 Screening tests

Because our participants with impaired inner speech were all stroke survivors with aphasia, we could not assume that their cognitive functioning was in other respects normal. Consequently, the PWA completed basic vision and hearing screening exams, an out-loud rhyme judgment task, the Western Aphasia Battery-Revised (WAB-R) (Kertesz, 2006), and the Cognitive Linguistic Quick Task (CLQT) (Helm, 2003).

All PWA passed the vision and hearing exams. The out-loud rhyme judgment task consisted of asking the participant whether two words rhymed, for each of 10 pairs of one-syllable words spoken aloud by the experimenter. This was done so as to investigate the relation

between judging rhymes that are heard to judging rhymes silently, through inner speech. The mean for PWA on the out-loud rhyming task was 8.67 (out of 10).

The WAB-R was used to confirm aphasia severity and type (see Table 1b). And the CLQT (Table 1b) was used to rate participants on five different cognitive measures: *attention*, *executive function*, *visuospatial skills*, *language*, and *memory*. Because the CLQT sub-tests assessing language and memory were heavily language-dependent, the scores of PWA on these portions were disregarded. (The WAB-R provided a more comprehensive picture of the language deficits of the PWA.) We were mainly interested in participants' performance on the non-linguistic sub-tests of the CLQT that were relevant to assessing attention, executive function, and visuospatial skills. Such tasks included navigating mazes, connecting lines to shapes in a specified order, and generating novel designs using four lines, obeying stipulated constraints. For each measure, the CLQT rates participants as falling into one of four groups: *within normal limits*, *mildly impaired*, *moderately impaired*, or *severely impaired*.

When two tasks explicitly requiring language generation were subtracted from the CLQT sub-tests used to assess attention[3], executive functions, and visuospatial skills, all nine of the remaining aphasia participants scored within normal limits on those measures, with the following exceptions: one participant (#5, table 1b) was mildly impaired on attention and visuospatial reasoning and moderately impaired on executive function; a second (#8, table 1b) was mildly impaired on attention and visuospatial skills. We did not see the mild impairments by

[3] To give an example of how this subtraction was carried out, in the case of Attention there was one language-involving subtask (Story Retelling) that contributed a possible 12 points to one's Attention score. The normal range for falling within normal limits on Attention is 215-180. Thus, to calculate whether a person fell within normal limits on non-linguistic tests of attention, we first lowered the range that counts as within normal limits by 12 points to 203-168, and then lowered the participant's score on Attention by whatever amount was contributed by their performance on the Story Retelling task (for sometimes they scored points on the language-involving tasks, despite their aphasia). If that adjusted score then fell within the adjusted within-normal-limits range, they were judged to be within normal limits on Attention.

themselves as reason to exclude these participants from analysis. And while participant #5 showed moderate impairment on executive functions, her score was only one point from qualifying as mildly impaired, for her age group; further, she was close to within normal limits on both attention and visuospatial skills.

Table 1a

Demographic Information for Control Participants

	Gender	Age	Level of Education
1	Female	56	Some College
2	Female	78	Bachelor's
3	Female	59	High School
4	Female	57	Some College
5	Female	50	Bachelor's
6	Female	60	Some College
7	Female	47	Some College
8	Male	59	Master's
9	Male	60	Master's
10	Female	67	Bachelor's
11	Male	51	Some College

Table 1b

Demographic Information for Participants with Inner Speech Impairments

	Gender	Age	MPO[1]	Level of Education	Aphasia Type and Severity[2]		CLQT[3] Attention	CLQT Executive Functions	CLQT Visuospatial
1	Female	59	72	Some College	Conduction	81.8	191	27	101
2[4]	Male	44	112	Bachelor's	Broca's[4]	71.2	191	27	101
3	Male	58	76	Master's	Anomic	68.9	185	25	87
4[5]	Female	68	175	Bachelor's	Broca's[4]	43.9	172	19	87
5	Female	76	315	Some College	Broca's	66.1	129	8	59
6	Female	56	92	Some College	Broca's	50.6	191	27	101
7	Male	60	101	Bachelor's	Broca's	70.9	180	25	97
8	Female	62	172	Bachelor's	Conduction	81.1	151	19	80
9	Male	59	15	Associate's	Broca's	59.2	183	22	94

Note. [1]Months post-onset. [2]*Western Aphasia Battery-Revised* used to determine type and severity (Aphasia Quotient), total possible points = 100 (≤ 93 indicates presence of aphasia). [3]*Cognitive-Linguistic Quick Test*: Ranges for participants up to 69 years old: Attention: Within Normal Limits=203-168, Mild=167-113, Moderate=112-38, Severe=37-0 (range reflects 12 point adjustment, due to deleting Story Retelling task from sum); Executive Functions: WNL: 35-19, Mild: 18-15, Moderate: 14-11, Severe: 10-0 (range reflects 5 point adjustment, due to deleting Generative Naming task from sum); Visuospatial Skills: WNL: 105-82, Mild: 81-52, Moderate: 51-42, Severe: 41-0. For participants over 70 years and older, the adjusted ranges are: Attention: WNL: 203-148, Mild: 147-88, Moderate: 87-28, Severe: 27-0; Executive Functions: WNL 35-14, Mild: 13-9, Moderate: 8-3, Severe 2-0; Visuospatial Skills: WNL 105-62, Mild: 61-37, Moderate: 36-22, Severe 21-0. [4]Apraxia of Speech present based on clinical judgment. [5]Left-handed (pre-stroke).

2.3 Apparatus

Trials were presented on an Asus 8A-Series computer with a 21-inch touch-sensitive screen. 530 digital photographs and drawings were used to create 53 trials (8 of which were demonstration or training trials). The program was written in C++ and recorded responses and response times automatically (10 ms resolution). Metal washers were used as game tokens, to be won or lost. Auditory feedback accompanied correct, incorrect, or opt-out responses, as explained below.

2.4 Procedure – Silent Rhyming Task

Participants were shown pairs of pictures on a touchscreen computer and asked to indicate, without speaking aloud, whether the words for the pictures rhymed. After seeing four preliminary sets of pictures used for explanation and training, participants were shown forty sets of two pictures, one set at a time, and were asked to indicate, silently, whether the words for the pictured items rhymed (Figure 1). They could either answer "yes", by touching a green check, or answer "no", by touching a red X, or indicate that that they did not know, by touching a blue question mark. Touching the blue question mark was counted as an incorrect answer for purposes of scoring. Half of the prompts where "yes" was the correct answer involved pictures of items whose linguistic labels rhymed but did not share similar endings (e.g., "box" and "socks"). This was to prevent participants from answering via visual images of written words as opposed to auditory-phonological cues.

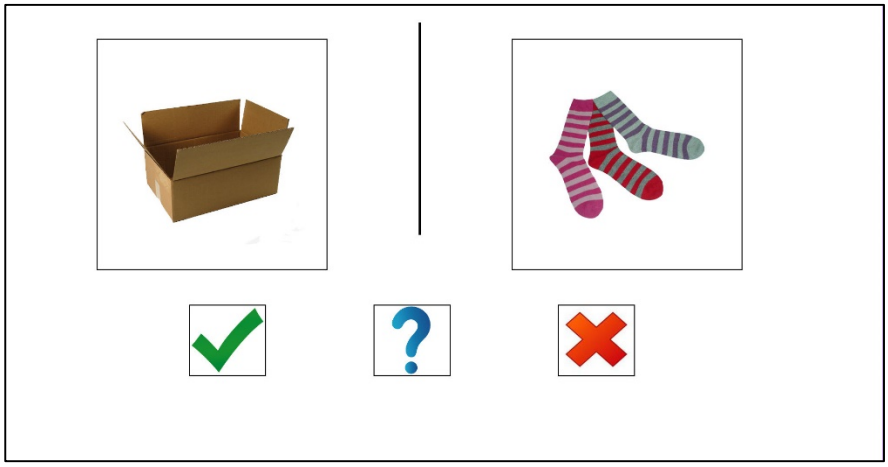


Figure 1: Example of a silent rhyming task trial

As noted above, the production of inner speech can be analyzed into several components. A poor performance on the pictorial rhyming task may result from a deficit in any one or more of these components. This study does not attempt to determine which sort of deficit might be responsible when a participant performs poorly on the silent rhyming task. "Inner speech" in our usage refers to whatever covert language-related impairment is responsible for poor performance on the silent rhyming task.[4]

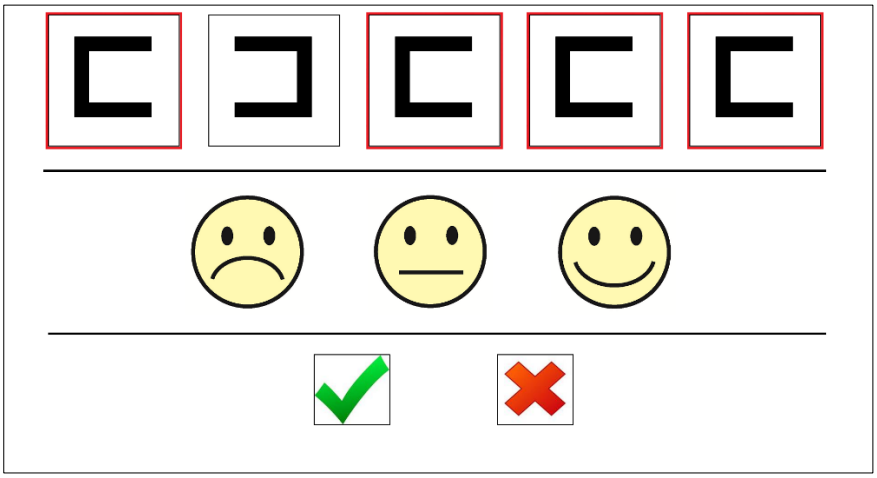
2.5 Procedure – Categorization and Metacognition Task

Participants were tested individually and told they would play a game on a touchscreen computer. The point of the game would be to win as many tokens as possible. It was emphasized that the tokens were not worth real money. Experimenters explained the rules of the game by modeling playing the game for the duration of four trials, discussing their thought processes out loud. The modeling was scripted so as to clearly reveal the different aspects of the

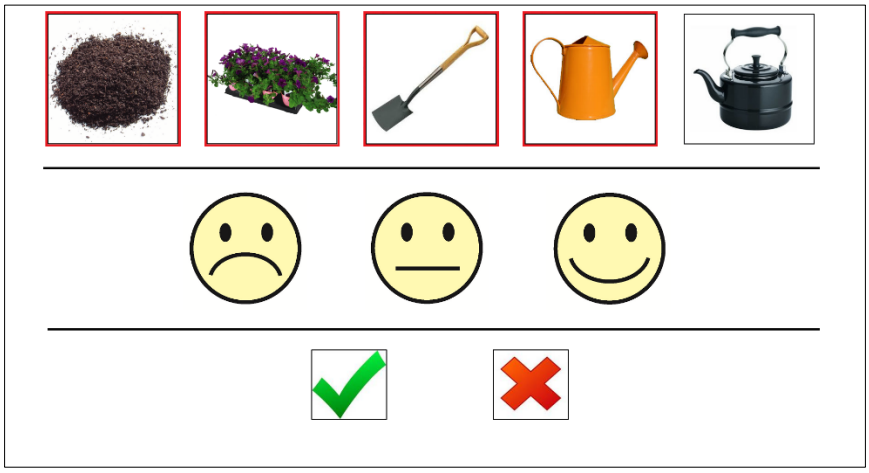
[4] Geva, Bennett, Warburton, Patterson, 2011, used written words as stimuli, not pictures, as we did. While this may obviate the possibility that the participant cannot find the word (since it appears in written form in front of them), it introduces the possibility that a poor rhyming performance stems from a reading deficit (Ullman, et al., 2005).

game, and to ensure that it was modeled in the same way for each participant. Each participant then completed 49 trials, taking as long as they needed. However, they were told they could not speak aloud while completing the trials, other than to ask for clarification about the rules of the game. The first four trials served as training trials, and were not included in the analysis. After the training trials, all participants completed the same 45 trials, in randomized order.

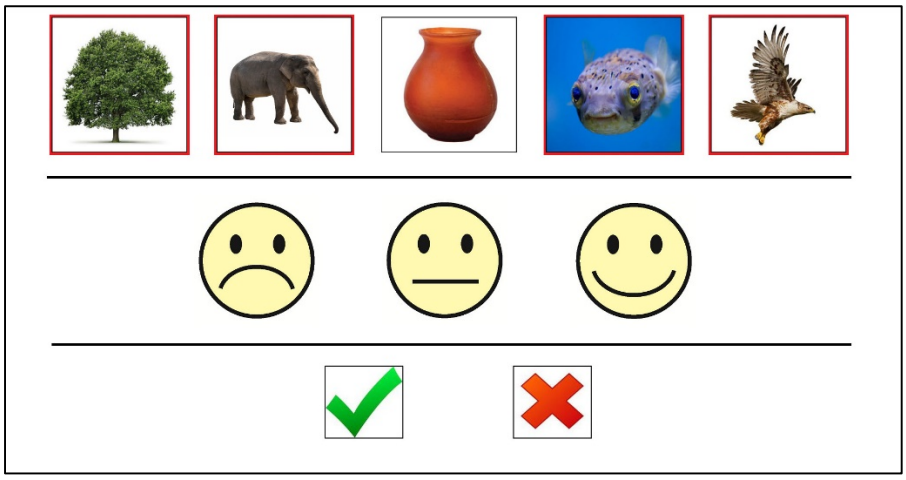
At the beginning of each new trial, the participant saw five pictures of different objects across the top of the screen. Below them, across the middle of the screen, were three *confidence faces* (one smiling face, one neutral, one frowning). Below the confidence faces, at the bottom of the screen, were a green check mark and red X. This part of the trial will be referred to as stage A of the trial. Sample stage A screens are shown in Figure 2. Participants were told that the first step was to touch the four out of the five pictures at the top that go together. When a picture was touched, a red box appeared around it. If the picture was touched again, the red box would disappear, allowing one to change one's mind. Participants were instructed that, in cases where they were not sure which four went together, they should go ahead and select four while giving it their best guess.



A geometric trial



A thematic trial ("gardening")



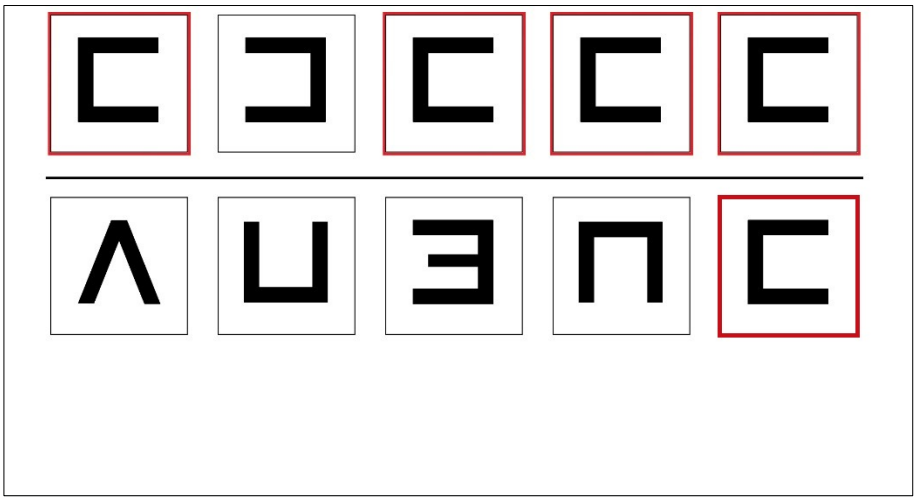
A categorial trial ("living beings")

Figure 2: Examples of Stage A stimuli, with correct four items selected

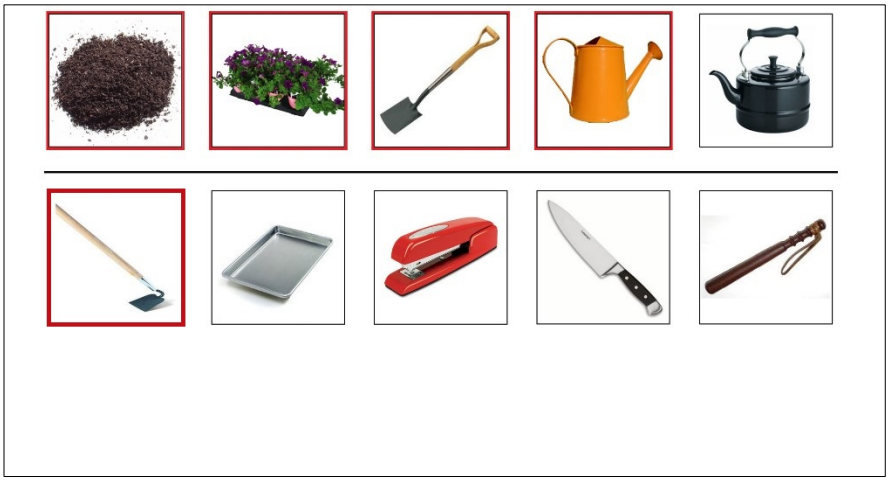
Once four objects were touched, with a red box showing around each, participants were directed to touch one of the confidence faces to indicate "how confident you are that you selected

the right four pictures." They were told that touching the smiling face served to indicate that they were "very confident," the neutral face to indicate that they were "a little bit confident," or the frowning face to indicate that they were "really not sure" which four items go together. We had no reason to doubt, on the basis of either their CLQT performance (see below) or their reactions *in situ*, that the participants understood these instructions. Once a confidence face was selected, participants were no longer able to change their selections with respect to the four items selected as going together.

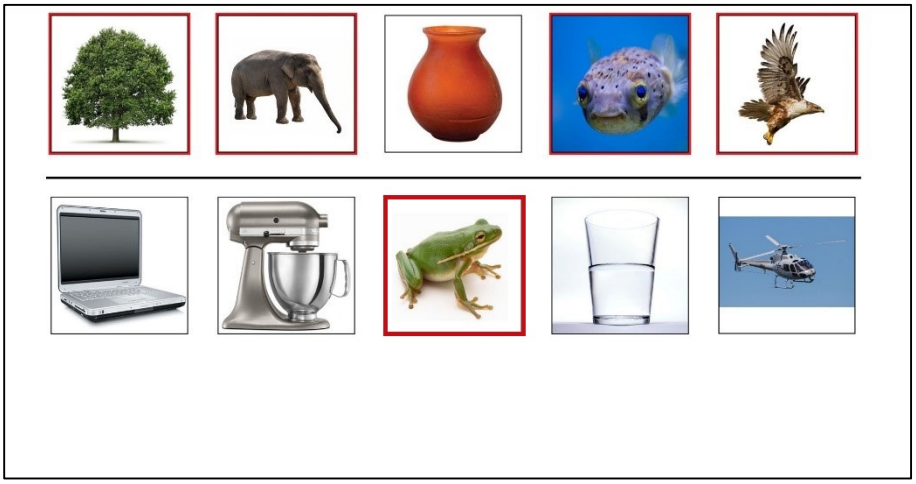
Participants were instructed that after they had selected one of the confidence faces, they should touch either the green check or the red X. If they touched the green check, they were taken to a new screen. This part of the trial will be referred to as stage B (Figure 3). At the top of the stage B screen appeared the same five images they saw at the top of the screen during stage A of the trial. Red boxes still appeared around the four pictures they selected. Below those five pictures appeared five new pictures. A participant's task was then to select the one picture out of the five new pictures that "goes with" the four they initially selected. Figure 3 shows examples of a stage B screens. If participants then touched the correct picture (for example, selecting a dog when the category uniting the first four pictures was *dogs*), they received three tokens. In addition, a pleasant winning sound, similar to bells chiming, was emitted from the computer. If, however, a picture was selected that did not go with the four from the first stage, participants lost three tokens, and the computer made an unpleasant buzzing sound (similar to a game show buzzer).



A geometric trial



A thematic trial ("gardening")



A categorial trial ("living beings")

Figure 3: Examples of Stage B stimuli

Finally, it was explained, through modeling, that if they touched the red X during stage A of the trial, instead of the green check, they would receive one token and would advance immediately to stage A of the next trial, without completing stage B. This in effect gave them the opportunity to *opt-out* of risking winning or losing three tokens on Stage B, by taking instead the lesser reward of 1 token and moving to the next trial. A neutral *ding* sound was emitted when the red X was selected.

The rationale for retaining the five pictures from stage A on the screen at stage B, along with the red outlines indicating which four had been chosen, was to ensure that the task at stage B was not a memory task but just a categorization task of the same kind as the participant had encountered at stage A.

2.6 Three types of trial

There were three types of trial, corresponding to three different kinds of category: *geometric*, *thematic*, and *categorial*. Four trials were used in modeling the task, and four were

practice trials for the participant. Not including their four training trials, participants completed 15 trials of each type, in random order. As there are always innumerable ways of grouping four objects into a category so as to exclude a fifth (provided the categories can be esoteric and artificial), stimuli were chosen with a goal of lessening the likelihood that multiple salient interpretations of the correct category would present themselves to participants.

The three trial types were distinguished as follows. Geometric trials used stimuli involving geometric shapes and figures. The unifying category for such trials always pertained to some visible feature of the stimuli. For 12 out of 15 of these trials the unifying feature was a shared color and/or shape, such as *right triangle*. For the other three, it was participation in a geometric motif, such as being a solid shape that is cross-cut by a straight line in two places, or being a solid shape with a small indentation on one side. A simple geometric trial involved four large green circles and one small green circle. To give the correct answer in stage A, participants had to touch the four large green circles. And, if they then chose to go forward to stage B, participants had to select the one large green circle among four distractors to win three tokens. A more difficult geometric trial involved four right triangles at various orientations and one non-right triangle in stage A. Some of the geometric trials were purposefully made difficult in the expectation that they would encourage participants to select a low confidence face and opt-out of stage B (i.e., select the red X).

For thematic trials, four of the five pictures were unified into a category in virtue of their being things one commonly finds together in a certain environment or setting. These trials used photographic images of actual objects. Examples of thematic categories included: things you take to the beach, kitchen items, picnic items, and things you see along the highway. For example, the picnic items included a cooler, a checkered tablecloth, picnic table, and plastic food

utensils. The distractor in this case was a dog. Again, efforts were made to make some thematic trials more difficult and others less difficult, by making the theme more or less obvious.

Categorial trials were trials in which the four objects that go together do so by virtue of belonging to a category that is defined neither by some visually perceptible feature nor by their being found together in a certain kind of common setting. Unlike geometric trials, the feature unifying four of the five items was not a perceptually salient feature. Nor, in contrast to thematic trials, was there a particular setting or scenario in which the four were typically found and the fifth was not. In some of the categorial trials, the relevant category was a commonly recognized taxonomic category, such as *living being*. The five images for the *living being* trial were of a tree, an elephant, a curved orange vase (the distractor), a fish, and a bird (Fig. 2). In some of the categorial trials, the categories were *functional* in nature, such as *energy source*. The five images for the *energy source* trial were of a tractor (the distractor), batteries, a solar panel, a windmill turbine, and a can of gasoline. Yet other categorial trials involved categories that might be described as *affordance-based*, such as *things that make a loud sound*. For the *things that make a loud sound* trial, the five pictured objects were: a stereo speaker, a rock (the distractor), a bird, a bicycle horn, and a dog. (Probably none of categories in the categorial trials should be described as merely *ad hoc* categories in the sense of Barsalou, 2003.) The goal in creating categorial trials was to force the participant to abstract-away from any visually salient perceptual similarity—be it a common shape, or frequent grouping in a setting one might see—in order to arrive at the property uniting the objects. We thereby sought to minimize chances that participants could answer the trial through the use of perceptual memory—e.g., visualizing a typical scene where the objects are found together—or through visually discriminating the shared feature.

See Appendix A for a complete list of the thematic and categorial categories used. The geometric categories are not included in this table, because they are often difficult to describe. Additional examples of the stimuli used, including the geometric stimuli, can be found in Appendix B.

2.7 Scoring

Participants were judged to have given a correct stage A response if they selected the four out of five images corresponding to the correct category (as defined by the experimenters) for that trial. In assessing stage A responses, we were assessing basic categorization abilities within each type of trial. To measure participants' abilities to assess their success in stage A, we looked at two main ratios. The first is the *reliability of subjective self-assessment* (RSSA). The RSSA is an assessment of the reliability of the subjective confidence levels reported by participants using the confidence faces. To calculate the RSSA for a particular type of trial, we summed the number of times that a participant both gave the correct stage A response and selected the smiling face with the number of times the participant both gave the incorrect stage A response and selected either the neutral or the frowning face, and divided that sum by 15 (as there were 15 trials of each type).[5] An ideal RSSA score of 1 would be received if every time the participant gave the correct categorization he or she reported high confidence (with the smiling face) and every time the participant miscategorized the stimuli in stage A he or she indicated a lack of full confidence (with a frowning or neutral face).

[5] In the notation we used, RSSA = (A1&H + A0&L)/15. Here "A1&H" stands for the number of times the participant correctly identified the four out of five and indicated high confidence by touching the smiling face, and "A0&L" stands for the number of times the participant incorrectly selected four objects and selected either the neutral face or the frowning face.

In thus calculating the RSSA score, we grouped the neutral face and the frowning face together as indicators of lack of high confience. So defined, the RSSA score serves as a measure of the degree to which a participant's high confidence (indicated with smiley face) reliably tracked his or her actual stage A success, and whether a participant's less than high confidence reliably tracked errors at stage A. The rationale for grouping the neutral face with the frowning face, as opposed to with the smiley face, is examined and further explained in the discussion section, below.

The second ratio we used to assess participants' metacognitive abilities is what we called the *reliability of active self-assessment* (RASA). This measure was inspired by the aforementioned animal metacognition studies (e.g., Smith et al., 2008). To calculate the RASA for a particular type of trial, we summed the number of times the participant both gave the correct stage A response and selected the green check (going forward to stage B) with the number of times the participant both gave the incorrect stage A response and selected the red X (opting out), and divided that sum by 15.[6] An ideal RASA score of 1 would be received for a type of trial if every time the participant correctly categorized at stage A he or she chose to go forward to stage B and every time the participant categorized incorrectly at stage A he or she opted out with the red X.

The RSSA and RASA are two different ways of measuring a participant's ability to assess his or her success in categorization at stage A. Because we were dealing with human participants with relatively intact language comprehension abilities (unlike the animal studies cited above), we could simply ask them to report their degree of confidence, using the

[6] In the notation we used, RASA = (A1&G + A0&R)/15. Here "A1&G" stands for the number of times the participant correctly identified the four out of five and touched the green check, and "A0&R" stands for the number of times the participant incorrectly selected four objects and selected the red X.

confidence faces. These responses were used to generate the RSSA. However, one might be skeptical of participants' abilities (or willingness) to accurately report their own confidence levels. For this reason, one might favor the RASA as a measure of ability to self-assess, as it measures this ability as a function of the participant's active betting behavior. It is because such betting games can be taught without using language that they are used to assess ability to self-assess in comparable experiments with animals. Even so, one might worry that RASA does not always track ability to self-assess, as some participants may, as a matter of personality, be less risk averse than others, or may simply be driven by curiosity to see the stage B slides; in that case, betting behavior will not line up neatly with ability to self-assess. Because both RSSA and RASA have advantages and disadvantages as measures of ability to self-assess, we thought it best to include both.

Whether participants in fact correctly selected the fifth matching object at stage B of a trial was not relevant to our calculations of metacognitive reliability, even if success at this stage was a main goal from the perspective of the participants. Success at stage B is also not a valid measure of a participant's categorization abilities, because it screens out all of those trials in which the participant decided, correctly or incorrectly, not to go on to stage B. For instance, someone might have high success at Stage B, even if she very often opts out when she should have gone forward (and therefore, often fails to know that she knows). Thus, mere success at Stage B has less theoretical interest than the more sensitive RSSA and RASA measures. The possibility of winning or losing tokens at stage B was included to provide participants motivation for carefully considering their confidence with respect to their categorizations at stage A.

3. Results

3.1. Silent Rhyming Task

An independent-samples *t*-test was conducted to compare silent rhyming task performance (i.e., hits) in controls and PWA conditions. This analysis revealed a significant difference between controls (mean = 37.18, SD = 2.04) and PWAs (mean = 20.78, SD = 4.24); $t(18) = 11.38$, $p < 0.001$ (see Figure). Note that a mean of 20 would be expected of a group that was simply guessing whether the words rhymed, for the 40 trials. Thus, despite the PWA generally scoring within normal limits on non-linguistic cognitive tasks of the CLQT (see Table 1b), there was a striking difference between the control participants and the PWA with respect to the silent rhyming task. By our operational definition of inner speech, the PWA were shown to have significant inner speech deficits compared to controls.

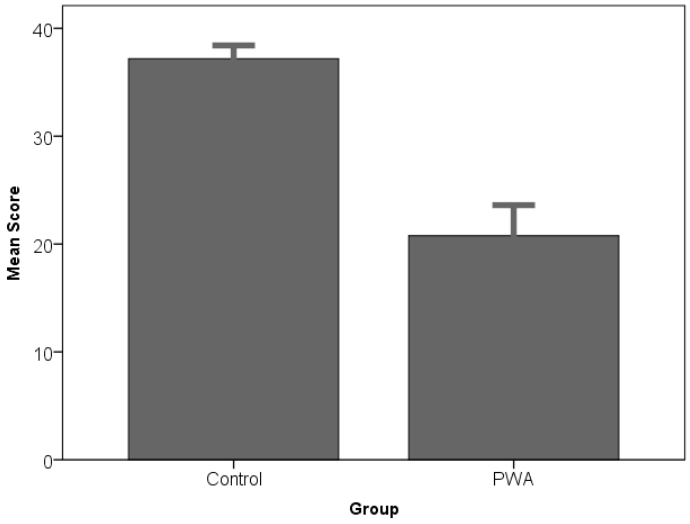


Figure 4: Number of hits on the silent rhyming task for controls and people with aphasia (PWA); error bars correspond to +/- 2 standard errors of the mean.

3.2 Correct responses in stage A

To determine whether there was a difference between controls and people with inner speech impairments in the number of times participants picked the correct four stimuli in stage A for the different stimulus conditions, a 2 (group: control, inner speech impaired) × 3 (condition: geometric, thematic, categorial) mixed design ANOVA with condition as the repeated measure was conducted. The analysis only revealed a significant main effect for condition, $F(2, 36) = 41.82$, $p < .001$, ${}_p\eta^2 = .699$. There was no main effect of group, $F(1, 18) = 1.75$, $p > .203$, ${}_p\eta^2 = .088$, nor was there an interaction between group and condition, $F(2, 36) = 1.21$, $p > .311$, ${}_p\eta^2 = .063$. As can be seen from an inspection of Figure 5a, the performance of the PWA was almost identical to that of controls. Bonferroni post hoc comparisons verified that both groups were significantly better at correctly matching the thematic stimuli in comparison to the geometric and categorial stimuli (both $p < .01$). There was no significant difference between the geometric and categorial stimulus conditions ($p > .11$).

3.3 Reliability of subjective self-assessment (RSSA)

The average RSSA score as a function of group and stimulus condition is presented in Figure 5b. An analysis of RSSA using a 2 (group) × 3 (condition) mixed design ANOVA resulted in a main effect for condition, $F(2, 36) = 8.32$, $p < .01$, ${}_p\eta^2 = .316$, and a significant interaction between group and condition, $F(2, 36) = 8.96$, $p < .01$, ${}_p\eta^2 = .332$. There was also a significant main effect of group, $F(1, 18) = 4.75$, $p < .043$, ${}_p\eta^2 = .209$. To better understand the interaction between group and condition, a means contrast analysis of group using between-subjects *t*-tests for the different stimulus conditions was conducted (adjusting degrees of freedom for unequal variances where necessary). This analysis found there to be no difference between the groups for the geometric, $t(18) = 1.09$, $p > .28$, and thematic conditions, $t(9) = 1.17$, $p > .29$

(with Levene's test indicating unequal variances, $F = 7.61$, $p = .013$), but found there to be a significant difference between the groups for the categorial stimulus condition $t(18) = 4.30$, $p < .01$. That is, the PWA performed significantly worse ($M_{PWA} = .70$, $SD_{PWA} = .12$) in the categorial stimulus condition compared to controls ($M_{Control} = .88$, $SD_{Control} = .07$). In contrast, the PWA performed similarly to controls in the geometric and thematic stimulus conditions.

3.4 Reliability of active self-assessment (RASA)

The average RASA score as a function of group and condition is presented in Figure 5c. Similarly to the results found for RSSA above, an analysis of participants' RASA scores using a 2 (group) × 3 (condition) mixed design ANOVA resulted in a significant main effect for condition, $F(2, 36) = 5.42$, $p < .01$, ${}_p\eta^2 = .231$, as well a significant interaction between group and condition, $F(2, 36) = 9.36$, $p < .01$, ${}_p\eta^2 = .342$. There was no main effect of group, $F(1, 18) = 3.49$, $p > .08$, ${}_p\eta^2 = .162$. Again, a means contrast analysis of group using between-subjects *t*-tests for the different stimulus conditions found there to be no difference between controls and PWA for the geometric, $t(18) = 1.79$, $p > .09$, and thematic conditions, $t(18) = 1.42$, $p > .17$, but a significant difference between the two groups for the categorial stimulus condition emerged, $t(11) = 3.39$, $p < .01$ (Levene's test indicating unequal variances, $F = 5.22$, $p = .035$). Specifically, the PWA were much less reliable in their active self-assessments for categorial trials ($M_{PWA} = .70$, $SD_{PWA} = .15$) compared to controls ($M_{Control} = .88$, $SD_{Control} = .07$).

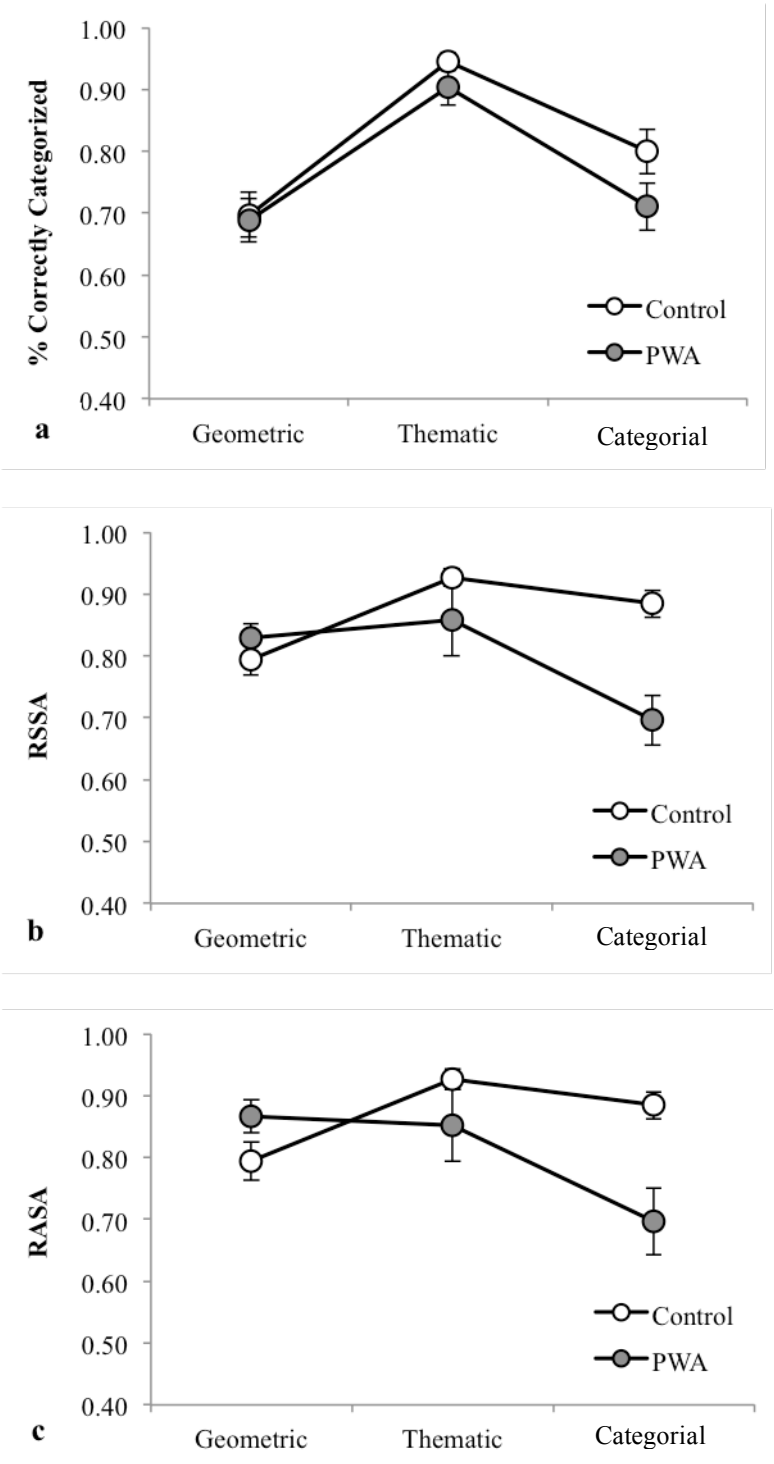


Figure 5. The average (a) A1, (b) RASA, and (c) RSSA score for control and PWA participants as a function of stimulus condition. Error bars indicate standard-errors from the mean.

3.5 Response times

The timing of participants in performing the different responsive actions was also recorded. Two measures of potential interest were *time to green* and *time to red.* Time to green is the amount of time between the start of a trial and the moment when the participant selected the green check. Time to red is the amount of time between the start of a trial and the moment when the participant selected the red X. The mean response times for these two measures, as a function of group and stimulus condition, are reported in Table 2. Note that for some control and PWA there were very few or no red-check responses for the various stimulus conditions. The difference in the overall mean response time of controls and PWA (averaged across response type and stimulus condition) was examined using a between subjects *t*-test. This analysis revealed that the PWA performed the task significantly more slowly than controls, $t(18) = -3.92$, $p < .01$. The table shows that for both groups and in all conditions time to red was longer than time to green.

Table 2. *Mean response time in seconds as a function of group, stimulus condition and response type. (Standard deviations are reported in parenthesis.)*

	Time to green			Time to red		
	Geometric	Thematic	Categorial	Geometric	Thematic	Categorial
Controls	10.64 (3.98)	11.01 (2.92)	12.12 (3.16)	20.02 (8.65)	20.25 (6.53)	30.27 (17.13)
PWA	18.96 (6.59)	17.99 (4.80)	22.75 (7.92)	40.63 (12.38)	32.64 (9.39)	44.95 (17.17)

4. Discussion

As expected, we found that the overt language production impairments of our PWA population were mirrored by corresponding inner speech impairments, as assessed by the silent rhyming task. With respect to the main experimental task, we then found two results of interest. First, the participants with inner speech impairments did not show deficits, relative to controls, in their ability to select the four objects that go together in stage A of the main task. This was true for all three types of trial: geometric, thematic and categorial. Thus, we did not find evidence of an association between inner speech impairment and impairments in categorization for any of the specific kinds that we tested. Yet, even supposing that the participants with inner speech impairments were as successful as controls in categorization, we cannot conclude that they identified the four that go together *in the same way* as controls. The longer reaction times of the PWA, even in the geometric and thematic trials, suggest that they may have developed compensatory strategies as a result of their stroke. These response time results conform to a general pattern (e.g. (Purdy, 2002; Murray, Holland & Beeson, 1999)), in which PWA perform similarly to controls in terms of the number of correct answers they give on non-verbal experimental tasks, while taking more time to do so.

Second, in the geometric and thematic trials, the PWA performed as well as the controls in assessing their categorizations in stage A. This was confirmed both by the subjective (RSSA) and active (RASA) measures of reliability of self-assessment. However, in the categorial trials, the PWA performed significantly worse than the controls in assessing their categorizations in stage A, as judged by both the RSSA and RASA measures. Interestingly, whatever ability it is that enables the PWA to categorize in stage A as well as the control participants, that ability does not appear to ensure equal performance in assessing their own success in the categorial trials.

The ability to categorize correctly, and the ability to accurately assess one's categorization abilities appear to draw on different cognitive resources—at least with respect to categorial trials.

While we did not find evidence for the hypothesis that inner speech impairment leads to deficits in metacognitive self-assessments categorization abilities *in general*, our results offer some preliminary support for the hypothesis that inner speech facilitates metacognitive self-assessments when they concern categorial kinds. The evidence is only preliminary due to the relatively small sample size, and the potential confounds inherent in working with a non-neurotypical population. However, steps were taken to confirm that the PWA were within normal limits on non-linguistic cognitive tasks.

Further, the fact that the PWA showed metacognitive deficits only on *one* of the three trial types calls for a more precise explanation than a broad appeal to their atypical neurological condition. Moreover, there were no significant, or near-significant, correlations between participant scores for RASA or RSSA on categorial trials and their performance on CLQT sub-tests assessing attention, executive control, and visuospatial reasoning.[7] Nor do the reaction time data offer a basis for an explanation of the data; for the slower overall performance of the PWA group on *every* kind of trial does not explain their showing metacognitive differences *only* on categorial trials.

With respect to the RSSA scores, it may be objected that, instead of grouping neutral face responses with frowning face responses as indications of a lack of confidence, we could have alternatively grouped neutral face with smiley face responses as indications of positive confidence. To consider this objection concerning the proper interpretation of neutral face responses, we calculated a second measure—RSSA2—which assigns a value of .5 to every trial

[7] These correlations were assessed using both Spearman's rank order correlations and Pearson's correlations. For all rank order correlations, $r < .300$ and $p > .44$. For all Pearson correlations, $r < .360$ and $p > .35$.

where a participant selected a neutral face, regardless of whether the stage A response was correct or incorrect. And (as before), a value of 1 was assigned to every trial where a participant both answers stage A correctly and selects the smiley face, or answers stage A incorrectly and selects the frowning face.[8] By this calculation, a neutral face response is never worth as much as a smiley face response when the participant is correct at stage A, nor as much as a frowning face when the participant is wrong at stage A. At the same time, by this measure, a neutral face response is still worth more than a smiley face response when the stage A response was incorrect, or a frowning face when the stage A response was correct (both of which are worth 0). Independent-samples t-tests were conducted to compare performance on this new measure ("RSSA2") between PWA and controls on different trial types. Echoing the RSSA scores reported above, there were significant differences in RSSA2 between controls ($M_{Control}$ = .87, $SD_{Control}$ = .07) and PWA (M_{PWA} = .72, SD_{PWA} = .10) on categorial trials, $t(18) = 3.99$, $p < .01$. And there were no significant difference in RSSA2 between controls and PWA on geometric trials ($p > .20$) or thematic trials ($p > .10$).

What might explain the association between the inner speech impairment and the deficit in metacognition on categorial trials? Taking a step beyond our operational definition of inner speech, one might conceive of the deficit in inner speech as involving an inability to experience auditory imagery of words. As noted in section 1.2, this inner speech deficit might be due to an impairment of any of several stages in the production or inner comprehension of the auditory

[8] In the notation we used, RSSA2 = (A1&S + .5N + A0&F)/15. Here "A1&S" stands for the number of times the participant correctly identified the four out of five and indicated high confidence by touching the smiling face, "N" stands for the number of times the participant selected the neutral face (regardless of whether they the participant was correct at stage A), and "A0&F" stands for the number of times the participant incorrectly selected four objects and selected the frowning face. The denominator is 15 because RSSA2 (like RASA and RSSA) was calculated separately for each trial type, there being 15 trials of each type. This enabled a perfect RSSA2 score of 1 if every time the participant was correct at stage A she/he selected the smiley face, and if every time the participant was incorrect at stage A she/he selected the frowning face.

imagery of words. Given such a deficit, we might entertain the following hypothesis: In non-impaired participants an episode of auditory verbal imagery may serve as a cue that one has successfully identified the four items that go together. For example, hearing the phrase "living beings" or "things with a handle" in auditory imagery might indicate that one has found the criterion by which the four that go together are distinguished. So a possible reason why inner speech impaired PWA are not as reliable as control participants in assessing their success in categorial trials is that they cannot experience this auditory imagery.

Still to be explained, however, would be why the metacognitive deficits of the PWA pertained only to the categorial trials. Here a possible answer is that, for the geometric trials and the thematic trials, something other than a verbal label in auditory imagery is available to play the role of the mental cue indicating success. In the case of the geometric trials, the cue might be a mental image of a representative member of the four that go together. In the case of the thematic trials, the cue might be a visual image of a scenario in which the four objects might typically be seen. Since the PWA were not impaired in their ability to generate such cues, they could potentially exploit such visual cues to perform as well as controls in assessing their success in geometric and thematic trials.

A hypothesis along these lines leaves several questions open about the nature of the metacognition involved. Supposing that it is the presence or absence of a cue such as we have described that enables successful assessment, one could offer various accounts of how the cue functions. If we think of the cue as "telling" the thinker something, then we could give various accounts of the content of the message, ranging from a highly self-reflective content, such as *I am confident that those four are all X's*, to a very minimal content, such as *That was easy* (cf., Perner, 2012). (It is well known that fluency or ease of processing is often used by humans as a cue for cognitive success. Compare Proust, 2013; Undorf & Erdfelder, 2011; Alter &

Oppenheimer, 2009; Koriat, 1997.) However, if the presence of the cue serves as an indicator of fluency, then the indicator of fluency is not the participant's attention to an external feature of the stimulus, such as its size or volume. This marks an interesting difference with the kinds of fluency-modulating factors explored by others (see, e.g., Kornell, Rhodes, Castel, & Tauber, 2011, and Rhodes & Castel, 2009).

5. Future research

Suitable test participants for the present study were limited to individuals with expressive language production challenges. Consequently, only nine such participants could be recruited. Accordingly, it would be useful to find other ways of investigating the role of inner speech in similar tasks. In a larger non-patient population, for instance, one could look for correlations between the strength of their dispositions toward inner speech and the reliability of their metacognitive self-assessments. Dispositions toward inner speech could be measured using the Descriptive Experience Sampling paradigm developed by Hurlburt and colleagues (Heavey & Hurlburt, 2003; Hurlburt & Akhter, 2006), and then compared to performance on metacognitive tasks of the kind employed here. Or one could use a verbal interference paradigm (Lupyan, 2009; Gilbert, Regier, Kay, & Ivery, 2006) with a healthy population to see whether taxing covert language resources impairs concurrent categorization or metacognitive performance in neurotypical participants. (During verbal interference, participants are required to repeat a string of words aloud (e.g. reciting the days of the week) while completing a primary task.) This would help confirm that the differences observed here did not pertain to some aspect of the PWA's condition other than their inner speech impairment.

Finally, it would be of interest to use a similar betting paradigm—with either people with aphasia or healthy controls under overt speech verbal interference—to look at the role of inner

speech in other kinds of metacognitive tasks. For instance, in "metamemory" tasks (Dunlosky & Bjork, 2008) participants are asked to judge the likelihood that they will remember a certain piece of information. (For example, they may be told to remember a seven digit numeral. Five minutes later, they may then be asked if they will be able to identify that numeral on a list of five similar seven digit numerals that are about to be shown. Betting behavior and opting out behavior can again be taken as an indication of metacognitive confidence.) In gaining a more detailed picture of the kinds of metacognitive tasks where inner speech makes a contribution, we may come to better understand *why* it makes that particular contribution.

Acknowledgements: Jonathan Martin, Kristen Grevey, and Heather Bolan provided administrative assistance on this project.

Funding: The work of Frank Faries and Peter Langland-Hassan was supported in part by a grant from the John Templeton Foundation, via a sub-grant from the New Directions in the Study of Mind project at Cambridge University.

APPENDIX A:

Complete list of thematic and categorial categories, excluding warm-up trials, with one example from each.

Thematic categories	Categorial categories
baby items (pacifier)	game (playing cards)
fishing (fishing rod)	musical instrument (saxophone)
picnic (cooler)	made of paper (paper plate)
beach (flip flops)	weapon (hand grenade)
gardening (watering can)	fruit (pear)
birthday party (balloons)	has a handle (brief case)
carpentry (nails)	living being (tree)
house cleaning (sponge)	power source (windmill turbine)
baseball game (hot dog)	sport ball (baseball)
living room (sofa)	lens (binoculars)
dog items (dog collar)	root vegetable (carrots)
kitchen (frying pan)	makes a noise (bicycle horn)
rain (umbrella)	liftable (electric toaster)
highway (empty billboard)	transparent (glass of wine)
lawn care (lawn mower)	tool (screwdriver)

APPENDIX B:

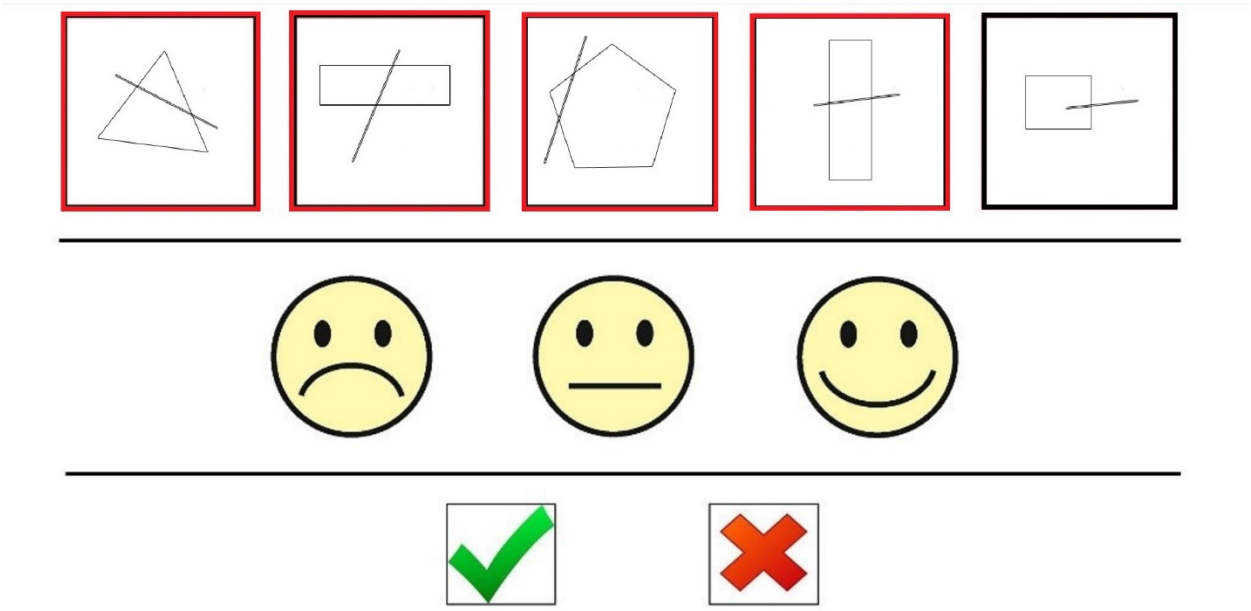


Figure 1: Geometric Trial 109, Stage A

APPENDIX B (continued)

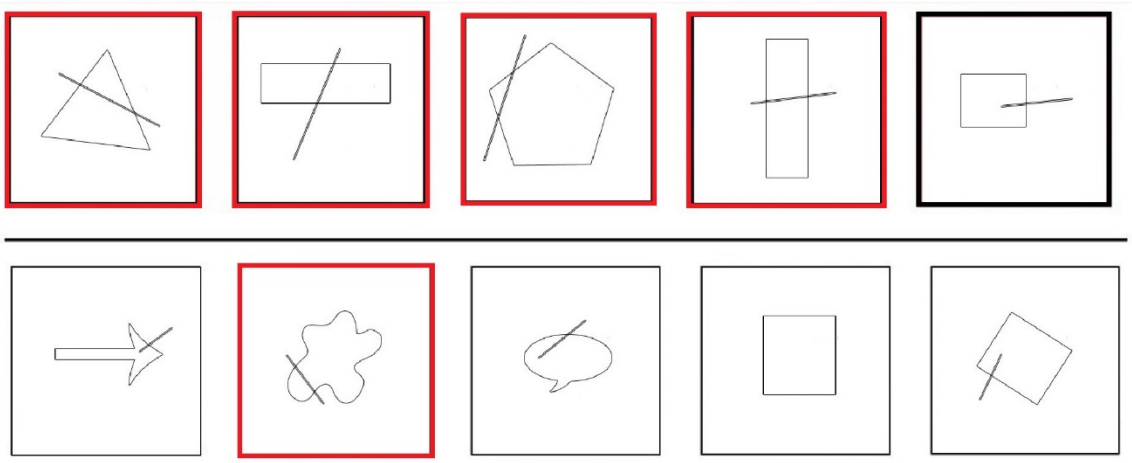


Figure 2: Geometric Trial 109, Stage B

APPENDIX B (continued)

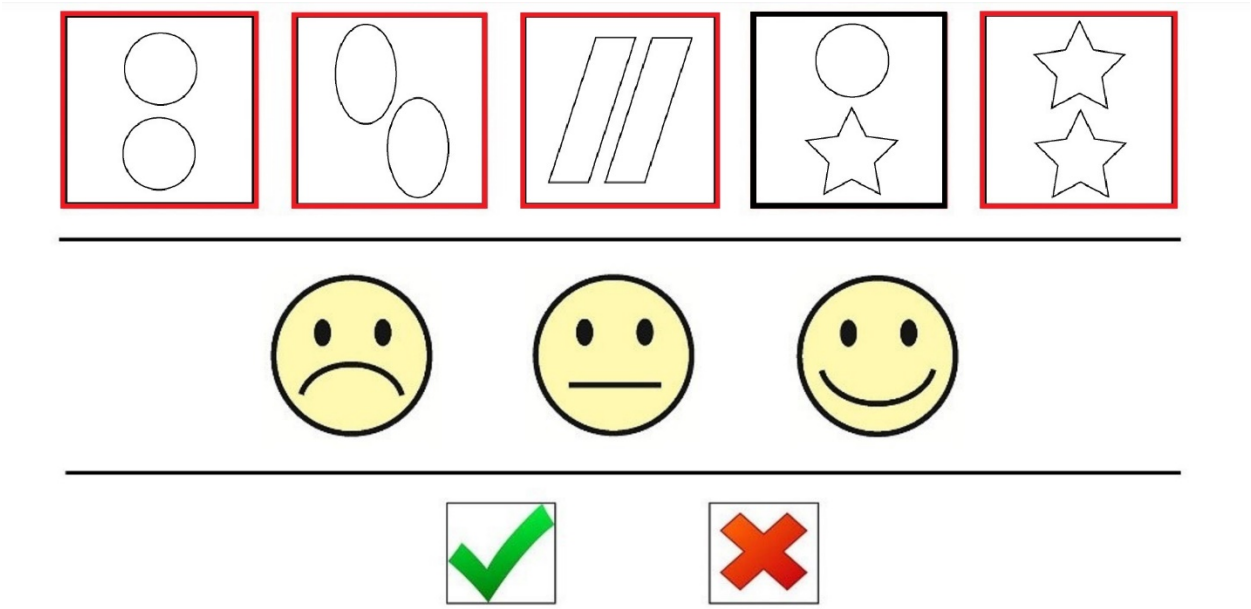


Figure 3: Geometric Trial 110, Stage A

APPENDIX B (continued)

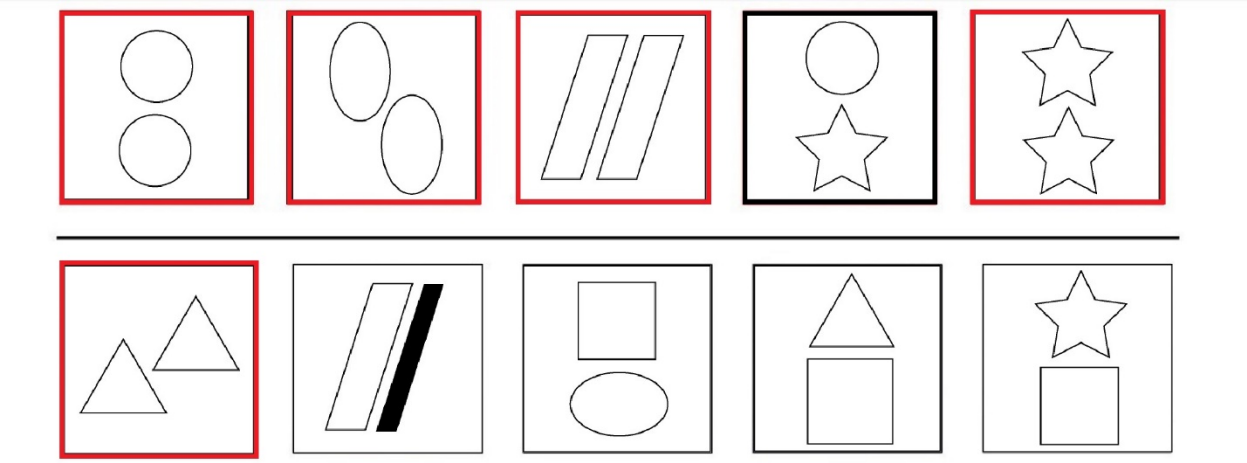


Figure 4: Geometric Trial 110, Stage B

APPENDIX B (continued)

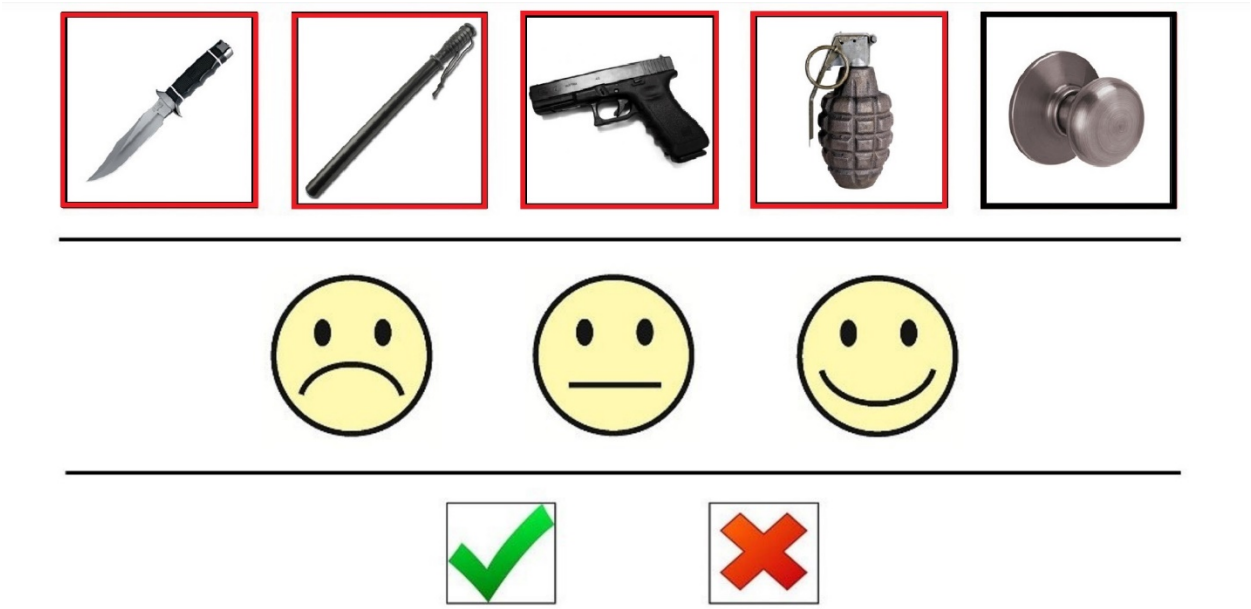


Figure 5: Categorial Trial 201 ("Weapon"), Stage A

APPENDIX B (continued)

Figure 6: Categorial Trial 201 ("Weapon"), Stage B

APPENDIX B (continued)

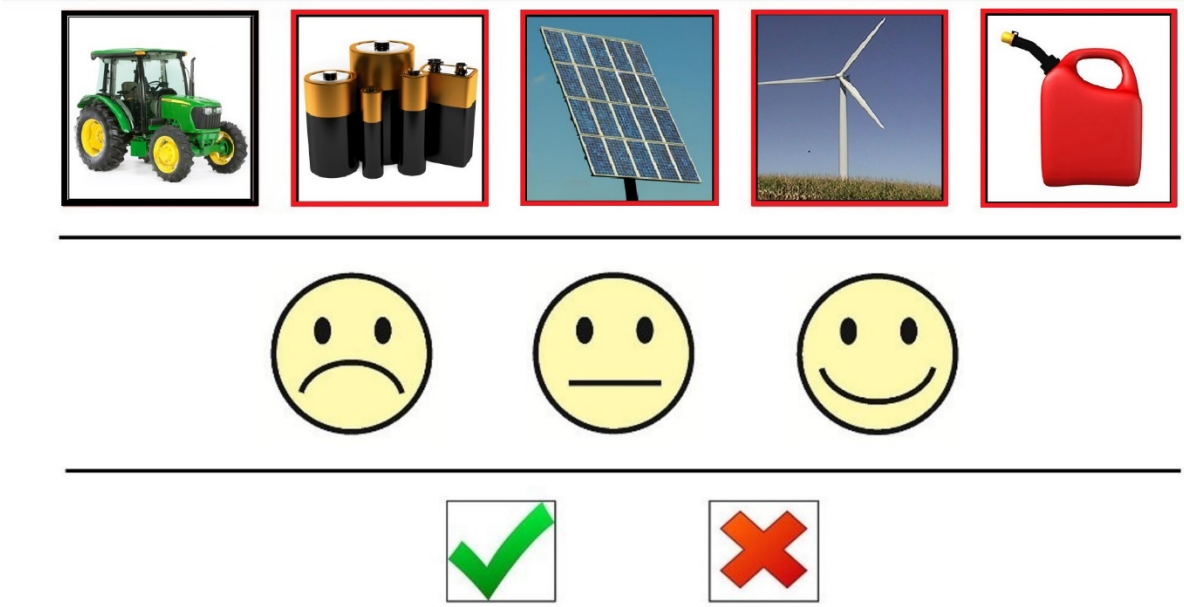


Figure 7: Categorial Trial 207 ("Energy Source"), Stage A

APPENDIX B (continued)

Figure 8: Categorial Trial 207 ("Energy Source"), Stage B

APPENDIX B (continued)

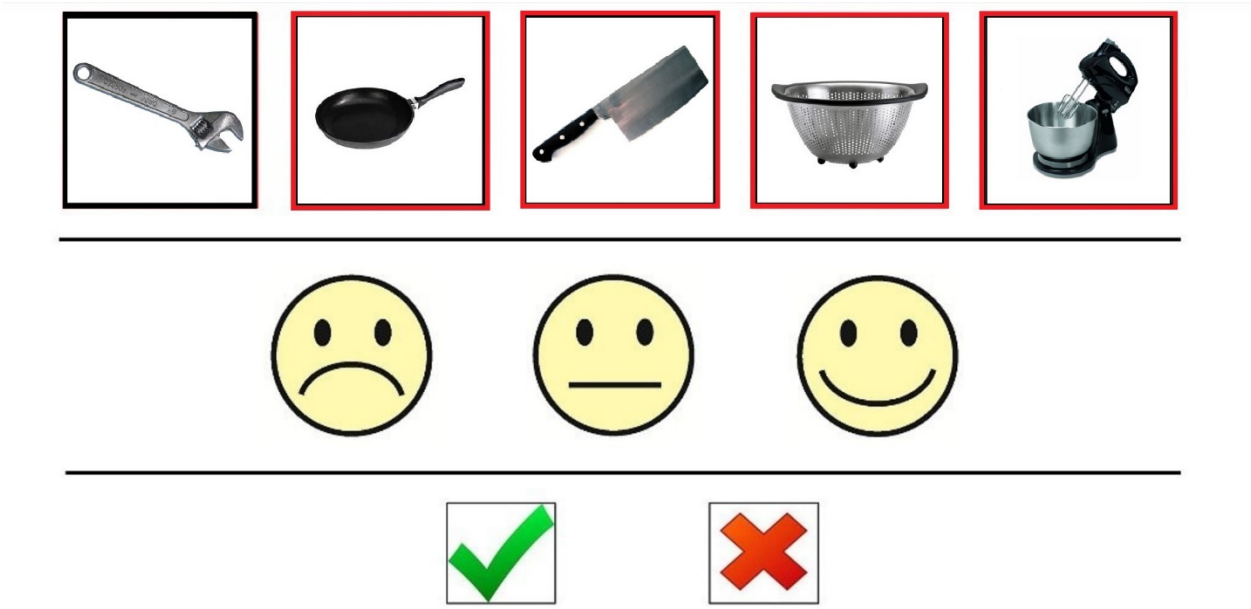


Figure 9: Thematic Trial 306 ("Kitchen"), Stage A

APPENDIX B (continued)

Figure 10: Thematic Trial 306 ("Kitchen"), Stage B

APPENDIX B (continued)

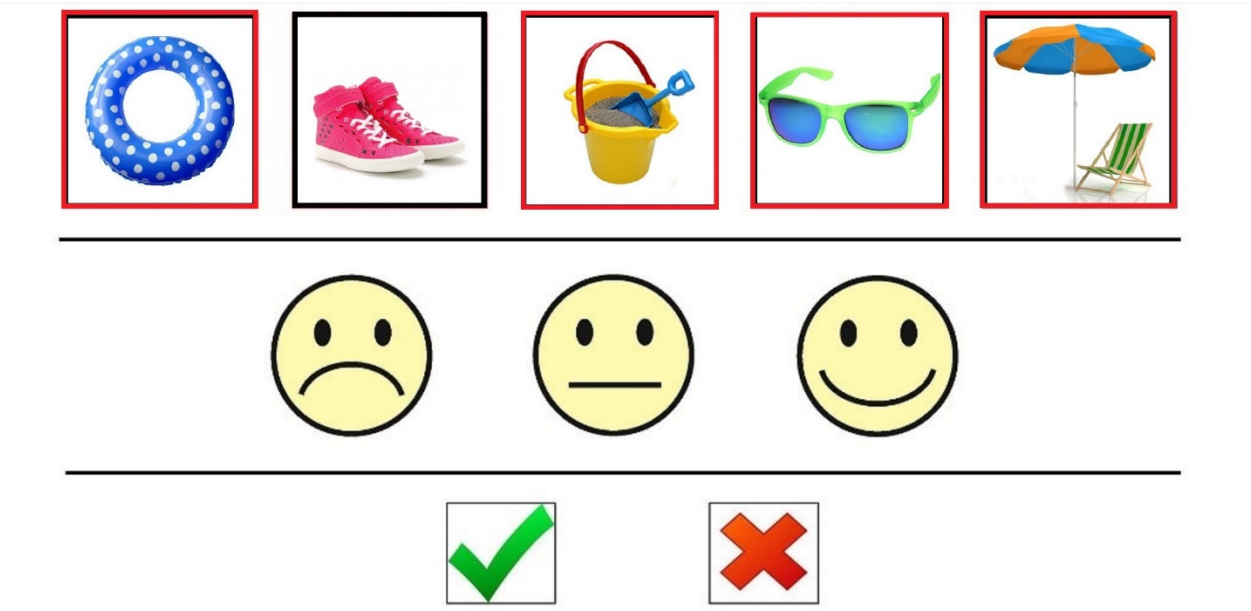


Figure 11: Thematic Trial 314 ("Beach"), Stage A

APPENDIX B (continued)

Figure 12: Thematic Trial 314 ("Beach"), Stage B

References

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

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13(3)*, 219–35.

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes* 18: 513-562.

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

Carruthers, P. (1996). *Language, thought, and consciousness*. (Cambridge, Cambrdige University Press).

Carruthers, P. (2008). Metacognition in animals: A skeptical look. *Mind and Language, 23(1)*, 58-98.

Carruthers, P. (2011). *The Opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Caspari, I., Parkinson, S.R., LaPointe, L.L., Katz, R.C. (1998). Working memory and aphasia. *Brain and Cognition*. 37, 2: 205-223.

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

Cohen, R., Kelter, S., & Woll, G. (1980). Analytical competence andlanguage impairment in aphasia. *Brain & Language*, *10*, 331-347.

Dunlosky, J., Bjork, R.A. (2008). *Handbook of Metamemory and Memory*. Taylor & Francis: New York.

Fama, M. E., Hayward, W., Snider, S. F., Friedman, R. B., & Turkeltaub, P. E. (2017). Subjective experience of inner speech in aphasia: Preliminary behavioral relationships and neural correlates. *Brain and Language, 164*(Supplement C), 32-42. doi:https://doi.org/10.1016/j.bandl.2016.09.009

Feinberg, T. E., Gonzalez Rothi, L. J., & Heilman, K. M. (1986). "Inner speech"s in conduction aphasia. *Archives of Neurology, 43*, 591-593.

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

Frith, C. D., (1992). *The Cognitive Neuropsychology of Schizophrenia.* Psychology Press.

Geva, S., Bennett, S., Warburton, E.A., & Patterson, K. (2011). Discrepancy between inner and overt speech: Implications for post-stroke aphasia and normal language processing. *Aphasiology*, *25(3)*, 323-343.

Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, *103*, 489-494.

Glosser, G. & Goodglass, H. (1990). Disorders in executive control functions among aphasic and other brain-damaged patients. *Journal of Clinical and Experimental Neuropsychology.* 12,4: 485-501.

Hampton, R. (2001). Rhesus monkeys know when they remember, *Proceedings of the National Academy of Sciences of the United States of America* 98: 5359– 5362.

Heavey, C.L. & Hurlburt, R.T. (2008). The phenomena of conscious experience. *Consciousness and Cognition.* 17: 798-810.

Helm, N. (2003). *Cognitive linguistic quick test*. Pro-Ed, Incorporated.

Helm-Estabrooks, N. (2002). Cognition and aphasia: a discussion and a study. *Journal of Communication Disorders.* 35: 171-186.

Hinckley, J. & Nash, C. (2007). Cognitive assessment and aphasia severity. *Brain and Language,* 103: 8-249.

Hurlburt, R.T. & Akhter, S.A. (2006). The descriptive experience sampling method. *Phenomenology and the Cognitive Sciences*. 5 (3-4): 271-301.

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

Kertesz, A. (2006). *Western aphasia battery-revised (WAB-R)*. Austin, TX, Harcourt.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370.

Kornell, N., Rhodes, M.G., Castel, A.D., & Tauber, S.K. (2011). The ease of processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22(6)*, 787-794.

Kornell, N., Son, L., & Terrace, H., (2007). Transfer of metacognitive skills and hint-seeking in monkeys. *Psychological Science, 18*, 64-71.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797-2822.

Langland-Hassan, P., (2008). Fractured phenomenologies: Thought insertion, inner speech, and the puzzle of extraneity. *Mind and Language*, *23(4)*, 369-401.

Langland-Hassan, P., Faries, F.R., Richardson, M.J. & Dietz, A. (2015). Inner speech deficits in people with aphasia. *Frontiers in Psychology*, 6: 528. *doi: 10.3389/fpsyg.2015.00528*

Laurent, L., Millot, J.-L., Andrieu, P., Camos, V., Floccia, C., & Mathy, F. (2016). Inner speech sustains predictable task switching: direct evidence in adults. *Journal of Cognitive Psychology, 28*(5), 585-592. doi:10.1080/20445911.2016.1164173

Levine, D. N., Calvano, R., & Popovics, A. (1982). Language in the absence of inner speech. *Neuropsychologia, 20*(4), 391-409.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, **50**, 315-353. doi:10.1016/j.cogpsych.2004.09.004

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

Lupyan, G., & Mirman, D. (2013). Linking language and categorization: Evidence from aphasia. *Cortex*, 49, 1187-1194.

Martínez-Manrique, F & Vicente, A. (2015). The activity view of inner speech. *Frontiers in Psychology*, *6(232)*. doi: 10.3389/fpsyg.2015.00232

Morin, A. (2009). Self-awareness deficits following loss of inner speech: Dr. Jill Bolte Taylors case study? *Consciousness and Cognition* 18, 524–529. doi: 10.1016/j.concog.2008.09.008

Murray, L.L. (1999). Attention and aphasia: theory, research and clinical implications. *Aphasiology.* 13, 2: 91-111.

Murray, L.L. (2012). Attention and other cognitive deficits in aphasia: presence and relation to language and communication measures. *American Journal of Speech-Language Pathology.* 21: S51-S64.

Murray, L.L., Holland, A.L. & Beeson, P.M. (1997). Auditory processing in individuals with mild aphasia. *Journal of Speech, Language, and Hearing Research.* 40: 792-800.

Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: Adults fail nonverbal false-belief reasoning. *Psychological Science*, *18*, 574-579. doi:10.1111/j.1467-9280.2007.01942.x

Noppeney, U., & Wallesch, C. (2000). Language and cognition—Kurt Goldstein's theory of semantics. *Brain and Cognition*, *44*, 367-386.

Oppenheim G.M. & Dell G. S. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory and Cognition 38*, 1147-60.

Papafragou, A., & Selimis, S. (2010). Event categorisation and language: A cross-linguistic study of motion. *Language and Cognitive Processes, 25*, 224-260.

Papafragou, A., Hulbert, J., and Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition* 108, 155–184. doi:10.1016/j.cognition.2008.02.007

Perner, J. (2012). MiniMeta: In search of minimal criteria for metacognition. In M. J. Beran, J. Brandl, J. Perner & J. Proust (Eds.), *Foundations of Metacognition* (pp. 94-116). Oxford: Oxford University Press.

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. *Behavioral Brain Research, 261*(Supplement C), 220-239. doi:https://doi.org/10.1016/j.bbr.2013.12.034

Pickering, M.J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*. 36: 329-347.

Plunkett, K., Hu, J., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*, 665-681. doi:S0010-0277(07)00108-4

Purdy, M. (2002). Executive function ability in persons with aphasia. *Aphasiology.* 16, 4-6: 549-557.

Proust, J. (2013). *The Philosophy of Metacognition: Mental Action and Self-Awareness.* Oxford, Oxford University Press.

Rhodes, M.G. & Castel, A.D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16(3)*, 550-554.

Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review, 15*(4), 679-691.

Smith, J. D., Shields, W. E., Schull, J., & Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, *62*, 75-97.

Smith, J.D. & Washburn, D.A. (2005). Uncertainty monitoring and metacognition by animals. *Current Directions in Psychological Science*, *14(1)*, 19-24.

Stark, B. C., Geva, S., & Warburton, E. A. (2017). Inner speech's relationship with overt speech in poststroke aphasia. *Journal of Speech, Language, and Hearing Research*, 1-10. doi:10.1044/2017_JSLHR-S-16-0270

Studtmann, P. (2017). Aristotle's categories. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, URL = https://plato.stanford.edu/archives/fall2017/entries/aristotle-categories/ (Last accessed on 15.09.2017.)

Ullman, M.T., Pancheva, R., Love, T., Yee, E., Swinney, D., & Hickok, G. (2005). Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain and Language*, 93(2), 185-238.

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. Journal of Experimental Psychology: *Learning, Memory, and Cognition*, *37(5)*, 1264–9.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*, 223-250.