Semantic Integration as a method for investigating concepts

Derek Powell, Zachary Horne, and N. Angel Pinillos

I. Introduction

The last ten years in philosophy have been marked by widespread interest in the psychology of philosophy. Much of this work has been carried out by Experimental Philosophers, who aim to better understand the contours of philosophical concepts and intuitions by importing the methods of the empirical sciences. Their hope is that through a better understanding of the psychology of philosophical concepts such as KNOWS, GOOD, FREEDOM and many others, we will be better able to assess philosophical arguments which utilize such notions.[1]

Experimental philosophers have amassed many interesting results, but compelling concerns have been raised about the survey methods that they typically employ. Here we argue, on the basis of these criticisms and our own, that the possibility of experimental artifacts is good reason to adopt a new experimental paradigm that we call Semantic Integration. This methodology uses a memory task as an implicit measure of the degree to which different situations instantiate concepts. This measure avoids the methodological problems researchers must address if they are to continue to use surveys.

The plan of the paper is as follows: First, we consider objections to survey methodology (section II) and then we describe how Semantic Integration tasks can be used to examine people's concepts (section III). Next, we argue that because semantic integration is an implicit measure it has important advantages over more explicit survey methods (section IV). Finally, we discuss caveats regarding the method (section V), variations on it, and briefly consider how it might be used alongside survey-based research (section VI).

II. The Experimental Methods of Experimental Philosophy

Studies in philosophy differ significantly in their details, but nearly all experimental philosophers employ the same experimental paradigm. Experimental philosophers present their participants with short passages and afterwards ask them questions about what they have read. These passages are designed to test whether certain features are parameters for instantiating a philosophical concept. The experimental case of interest is often derived from a philosophical thought experiment and is (ideally) paired with two control cases: one that clearly does not instantiate the concept, and one that clearly does. Then, researchers ask their participants to judge whether the concept applies using explicit survey questions, and compare participants' responses in the experimental condition(s) with the control cases. Studies using these methods have improved substantially when compared to early research that lacked proper control conditions,

[1] We follow tradition in using caps to denote concepts.

but the survey paradigm is still limited in important ways. In this section we review some of the problems associated with surveys and the challenges faced by researchers that use surveys. We discuss the issues raised by Cullen (2010), as well as other other limitations of survey methodology. In our view, the issues we raise warrant the adoption of different experimental methods that avoid these criticisms.

a. Pragmatic cues in experimental materials

In a recent critique, Simon Cullen (2010) discusses problems with survey-based investigations (Also see Clark & Schober, 1992; Schaeffer & Presser, 2003; Schwarz, 1994). In his discussion, Cullen stresses the importance of the semantic and pragmatic features of experimental materials. His overarching claim is that when people attempt to comprehend some utterance of natural language, they do not simply attempt to simply comprehend the exact meaning of the words as spoken or as written on the page. Rather, they attempt to comprehend the speaker's meaning, or the experimenter's meaning.

According to Grice (1975), people accomplish this by making a set of assumptions about the requirements for rational communication. These assumptions, sometimes referred to as Gricean norms, allow listeners or readers to grasp what a speaker means to convey or what they think speakers mean to convey. People assume that speakers are "cooperative communicators"--that their utterances are true, orderly, relevant, and non-redundant. Typically, speakers are themselves aware that their interlocutors make said assumptions and so they exploit these assumptions to get conversational participants to understand them. Sometimes the best way of making sense of someone's behavior, given that they are following Gricean norms, is by inferring that they mean something that goes beyond what they said or stated with their utterance. For instance, if Speak asks, "Has the number two bus come by yet?" a listener, Hear, can rightly infer that there is a number two bus, that its route passes by this location, and that Speak is hoping to catch the bus. Of course, none of these facts are explicitly stated in Speak's question. Hear can infer these things because she assumes that Speak is following the Gricean norms. For example, Speak would not be following the norm of relevance if he was not planning to get in the bus. Moreover, Speak can count on Hear to infer these things about his utterance because Speak knows that Hear will assume Speak is following the Gricean norms. This process normally leads to the computing of a "conversational implicature". Roughly, the propositions that the speaker means to convey (but go beyond what is said) and are inferable in a conversation applying the Gricean norms, are called "conversational implicatures" (as opposed to propositions that are conventionally associated with the words used). It is widely accepted that the deployment and computing of conversational implicatures is pervasive in human communication. For this reason, Cullen (2010) argues that if experimental philosophers ignore conversational implicatures, then their instructions, stimuli, response options, and other experimental materials may not convey the meanings they intend.

For an illustration of how conversational implicatures can affect survey results, consider research on base-rate neglect (Kahneman & Tversky, 1973). This is the tendency for people to ignore relevant statistical base-rates when judging the probabilities of events and to instead rely on heuristics (for a review, see Nisbett & Ross, 1980). In one study, researchers presented people with a description of a fictional college student and asked them to estimate the probability that the student majored in various fields. If the descriptions include traits that seem stereotypical of an engineering student (e.g., introverted, enjoys solving problems), then people estimated that it is highly probable that the student is an engineer. Researchers found that people make similar probability estimates even when they are told that only a small percentage of students study engineering. This finding showed that people ignore base-rate information when making their probability estimates and instead employ a representativeness heuristic: since the student resembles an engineer, they judge that it is probable he is one, and they ignore the base-rate information which would suggest that any individual student is most likely not an engineer (Kahneman & Tversky, 1973).

However, more recent research suggests that base-rate neglect may be due, at least in part, to conversational processes rather than decision processes. If participants assume that experimenters are cooperative communicators, then they assume that the information they've been given is the most relevant to the task at hand. This may lead them to place a greater weight on the descriptions given than they would have otherwise. Schwarz and colleagues (1991) explored this prediction by examining base-rate neglect while manipulating this guarantee of relevance. Participants in one condition were told that the descriptive information presented to them had been compiled by psychologists (as in the original experiments of Kahneman & Tversky, 1973), while in the other condition they were told that the same description had been compiled by a computer that randomly sampled from a database of information. Whereas communication from another person comes with an implied guarantee of relevance, computer-generated text provides no such guarantee. As predicted, researchers found that participants were significantly less influenced by computer-generated descriptions than human-generated descriptions (Schwarz et al., 1991). What's more, relevance effects can be quite subtle: even the order that information is presented can affect how participants interpret it. In particular, people place greater weight on the last source of evidence they are shown. For example, Krosnick and colleagues (1991) found that base-rates had a stronger effect on participants' judgments when they were the last piece of information participants read before making their response. The guarantee of relevance suggests that if experimenters present an apparently sufficient source of evidence (e.g., base-rate information), and then present another source of evidence (e.g., an individual description), then people consider this second source as non-redundant and thus highly relevant to the task at hand. In general, even relatively subtle pragmatic cues can have important effects on people's responses to survey questions. Perhaps unsurprisingly, Cullen

(2010) demonstrated that pragmatic cues can likewise affect people's responses to philosophical thought experiments.

Cullen argues that researchers can overcome these problems if they are sensitive to the context in which participants interpret their experimental materials, and the norms that govern these interpretations. Following Schwarz (1994), he argues that experimenters and participants are engaged in a conversation governed by the norms of cooperative communication (Cullen, 2010; Grice, 1975; Schwarz, 1994). Since participants abide by these norms, and expect researchers to abide by them as well, experimental materials must be constructed with pragmatic cues in mind. While addressing the pragmatic features of experimental materials would improve the conclusions that can be drawn from surveys, the problems faced by survey-researchers are more challenging than Cullen (2010) claims. As researchers, we still need to determine exactly how materials and questions ought to be phrased, and what implicatures they ought to contain. To make matters more challenging, this would need to be determined for each concept that experimental philosophers intend to examine.

Consider the concerns that cognitive psychologists faced when studying causal learning: An important construct in research on causal learning is *causal strength*, normatively defined as the probability that some cause produces an effect (Cheng, 1997). Although people often make judgments about causal strength, researchers could ask participants to report such a judgment in any number of ways, and it is not obvious which way is optimal. Keeping things as simple as possible, in one experiment Buehner and colleagues (2003) asked their participants to make a rating on a scale from 0 (X does not cause Y at all) to 100 (X causes Y every time). Interestingly, they found that participants' judgments tended to cluster into two groups: one group of participants made judgments consistent with the normative definition of causal strength, whereas the others made non-normative judgments consistent with simpler heuristic algorithms. As causal learning is often taken to be a relatively basic cognitive mechanism (Cheng, 1997), it would be remarkable if certain people learned causal relationships via different cognitive mechanisms. However, while the finding that some people make non-normative causal strength judgments could indicate that they utilized a heuristic algorithm in computing causal strength, it could also be due to ambiguities in the question used to probe their judgments. Indeed, Buehner and colleagues note that their causal strength question can be interpreted as applying in two different contexts: 1) the experimental learning context where the effect is also produced by other background causes or 2) a counterfactual context where only the cause of interest is present. Their causal strength question only asks for a true causal strength judgment under the second interpretation. Buehner and colleagues concluded that participants had interpreted the question differently.

In light of their findings, they concluded that the best way to eliminate ambiguity in the causal strength question was to phrase it counterfactually and in terms of frequencies (Buehner,

Cheng, & Clifford, 2003).[2] This wording makes the context clear, and participants do not need to make any inferences beyond what is stated. Importantly, in developing this causal strength question, researchers were guided by a detailed model that defined the construct (Cheng, 1997). They were able to conclude that their question was adequate when it lead participants to give normative answers--owing to the fact that people are capable of making such judgments. In contrast, experimental philosophers will rarely have clear definitions of their constructs to aid in crafting their questions and materials because such definitions are precisely what they hope to discern through experimentation.

b. Demand Characteristics

There is another set of problems that can affect the results of surveys, often referred to as *demand characteristics* (Orne, 1962). Crudely put, demand characteristics are artificial features of an experimental task that lead participants to perform some task other than what the researchers intended them to.[3] Demand characteristics can occur when participants are apprehensive about being evaluated (Weber & Cook, 1972). Apprehension can lead participants to respond in ways they perceive as either socially desirable, or "correct," irrespective of their actual attitudes or intuitions. Demand characteristics can also occur when participants assume the role of a *faithful participant*, eschewing all pragmatic cues and following instructions exactly to the letter (Weber & Cook, 1972).

Survey materials in experimental philosophy studies are particularly likely to exhibit demand characteristics because experimental philosophers often present naive participants with bizarre thought experiments. Although the uniqueness of thought experiments is harmless in professional philosophy, there is evidence that folk participants are more likely to assume a faithful role, ignoring pragmatic and contextual cues, when experimental materials are unrealistic (Weber & Cook, 1972). That is, if experimental materials are too convoluted or unrealistic, then participants might ignore the contextual cues in experimental materials, or interpret them under different assumptions. Additionally, if participants are apprehensive about being evaluated, then they are more likely to try to guess at desirable or "correct" response. When the cases they have been presented with are particularly bizarre thought experiments, they may engage in a kind of amateur philosophizing, diverging from the aims of experimental philosophers. If demand

---

[2] For example, "Imagine 100 healthy people started smoking cigarettes, how many do you think would develop cancer?"

[3] Under this definition, there is some overlap with the concerns raised previously. For instance, pragmatic cues in base-rate experiments may have led participants to focus only on the descriptions they were given and to suppress information about base-rates. Since the experimenters were interested in how participants would use all the information they were given to produce the most accurate judgment they could, participants who interpret the instructions in this way clearly did not perform the intended task.

characteristics cannot be ruled out, then it is unclear how to interpret the results of surveys. And experimental philosophers cannot simply assume that their results will not be so affected.

III. Semantic Integration

In this section, we propose a new methodology for investigating concepts that we call Semantic Integration. Inspired by research in cognitive psychology, this method uses memory tasks as an implicit measure of how concepts are activated by different situations. As we discuss, semantic integration has important advantages over survey-based research, minimizing the influence of pragmatic cues and greatly reducing the possibility of demand characteristics. It also provides a more direct way of examining concepts. First, we introduce research on memory and language processing that inspired this methodology. Then, we describe the components of a semantic integration task, and two of our experiments that employ the method.

a. Memory and Language Processing Research

People tend to think of errors in memory as errors of omission-- they acknowledge that we sometimes forget things that have happened to us, but assume that we can only form memories for things that really have happened. But psychologists have found that people sometimes report memories for events that have never actually occurred (for a review, see Schachter, 1995), indicating that memory is not entirely dependent on external inputs. Bartlett (1932) is often credited with reporting the first experimental evidence for the formation of false memories. In his research, he had participants read a story and then recall it several times after subsequent delays. His analyses were informal, but he reported that memories grew increasingly distorted after each recall. Since Bartlett, researchers have found evidence for the formation of false memories in list learning paradigms (Deese, 1959; Roediger & McDermott, 1995), and in retention of sentences (Bransford & Franks, 1971), longer prose passages (Sulin & Dooling, 1974), image sequences (Loftus, Miller, & Burns, 1978) and videos (Loftus & Palmer, 1974). These phenomena are more than just curiosities; researchers have leveraged false memory to investigate the nature of our mental representations as well as our language comprehension processes.

Psychological research indicates that people's memories are influenced by semantic processing, and that people's memory is better for semantic information than for specific episodes or verbatim utterances (Anderson, Bjork, & Bjork, 1994; Anderson & Ortony, 1975; Deese, 1959; Loess, 1967; Roediger & McDermott, 1995; Sachs, 1967). Even in a simple experimental context (e.g., learning lists of words), experiences are processed and given semantic representations. In one study, Roediger and McDermott (1995; also see Deese, 1959) asked participants to memorize lists composed of different words that were semantically related to a single target word. When participants were asked later to recall the words they had been

presented with, they were just as likely to falsely recall the target word, which had never been presented, as any of the other words that actually appeared in the list. For example, when presented with a list made up of words like “glass”, “pane”, and “shade,” people are just as likely to recall the target word “window,” even if the word never appeared in the list. To introduce some terminology, the list *semantically activates* the word “window” -- it causes people to form or retrieve stored mental representations for the word “window.”

Cognitive psychologists and psycholinguists have leveraged false memory phenomena to examine language processing (e.g., Bransford & Franks, 1971; Flagg, 1976; Gentner, 1981). In particular, prior research investigated how semantic information is combined to form meaningful structured representations, or *discourse meanings*. This process, sometimes called *semantic integration* (Bransford & Franks, 1972), enables people to comprehend complex ideas communicated through connected discourse. Early research by Sachs (1967) found that memory for the meanings of sentences is more robust than memory for their specific wordings. He asked participants to read passages and then tested their recognition for sentences either immediately, or after they had read different amounts of intervening material. Some of the tested sentences had actually appeared in the text, but others were altered semantically or syntactically. When the meanings of the sentences were changed, participants made few errors. Even after substantial distraction participants rarely reported memory for sentences that had not appeared in the passage. However, when the changes were syntactic (e.g., a shift from active to passive voice), participants often reported recognizing the new sentences. After distraction, their recognition performance was near chance. Sachs concluded that during language processing, the original form of presented material is stored just temporarily, only long enough to be comprehended, whereas the material’s meaning is encoded into long-term memory.

If semantic information is integrated during language processing, and it is the meaning of a passage that is encoded into memory, then memory ought to exhibit *productivity*. That is, it should be possible for exposure to several basic, interrelated sentences to produce false memory for a sentence that expresses the integrated representation. A number of studies have confirmed this prediction, indicating that people integrate simple sentences to form representations for more complex sentences during language comprehension (Bransford & Franks, 1971; Cofer, 1973; Flagg, 1976). Additionally, people have been found to integrate information from text passages read during an experiment with their general background knowledge, leading to false recall for additional information that was not experimentally presented (Owen, Bower & Black, 1979; Sulin & Dooling, 1974; Thorndyke, 1976).

To explain these findings, Gentner (1981) examined a model of language processing in which sentences are considered both individually and in the broader context of the story in which they appear. Her model assumes that when a sentence is read within the context of a larger

passage, the discourse meaning that a reader forms may incorporate information not contained in the original sentence. Following research in linguistics (e.g., Chafe, 1970), artificial intelligence (e.g., Schank, 1972; 1973), and psychology (e.g., Miller & Johnson-Laird, 1976; Stillings, 1975), Gentner hypothesized that complex verb meanings can be represented by networks of *subpredictates* that express semantic relationships. Crudely put, a verb's subpredicates are simpler verbs that function as components of the more complex verb. To illustrate, consider the relationship between the verb 'give' and the more specific verb 'pay'. On Gentner's analysis, 'giving' some item is to take some action that transfers ownership of that item to a recipient. 'Paying' is a more specific form of giving, in which the giver owes the recipient. Thus, a representation of 'gave' would include subpredicates like 'caused', 'changed', and 'possession', and a representation of 'paid' would add the subpredicate 'owed'. Gentner tested this hypothesis by asking her participants to read paragraph-long stories that each included a sentence with the verb of interest -- the *critical sentence*. For instance, one story contained the critical sentence, "Max finally gave Sam the money". In the experimental condition, additional context explained that Max owed Sam money, whereas the control condition lacked this context. After reading one version of the story, participants performed a recall task in which they were shown the critical sentence with the word 'gave' removed, and they were asked to fill in the word that had appeared in the story. In support of Gentner's predictions, participants who had been provided with the additional context were more likely to falsely recall the verb 'paid' than participants in the control condition.

b. Using Semantic Integration to investigate philosophical concepts

Whereas Gentner (1981) used a false recall paradigm to examine how verbs with known meanings are integrated during language processing, we propose that the same methods can be used to investigate the meanings of philosophically significant concepts. On a traditional view, many philosophical concepts are complex mental entities constituted by simple concepts. The simple concepts, jointly, provide a "definition" of the complex concept. This means, at the very least, that the constituent concepts express properties which provide necessary and (jointly) sufficient conditions for the instantiation of the complex concept. On this traditional conception, the job of the philosopher is to discover the definition of putative complex concepts such as KNOWS, GOOD, PERSONHOOD, FREE WILL, MORAL RESPONSIBILITY, etc. Moreover, the philosopher can achieve this task without investigating the external world since the task mostly involves reflection and thought experiments.

In terms of semantic integration research, the traditional view makes the prediction that a word which expresses a concept C will have subpredicates which are the constituents of C.[4] For example, a view which says that the concept KNOWLEDGE is a complex concept constituted by the concepts JUSTIFIED TRUE BELIEF (Plato, Chisholm, Ayer) will make the prediction that the constituent concepts, expressing necessary conditions, will be subpredicates for 'Knowledge'. Researchers can test whether including these subpredicates in a passage leads to false recall for the concept of interest, offering evidence that these subpredicates were integrated to trigger the concept. If the integration process produces recall for the target concept, researchers can infer that these subpredicates are part of the concept.

Although Semantic Integration has a straight-forward application to complex concepts, it can also be used to investigate simple concepts. This is because simple concepts may have non-trivial necessary conditions. More generally, a concept (complex or simple) can have a necessary condition which does not also correspond to a constituent concept. For example, a necessary condition for being water is that it contain Hydrogen molecules. But the concept HYDROGEN is not a constituent of the concept of WATER. Now in this case, Semantic Integration can be used to test whether an agent recognizes that containing hydrogen is a necessary condition for being water. For example, we might discover that a participant who is told that some substance is a liquid and that it is made of Oxygen and Hydrogen to integrate and falsely recall that they saw the word 'water'. The key point here is that there is no requirement that the word under investigation, 'water' express a complex concept in order for agents to semantically integrate concepts that yield the false recall of 'water'.

Now in the example we just considered, the concepts that are integrated to yield the false recall of 'water', LIQUID, HYDROGEN and OXYGEN are not actually jointly sufficient for instantiating water. What is missing is that the hydrogen and oxygen atoms formed a molecule of a specific type with two hydrogen atoms for every oxygen atom. This suggests that metaphysical exactness is not needed for Semantic integration to work. In Gentner's experiments (1981) the role of subpredicates are played by concepts which are not jointly sufficient for the target concept. For example, she shows that people integrate 'working on' and 'workers are carrying whitewash and rollers' to yield the false recall of 'painting'. But strictly speaking, these features are not jointly sufficient for painting. The workers might have carried the whitewash but ended up working on something unrelated to painting. Of course, we expect participants in semantic integration tasks to understand the story in a plausible way. As a result, there is no requirement that the items integrated to yield a concept actually form a jointly sufficient set for that concept in some strong metaphysical sense.

---

[4] In the description of Gentner's research we said that subpredicates were linguistic items. It is better to think of them as mental representations or concepts since an agent may lack the words which play a role in the integration.

Nor is there a requirement that the items playing the role of subpredicates in integration correspond to necessary conditions for the concept at issue. Recall Gentner's example: Clearly, carrying whitewash and rollers is not a necessary condition for painting (think of spray painting). What these thoughts suggest then is that semantic integration can not only be used to investigate simple concepts, but can also be used to investigate concepts which have no discernible or interesting conditions which are necessary or sufficient. What seems to be required instead is something much weaker. What may be required for semantic integration is that agents judge that the items make it *likely* that the target concept is instantiated in the particular context.

This is good news for the prospects of using Semantic Integration experiments in the study of philosophical concepts. It is good for at least four reasons. First, in the case where a philosopher is interested in studying a concept that happens to be complex, she can get a subject to integrate concepts which merely approximate the true constituent concepts. For example, suppose that a concept C in fact has constituents X, Y and Z and we want to test whether Z is constituent. We can still test for false recall of the lexicalization of C in the presence of Y and Z but without invoking X so long as we have something else that approximates it.

Second, philosophers have argued that many philosophically interesting concepts are simple (Fodor, "Concepts", Williamson "Knowledge and its limits"). Yet, these concepts may still have interesting necessary conditions. For example, Williamson holds that although KNOWLEDGE is simple, the concept still has philosophically important necessary conditions. For example, many philosophers hold that a necessary condition for X knowing P is that P be true. In the cases when we are wondering whether the necessary condition is something that laypeople accept, Semantic Integration can help here as well, as the example involving WATER and HYDROGEN reveal.

Third, some philosophers think that some concepts are sensitive to certain parameters and that this sensitivity is accessible to laypeople. For example, Joshua Knobe (cite BBS article) holds that competent folk mental state attributions are sensitive to the moral valence of the content attributed, and some epistemologists have claimed that competent folk knowledge ascriptions are sensitive to practical interests (Stanley and Sripada, Pinillos) and also moral properties (Beebe?). Yet in none of these cases, is it required that the concept at issue be complex or even that moral valence or practical interests constitute interesting necessary conditions for the concept. Yet, Semantic Integration is still apt for testing parameters. To do this, we develop vignettes varying with the parameter at issue. We choose a critical sentence whose truth, together with a parameter given a certain value is thought to yield the concept at issue. We then test for a false recall of the lexicalization of that concept (Henne and Pinillos in preparation; Waskan et al., submitted).

Fourth, Semantic Integration can also be used to study the semantic contours of a word even if the word does not express a unique concept. One way this can happen is if a word is context sensitive so that it expresses different concepts depending on the conversational setting.

For example, some philosophers accept that ‘knows’ is context sensitive so that the word expresses different concepts KNOWS1, KNOWS2, KNOWS3 etc. corresponding to the different standards associated with ‘knowledge’ (Cohen, Lewis). For example, if a conversation takes place in a casual setting, the standards for knowledge might be lower than in a conversation taking place in a philosophy classroom. In the former, a use of ‘knows’ might express the concept KNOWS1 while a use of ‘knows’ in the latter context may express a more demanding concept KNOWS4. Semantic Integration can be used here to test context sensitivity, though the mechanism involved is very different from what we have seen so far.

In the previous cases discussed, the concepts C1, C2,....Cn which are integrated into a concept C (yielding a false recall of a lexicalization of C) are approximately jointly sufficient for the concept to be satisfied in the vignettes. In context sensitivity, there may be two levels of semantic integration corresponding to the setting of the context and also the traditional conceptual integration. First, the concepts integrated A1,A2,...An do not integrate to the target concept C. Rather, they integrate to a concept B which sets the conversational context. Second, once the standard is set, further concepts C1, C2,.....Cn are integrated to yield the concept C in the traditional sense. For example, BEING IN A BAR and BEING WITH CASUAL FRIENDS may integrate to LOW EPISTEMIC STANDARDS. But now if we suppose that the critical sentence is ‘A believes P’, then the concepts LOW JUSTIFICATION and P is TRUE might integrate with LOW EPISTEMIC STANDARD to yield KNOWS4 which corresponds to a low standard version of ‘knowledge’. If this process account is on the right track, then we may exploit it to yield a new test for context sensitivity.

What these four points reveal is that the viability of the Semantic Integration method does not depend on any particular understanding of concepts. On the contrary, it looks like the method is applicable under a wide variety of assumptions about concepts. The versatility of the method is then especially useful for philosophers who themselves might disagree about the nature of the concepts under study.

c. Two Experiments Using Semantic Integration

In the remainder of this section, we discuss two experiments we conducted that demonstrate that semantic integration can be used to investigate philosophically significant concepts. In this research we focus on the concept KNOWLEDGE, but recently other researchers have adopted our methods in order to examine EXPLANATION (Waskan et al., submitted) and CAUSATION (Henne & Pinillos, in prep).

There are three main components in a semantic integration study. The first component is the passage containing the contextual information hypothesized as sufficient to semantically activate the target concept. In order to construct passages that lead to false recall of

KNOWLEDGE, we altered contextual information in different versions of a main story, controlling for word count, sentence length, and overall structure. In a preliminary study, we constructed two versions of a story about a detective (Jack Dempsey) who forms the belief that a suspect (Will) is guilty. In the experimental condition, the detective's belief is justified by legitimate evidence and his belief is true (the suspect is in fact guilty). In the control condition, the detective cannot find any evidence and participants are not told whether the suspect is guilty, but the detective forms the belief anyway.

In each of these stories, we included a critical sentence containing a critical verb. Recall that when sufficient contextual information licenses using a more specific verb, people will falsely recall the more specific verb as appearing in the passage. In our knowledge experiment, we chose a critical verb that is consistent with knowing that P, but does not *entail* knowing that P; in this case, we used the verb 'thought'. The idea is that when read in the right context, a sentence containing the word 'thought' will lead to false recall of the word 'knew'. We predicted this predominantly will occur in the experimental condition where the appropriate context is supplied, and will happen significantly less in the control condition.

> Critical sentence: "Whatever the ultimate verdict would be, Dempsey thought Will was guilty."

An additional consideration when choosing a critical verb is the frequency with which that word occurs in English communications. Generally, it has been found that recall performance is better for high-frequency than for low-frequency words, and that the opposite is true for recognition performance (Kintsch, 1970). That said, there is some evidence that low-frequency words might benefit at recall when they presented together with high-frequency words (Duncan, 1974; Gregg, 1976), as will likely be the case in Semantic Integration experiments. In any case, a good practice would be to ensure that critical and target words are matched for frequency of occurrence as closely as possible. 'Thought' and 'knew' are reasonably well-matched as the 179th and 300th most common words in the english language, respectively (Wolfram|Alpha, 2013).

The second component of a semantic integration study is a distractor story. In principle, this distractor task could consist of almost anything. The purpose of the distractor is simply to diminish the effect of episodic memory in the recall task. Importantly, however, distractors do not contain either the critical verb or the target word.

After reading the distractor, participants advanced to the third part of the experiment, the recall task. They were shown several sentences from the story, each with one word removed. Their task was to recall the word that appeared in the blank. We were interested in their recall performance for the critical sentence. During the recall task, participants were shown this sentence with the word "thought" replaced with a blank, as shown below:

> Recall Task: "Whatever the ultimate verdict would be, Dempsey _____ Will was guilty."

Participant typed in the word that they recalled as having appeared in the original story. Consistent with our predictions, participants were more likely to recall "knew" as having appeared in the sentence when the detective's belief was justified and true (Powell et al., Draft).

Clearly, this finding does not demonstrate anything particularly interesting about "knowledge", but it does demonstrate that semantic integration can be used to examine philosophical concepts. Consequently, we investigated Gettier cases, a clearly more substantive issue in philosophy. We adapted our detective story and added another character named Beth. Beth is Will's soon-to-be ex-girlfriend, who has it in for Will and interferes with Dempsey's investigation. We created three versions of the story, a false belief version, a Gettier version, and a justified true belief version.

> In the false belief condition, Will is innocent, but Beth framed him by committing the crime and planting evidence.
>
> In the Gettiered condition, Will committed the crime and disposed of all the evidence, but Beth makes sure Will gets caught by planting evidence for Dempsey to find.
>
> In the justified true belief condition, Will committed the crime and left behind evidence. Seeing his mistake, Beth does nothing and waits for Dempsey to arrest Will.

We found participants were more likely to falsely recall 'knew' as having appeared in the critical sentence in the justified true belief and Gettiered conditions than in the false belief condition. However, we also found *no* difference in recall between the Gettier and justified true belief conditions. Surprisingly, our participants did not distinguish between Gettiered and non-Gettiered justified true belief (Powell et al., draft).[5]

IV. Pragmatic Considerations and Demand Characteristics

Semantic integration tasks offer two important advantages over more explicit survey methods because they implicitly measure the semantic activation of a concept. For one, semantic integration tasks avoid the concerns raised by Cullen (2010) over pragmatic cues. Researchers using survey methods need to account for pragmatic cues in the stimuli that they present to participants as well as in their instructions, questions, and response options. In the semantic

---

[5] The details of the study and arguments made on the basis of this data can be found here.

integration paradigm, participants are told they are performing a memory task and nothing in the instructions, response prompts or options indicates otherwise. While these materials are not devoid of pragmatic cues, pragmatic factors in this context are considerably less problematic, and considerably better understood. Psychologists have studied memory since Ebbinghaus (1885), and have developed reliable methods for testing people's recollection of presented material. While it is clear that stimuli can still contain pragmatic cues and conversational implicatures, this fact is not in any way unique to semantic integration. For one, survey methods will also face these same concerns. Moreover, if one were skeptical about an experimental paradigm for this reason, one would also have to be skeptical about research on causal reasoning, decision making, psycholinguistics, or nearly any line of research that involves presenting text to participants. The pressing concern is that pragmatic cues in instructions will lead participants to approach the experimental task incorrectly, or to interpret their response options in a manner inconsistent with the researcher's intentions. Semantic integration tasks avoid these difficulties.

Second, semantic integration tasks preclude demand characteristics. Even if participants are apprehensive about being evaluated, their apprehension is unlikely to lead researchers to any erroneous conclusions. Evaluation-apprehension should motivate participants to perform the task well, but since there is no reliable way for participants to produce "desirable" answers except by probing their own memory, there is little risk of evaluation-apprehension leading to spurious findings. In addition, because the memory task is both intelligible and experimentally realistic, participants are less likely to take on the role of the faithful participant (Weber & Cook, 1972). Even if some participants do ignore experimenters' conversational implicatures, this is unlikely to affect their performance, as the instructions of a memory task can be made comprehensible even without many contextual cues.

V. Caveats

The interpretation of findings from semantic integration tasks can depend on resolving three questions:

(a) How are concepts structured?
(b) What mental process leads to integration of semantic information?
(c) Does "impure" semantic integration complicate matters?

a. The structure of concepts

If semantic integration directly measures the semantic activation of people's concepts, then one might wonder about the nature and structure of these concepts. As discussed, Gentner (1981) hypothesized that verb concepts are represented as structured collections of subpredicates. On the basis of this view, she made and confirmed very specific predictions about how

representations would be combined during the processing of connected discourse, lending support for her theory. Still, psychologists have attempted to describe concepts using a number of representational formats (e.g., Posner & Keele, 1968; Medin & Schaffer, 1978) and some philosophers have questioned whether some concepts can be analyzed at all (cite williamson). This may prompt some to doubt that Gentner's model of concepts is accurate, or to worry that, even if it accurately describes the representations of certain concepts, different types of concepts may be represented in other ways (e.g., natural kind terms, prototype or exemplar models, distributed representations). Although these possibilities may complicate the interpretations of semantic integration experiments, researchers who use semantic integration can remain agnostic to the "true" theory of concepts. The method rests on two basic assumptions: (1) semantic concepts are mentally represented in some fashion and (2) memory for the *meaning* of a passage is more robust than memory for its exact wording. The first claim is a fundamental assumption of modern psychology and one which we will not defend. The second is supported by a large body of research on memory, some of which we discussed in section III (e.g., Bransford & Franks, 1971; Brewer, 1977; Barclay, 1973; Cofer, 1973; Flagg, 1976; Sachs, 1967).

b. Mental Processes and Semantic Integration

Thus far we have assumed that integration occurs during comprehension and encoding, but another possibility is that integration actually occurs at recall. That is, during encoding people store the meanings of individual propositions separately. Then, at recall, they integrate these meanings by a process of inference to form a reconstruction of the memory for an individual sentence or proposition. Supposing this is true, it is worth noting that semantic integration still overcomes concerns about demand characteristics and pragmatic cues, but it can no longer be said to provide a direct a measure of semantic activation. Rather, in this case the responses that participants give are just as dependent on inferential processes as their responses to surveys. Fortunately, Gentner (1981) tested this possibility by inserting contextual information before or after the critical sentence in a passage. She found that false recall for critical items was greater when the inserted material came before the critical sentence, supporting the interpretation that meanings are integrated on-line during discourse comprehension rather than after the fact during recall. This supports our assertion that semantic integration isolates conceptual activation from downstream inferential processing.

c. *Impure Semantic Integration*

"Pure cases of semantic integration" (Gentner, 1981, p. 371) occur when the subpredicate structures of *n* propositions are directly combined to produce some unified structure. For example, Gentner describes 'gave' and 'owed' resulting in recall for the verb 'paid'. However,

Gentner also provides evidence that semantic integration can occur when the context does not directly specify any of the subpredicates in the new semantic structure. As discussed earlier, she found that people recall 'painting' in place of 'working' when they are told that the workers are carrying brushes, whitewash, and rollers. People need to infer that the workers are using these materials, and thus, that they are painting. Ideally, researchers can make inferences about the subpredicate structure of a concept based on people's integration performance. However, researchers must exercise caution in light of Genter's results. It would clearly be an error to infer on the basis of her findings that "carrying brushes, whitewash, and rollers" is really a component of the subpredicate structure of 'painting'. Consider the fact, for instance, that spray painting does not require brushes, whitewash, or rollers.

VI. Alternate Experimental Designs and Surveys

a. Similar experimental paradigms

In this paper we described an experimental method modeled after Gentner's (1981) work on the semantic integration of verb meanings, and detailed its use for examining people's concept of KNOWLEDGE. It bears noting that there are a number of other related experimental paradigms that have been used to examine semantic integration in discourse comprehension (e.g., Bransford & Franks, 1971; Brewer, 1977; Barclay, 1973; Cofer, 1973; Flagg, 1976; Sulin & Dooling, 1974; Thorndyke, 1976; Owens, Bower, & Black, 1979), and that some of these paradigms might also be put to use in experimental philosophy. However, Gentner's (1981) paradigm has several qualities that are desirable for the purposes of experimental philosophy, even relative to other semantic integration tasks. First, the use of a free recall task makes its results more compelling than tasks that rely on recognition judgments. Participants' responses to recognition tasks can be influenced by both true recollection as well as feelings of familiarity (Tulving, 1985). In contrast, explicit recall of the word 'knew' provides unambiguous evidence for the semantic activation of the concept KNOWLEDGE. Second, this paradigm focuses responses onto a single specific word of interest, whereas other semantic integration paradigms often ask participants to evaluate larger semantic units, such as phrases or sentences (e.g., Bransford & Franks, 1971; Sulin & Dooling, 1974). Specifying a target verb can reduce ambiguity in investigations of individual concepts. Thus, we argue that, where possible, the semantic integration tasks we describe here are the optimal method for examining the parameters involved with instantiating people's concepts.

Of course, not all concepts of interest will necessarily have a verb form ('knew'), with nearby synonyms ('thought', 'believe'). When this is not the case, other semantic integration tasks may be more appropriate. The disadvantages associated with semantic integration tasks measuring recognition for sentences or phrases (e.g., Bransford & Franks, 1971; Owens, Bower & Black, 1979) are not insurmountable. In particular, employing a remember-know procedure

(Tulving, 1985) can help distinguish between genuine recollection and familiarity. With sufficient care, it is possible to craft phrases or sentences that unambiguously express whatever concept may be of interest to researchers (e.g., Waskan et al., submitted).

b. Surveys and Semantic Integration

The methodological advantages of semantic integration are due to the implicit nature of the task. However, this also marks semantic integration tasks as importantly different from the explicit measures of survey tasks. Different research questions might warrant the use of either surveys or semantic integration depending on the aims of the researcher.

Many experimental philosophers hope that through a better understanding of the psychology of philosophical concepts, we will be better able to assess philosophical arguments that deploy those concepts. We have argued that, in general, semantic integration tasks are well suited to accomplishing this goal. Semantic integration tasks provide an implicit measure of conceptual activation, making them ideal for capturing these sorts of intuitive reactions. However, some philosophical concepts could also be applied to situations by more effortful cognitive processes. In these cases, explicit survey questions that elicit conscious consideration may be better suited, if these questions can be adequately constructed. Additionally, surveys may be more appropriate where experimental philosophers are interested in people's judgments. For instance, some researchers may not be interested in KNOWLEDGE per se, but in knowledge ascription behavior. In this instance, semantic integration would be inappropriate and surveys would be preferable.

VII. Conclusion

In this paper, we discussed the ways in which pragmatic cues and demand characteristics can affect the results of surveys. In light of these problems, we argued that experimental philosophers should adopt a new experimental paradigm that we call Semantic Integration. Our experimental investigations of KNOWLEDGE demonstrate how this method can be used to examine philosophical concepts. Semantic Integration can be applied to investigate complex concepts in a manner consistent with the aims of traditional conceptual analysis, and used to examine other parameters relevant to the instantiation of concepts. This method avoids concerns about pragmatic cues and demand characteristics because participants' conceptual activation is measured implicitly through a memory task. For these reasons, Semantic Integration represents an important methodological advance in experimental philosophy.

REFERENCES TO BE COMPLETED WHEN FORMATTING INFO IS GIVEN

Plato seems to be considering some such definition at Theaetetus 201, and perhaps accepting one at Meno 98.

Roderick M. Chisholm, Perceiving: A Philosophical Study (Ithaca, New York: Cornell University Press, 1957), p. 16.

A. J. Ayer, The Problem of Knowledge (London: Macmillan, 1956), p. 34.

word frequency cites

GREGG, V (1976). Word frequency, recognition, and recall. In J. Brown (Ed.),*Recall and recognition* (pp. 183–216). London: Wiley

Wolfram|Alpha (accessed March 3, 2013)
http://www.wolframalpha.com/input/?i="knew"&a=*C.knew-_*Word-
http://www.wolframalpha.com/input/?i=thought&a=*C.thought-_*Word-