Running head: Concept Appraisal

Concept Appraisal

Sapphira R. Thorne[a], Jake Quilty-Dunn[bc], Joulia Smortchkova[bc], Nicholas Shea[bc*], & James A. Hampton[a]

[a] Department of Psychology, City, University of London, London, EC1V 0HB, UK

[b] Institute of Philosophy, School of Advanced Study, University of London, London, WC1E 7HU, UK

[c] Faculty of Philosophy, University of Oxford, Oxford, OX2 6GG, UK

Author Note

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No. 681422 (MetCogCon).

* Corresponding Author: Nicholas Shea, Institute of Philosophy, School of Advanced Study, University of London, London, WC1E 7HU, UK. Telephone: +44 (0) 207 862 8824. Email: nicholas.shea@sas.ac.uk

Word count (excluding references, tables, and abstract): 13,061

Abstract

This paper reports the first empirical investigation of the hypothesis that epistemic appraisals form part of the structure of concepts. To date, studies of concepts have focused on the way concepts encode properties of objects, and the way those features are used in categorisation and in other cognitive tasks. Philosophical considerations show the importance of also considering how a thinker assesses the epistemic value of beliefs and other cognitive resources, and in particular, concepts.

We demonstrate that there are multiple, reliably judged, dimensions of epistemic appraisal of concepts. Four of these dimensions are accounted for by a common underlying factor of how well people believe they understand a concept. Further studies show how dimensions of concept appraisal relate to other aspects of concepts. First, they relate directly to the hierarchical organization of concepts, reflecting the increase in specificity from superordinate to basic and subordinate levels. Second, they predict inductive choices in category-based induction.

Our results suggest that epistemic appraisals of concepts form a psychologically important yet previously overlooked aspect of the structure of concepts. These findings will be important in understanding why individuals sometimes abandon and replace certain concepts; why social groups do so, for example during a 'scientific revolution'; and how we can facilitate such changes when we engage in deliberate 'conceptual engineering' for epistemic, social and political purposes.

Keywords: concepts; categorization; conceptual engineering; epistemic appraisal; metacognition; essentialism; dual-character

1. Introduction

Most words in a person's lexicon are associated in semantic memory with a meaning – with a lexical concept. Concepts organize information in ways that guide communication, perception and action. They are the fundamental basis of shared knowledge and culture. Much is known about how concepts encode the features or properties that characterise a category. The study of concepts should not, however, be limited to how categories are demarcated. Our aim is to investigate the structure of concepts from an entirely different perspective: is a given concept epistemically useful? How much does a thinker know about the category? How well does the category allow generalization? We want to find out whether people's conceptual knowledge includes appraisals of this kind – appraisals of the epistemic value of a concept.

Finding out how concepts are appraised is important because concept appraisals are likely to play a key role in the way we change our concepts, both individually and collectively. The history of science encompasses a series of 'scientific revolutions' (Kuhn, 1970). During a scientific revolution, as empirical findings accumulate that can no longer be accommodated by existing theories, new concepts are constructed to replace the existing conceptual scheme. For example, the concepts of heat and temperature were introduced to replace a concept that was less useful, because it ran together these two different thermodynamic properties (Carey, 2009). While the existence of scientific revolutions has long been recognised, we still do not know what kinds of epistemic appraisals drive scientists to abandon an existing concept and replace it with something new.

Away from science, there has been growing interest in deliberately reconfiguring social concepts in areas like race and gender (Haslanger, 2000). This process of 'conceptual engineering' is aimed at changing our concepts directly, so as to achieve improved public understanding and behaviour (Cappelen, 2018; Machery, 2017; Thomasson, 2017). If the

conceptual engineering enterprise is to succeed, a psychological understanding of concept appraisal will be critical. People will resist conceptual change when they feel that a concept is well-understood, and when they have the sense that a given category supports generalisations and forms a good basis for building knowledge. By contrast, undermining people's epistemic confidence in a concept, implicitly or explicitly, opens the way to its modification or replacement.

Since epistemic appraisal of concepts has not been investigated before, we turned to the philosophical literature for suggestions about the kinds of appraisal to look for. The predominant focus in philosophy (in epistemology) has been the epistemic appraisal of beliefs and of the cognitive mechanisms for forming beliefs (Carr, 2017; Hookway, 1994; Skorupski, 2010). We reflect on what we know, how much knowledge we have and whether the things we believe are accurate (Goldman & Beddor, 2015; Proust, 2013; Treanor, 2013). Similar assessments can be applied to the collection of information that is encoded in a concept. That is, we can reflect on our own individual understanding of a concept. A concept is a cognitive tool, and in the first case the thinker may be asking, 'how good am *I* with this tool, how well do *I* understand it?'. On the other hand, a thinker may focus on the tool itself and ask, 'how useful is this tool?', 'is it good for various purposes?' (Goldman, 1978), where the focus is not on the thinker's own understanding, but on the category itself. Some categories tell us a lot about an object (Boyd, 1991; Keil, 1989; Quine, 1970). Category members share many features and display them reliably (Égré & Ó Madagáin, 2019). The category is rich and homogeneous enough that someone who has learnt about it – whether the thinker herself, or some expert in the relevant domain (Kalish, 2015; Putnam, 1973) – can usefully rely on it (Osherson et al., 1990; Rips, 1975). When an object is correctly categorised under the concept, we can form a large number of reliable expectations about that object (Millikan, 2000). Both of these broad kinds or families of epistemic appraisal – of the

thinker's own understanding and of how useful the category is epistemically – potentially come in several varieties which we exploit in our studies. We draw these out in the next section (Section 1.1).

To develop a method to explore these questions we took inspiration from a study by Haslam, Rothschild, and Ernst (2000). They were interested in the different ways a social concept could be based on an "essence". To this end, they developed a set of nine scales which they considered to reflect different aspects of essentialism. Essentialised concepts were hypothesised, among other things, to be natural, to support induction, and to be immutable. Participants rated a set of 40 social categories on the nine scales.

Our method was to devise, in a similar fashion, a set of scales or 'dimensions' (as we will call them) to cover ways people may plausibly appraise their concepts epistemically. We chose eight dimensions, as explained below. A first, basic question, is whether or not these dimensions of appraisal are readily evaluated for common concepts. That people reliably agree on how to judge a set of concepts along a given dimension would lend credence to the idea that epistemic appraisals are properties of these concepts. If the dimensions prove reliable, then we go on to ask, still following Haslam et al. (2000), whether dimensions of appraisal fall together into correlated clusters, or whether, evaluated across a broad selection of concepts, people assess them independently.

A second broad question concerns the psychological importance of these forms of concept appraisal. Are they an isolated island, nothing more than a stable pattern in the way people answer various questions about concepts? Or are they integrated with other aspects of the structure of concepts? We began to probe that question by examining whether the way a concept is appraised predicts other aspects of conceptual knowledge. For this initial investigation we chose two well-studied phenomena: taxonomic structure (superordinate, subordinate and basic levels); and category-based induction.

1.1 Putative dimensions of appraisal

In this section we explain how, guided by existing literature, we arrived at a set of dimensions to form the basis of this initial exploratory study. Guided by the two broad kinds of epistemic orientation mentioned above, we chose eight putative forms of appraisal, which we call 'dimensions' (and mark with Italics). A full description of these dimensions is given in Table 1.

The first group of dimensions concern the thinker's estimation of their own grasp or understanding of a concept. This is a matter of 'epistemic self-audit' (Skorupski, 2010). Epistemologists have focused mainly on beliefs and knowledge, but some have also indicated that the same principles should apply more widely: 'concerning any element of our cognitive apparatus, we can raise questions about its reliability' (Hookway, 1994, p. 214). Just as the reliability of beliefs consists in their truth or accuracy (Goldman & Beddor, 2015), so too with concepts – we can assess how accurately they represent the world. A concept encodes a collection of beliefs and expectations. Their accuracy is epistemically valuable (Carr, 2017). Similarly, psychological studies of metacognition (mechanisms by which people monitor and control their psychological states and processes) show that people are able to keep track of accuracy in many domains, albeit only moderately well (Fleming et al., 2010; Koriat, 2015; Proust, 2013). So our first dimension of appraisal, *Accuracy*, asks whether the information encoded by a given concept is accurate.

Avoiding error cannot be the only aim. We could do that by having no beliefs. Another goal of cognition is to acquire truths (Goldman, 1978, 521), to extend our knowledge (Hookway, 1994, p. 211). Even if it is hard to quantify knowledge, people do make common-sense judgements about how much they know (Treanor, 2013). In studies of metacognition, people answer some questions by assessing how much information they are able to access

(Koriat, 2012). Similarly, they can ask themselves how much they know about a category – that is, how much information is encoded by their corresponding concept. So our second putative dimension of appraisal is *How Much Do You Know*.

Much of what we know is implicit rather than explicit (Davies, 2015). We cannot put it into words. So too with concepts: they encode some information implicitly. For example, in a categorisation task, an object may be categorized based on its features, without the thinker being able to identify which features are being used (Ashby & Valentin, 2017; Sloutsky, 2010). At the same time they may still have some explicit metacognitive awareness of how well they are likely to categorise new exemplars (Jacoby et al., 2010). A similar distinction is made between two kinds of metacognition. 'Procedural' metacognition is often based on implicit information (e.g. Wahlheim, Finn, & Jacoby, 2012), for example on fluency of processing (Proust, 2012; 2013); 'analytic' metacognition is often based on stored explicit beliefs ('theory-based metacognition': Koriat, 2007). Applied to concepts, we can ask how much of what the thinker knows about a category is represented explicitly, and how much is implicit. We operationalise that question by asking participants to tell us how well they could explain the category to someone else. So our dimension *Explain* is an appraisal of how much of the information stored in a concept is explicit.

Next we turn to our second group of dimensions. These are appraisals of how good a concept is as a cognitive tool – of how epistemically useful the category is. Epistemologist Alvin Goldman has argued that we can 'assign appropriate values to different kinds of mental representation' based on the goals they are used to obtain (Goldman, 1978, 522). Concepts are used for the goal of forming reliable expectations about the properties of objects based on what we have learned in the past (Millikan, 2000). A thinker sees an object, categorises it as a hazelnut, breaks it open and experiences its flavour, and stores that experience using their

HAZELNUT concept.[1] Then, when next encountering an object that falls under HAZELNUT, they can form an expectation about how it will taste. This ability to perform category-based induction (Rips, 1975; Osherson et al., 1990) requires that a property observed in one member of the category is likely to be shared by others, that its members are homogeneous in this respect (Égré and O Madagáin, 2019). Our *Induction* scale asks participants to rate different concepts along this dimension.

Usefulness for induction is not just a matter of how many members share a given feature, but also of how many features are shared. If many features are shared across all or most members of the category, categorising an object under the concept tells you a lot about that object. The most informative categories are natural kinds (Quine, 1970; Boyd, 1991; Millikan, 2000). We know that people treat natural kinds as having essences which determine their properties (Keil, 1989; Murphy and Medin, 1985). This essentialised way of thinking extends to social concepts (Haslam, Rothschild, & Ernst, 2000), including those formed around a normative ideal (e.g. the "true" scientist has a pure thirst for knowledge: Knobe, Prasada, and Newman, 2013). Concept-users appreciate that some categories are more kind-like and informative than others. Our *Informativeness* dimension asks participants to rate that directly.

Where there is a natural kind, there are often experts in that kind. The most epistemically useful concepts are often supported by a high level of expertise. Ordinary concept users may defer to experts about what members of the kind are like (Kalish, 2015) and rely on experts' ability to classify real cases (e.g. to tell elms from beeches, Putnam, 1973). Conversely, people's disposition to defer to experts may indicate that they take a concept to be a particularly epistemically useful one. Our *Deference* dimension asks participants to tell us directly about deference to experts.

[1] We follow the philosophical convention of using small caps to name concepts.

In the ‘guru effect’, thinkers take a piece of information communicated by a person possessing certain status (e.g. scientific, religious, intellectual) to be based on deep ideas that they do not grasp themselves (Sperber, 2010). More generally, when deferring to experts a person may feel there is a lot to learn about a category even though they don’t know much about it themselves. Our next dimension asks people how much they think there is to learn about a category (*How Much To Learn*), starting from scratch. This is somewhat akin to *Informativeness*, but is more objective, since *Informativeness* depends upon your own knowledge. You may think there is much to learn about amphibians, while not knowing much yourself, so that learning that a creature is an amphibian does not tell you much about what it is like. Égré and O Madagáin (2019) point out that some categories apply to more members than others (‘inclusiveness’). Having more members, as well as having more features, might increase *How Much To Learn*. That would make superordinate categories rate highly on this dimension (which we explore in the study of taxonomic structure, Study 3). On the other hand superordinate categories (like *amphibian*) are likely to have a diversity of members, reducing the extent to which falling under the concept tells you much about what an individual is like (i.e. potentially reducing *Informativeness*).

Finally, we come to a dimension, *Communicate*, which could exemplify either of both of our two broad kinds of epistemic appraisal. Concepts are an interpersonal tool. They are the basis on which we communicate with others and build a shared understanding. Previous work has focused on the epistemic value of accuracy in communication. Even young children keep track of the reliability of the people they communicate with (Koenig and Harris, 2005), and adults continually exercise ‘epistemic vigilance’ about the accuracy of the things they are told (Sperber et al., 2010). Enlarging the focus slightly, it is plausible that people have a sense of whether a concept builds on a shared understanding, whether they are ‘on the same page’ as those they are talking to. Most everyday concepts are a good basis for communication (e.g.

JOURNALIST, FROG), since people understand them in the same way (Murphy, 2002, 244). Other concepts are understood in different ways by different people (PEST, WEED), including because some people have specialist knowledge that others do not share (SCHIZOPHRENIA, ARTHROPOD). Our *Communicate* question asks participants to assess concepts along this dimension. When expressing the concept, are people likely to understand each other easily, or is it hard to predict what other people will have in mind?

The *Communicate* dimension falls within both of our broad families of concept appraisal. On the one hand, it is an appraisal of a cognitive tool – assessing a concept as a potentially shared cognitive resource. That would relate it to *Informativeness* and *Induction*, and potentially also to *How Much To Learn* and *Deference*. On the other hand, people may be alive to their own state of knowledge or ignorance about a category (what exactly *is* an arthropod?), and thus appraise a concept's suitability for communication based on whether they themselves understand the category well. That would make *Communicate* analogous to rating how well one could oneself categorise new exemplars under a concept (Jacoby, Wahlheim, & Coane, 2010), and would put *Communicate* close to our other understanding-related dimensions: *How Much Do You Know*, *Accuracy* and *Explain*. So the way *Communicate* relates in people's minds to other dimensions of appraisal, if at all, is an open question for our studies.

Our eight dimensions potentially overlap with one another in various other ways as well. One aim of studies 1 and 2 is to see whether they overlap in people's minds: whether people assess them independently (if indeed they are judged reliably at all), or whether some dimensions are just notational variants of one another, different ways of asking about a common underlying factor by which concepts are appraised.

1.2 The current experiments

The studies were designed to investigate the psychological reality of forms of concept appraisal (studies 1 and 2), and to probe their psychological importance (studies 3 to 5). As an initial study in a new field, the design was necessarily exploratory. At the outset we did not know whether there is any form of epistemic appraisal on which people would agree when evaluating common concepts. Studies 1 and 2 investigate that. By selecting eight dimensions widely across different potential forms of epistemic assessment, we hoped to discover reliable forms of appraisal if they exist.

The second question for studies 1 and 2 was whether there is duplication amongst our dimensions: whether some dimensions are just different ways of asking about a common underlying epistemic property. We explored that by asking for ratings for a wide range of common concepts drawn from several different domains (social categories, health conditions, recreations, plants and animals, artefacts and foodstuffs). Study 1 used concepts selected to maximise variability in our eight putative dimensions of appraisal, based on the experimenters' intuitions. Study 2 performed the same procedure on a list of concepts drawn from existing concept norms. All participants in a study rated the same concepts and each rated them on just one of the eight dimensions. We looked at whether average ratings on any of the eight dimensions would correlate between subjects across our broad sample of concepts. If several different dimensions are appraised in the same way by different participants, that is evidence for a common underlying factor. That in turn would suggest that it is this common factor, rather than a larger number of individual dimensions of appraisal, that is part of the underlying structure of concepts.

Since Studies 1 and 2 provided preliminary evidence that different people do indeed reliably judge the same concepts in the same way along various dimensions, we went on to explore whether these appraisals are psychologically important – are they embedded in other

well-established aspects of the structure of concepts? Study 3 looked at whether dimensions of appraisal predict taxonomic structure – the distinction between superordinate, subordinate and basic level concepts. Studies 4 and 5 examined the same question for category-based induction, which we take to be one of the most important cognitive functions of a concept. We investigated whether the concept appraisals established in Studies 1 and 2 would predict the choices made by a different group of participants when given an induction task, the task of extending features observed in some members of a category to a new instance.

2. Five Studies

2.1 Study 1 – Reliability and Structure of Dimensions of Concept Appraisal

Study 1 is the first step in establishing whether any forms of epistemic appraisal attach to concepts. We investigated (a) whether people would agree in their appraisals of common lexical concepts along our eight dimensions, and (b) whether people track many different dimensions, or whether the different scales collapse into a smaller number of factors. Ratings were obtained for 160 concepts, with each participant rating all concepts along just one of the eight dimensions. With the aim of capturing whatever variability exists across the dimensions, the experimenters used their own intuitions to generate a list of common concepts offering contrasting assessments along one or more dimensions, for example in: *How Much You Know* (DOG OWNER VS. LEGAL CLERK), *Explain* (STUDENT VS. CONNOISSEUR), *Induction* (GENIUS VS. RENTER), or *How Much To Learn* (POLICE OFFICER VS. TEA DRINKER).

2.1.1 Method

2.1.1.1 Participants. There were 426 participants (295 Female, 115 Male, 16 undeclared) recruited through Prolific Academic who participated in exchange for a small monetary reward (Age 18-85; M = 35.6 years). Of these, 369 self-reported as native English speakers, 43 as fluent non-native speakers, and 14 were undeclared. We excluded 27 participants (6%)

with extensive missing data on all four domains, leaving a final N of 399. Participants were randomly allocated to one of eight groups, each judging a single dimension. N within groups varied from 46 to 60.[2]

2 The sample size yielded an average Standard Error for item means of 0.2 on the 1-5 scale. In simulations, this translated into a 95%CI of around 0.04 for a typical correlation of 0.6 between dimensions.

Table 1. The eight dimensions used in Studies 1 and 2. Instructions are shown for the Biology domain.

DIMENSION	INSTRUCTIONS	SCALE ANCHORS
Accuracy	For some categories of plant or animal, you may be pretty sure that most of the things you believe about it are true. For other categories you might doubt some of the beliefs you have about that category. Other categories may lie in between.	*sure they are true* to *not sure they are true*
How Much Do You Know	Some kinds of plants and animals will be well known to you, and others may be less so.	*know everything* to *know nothing*
Explain	For some categories, you may feel confident that you could explain what it is to someone who knew nothing about the category. For other categories you may feel that you could not explain it very well to someone else. Other categories may lie in between.	*explain* to *not explain*
Induction	When you learn something new about some plants or animals which belong in a category, others in the category are likely to have the same new property. For some categories, it may seem very likely that category members will share a property. For other categories you would be very likely to dismiss this as a coincidence. Other categories may fall in between these extremes.	*new property will apply* to *new property will not apply*
Informativeness	Some ways of describing plants and animals allow you to infer many other things about it; knowing that some plant or animal fits the description tells us a lot about that plant or animal. Other ways of describing plants or animals allow hardly any inferences about what else they are like; knowing that some plant or animal fits the description does not tell you much about it at all. Other categories lie between these two extremes.	*very informative* to *very uninformative*
Deference	For some categories there are experts who could tell you everything you need to know about plants and animals in that category. For other categories there are no such experts and everyone is entitled to their opinion. Other categories may lie in between.	*defer to expert* to *do not defer to expert*

(CONTINUED)

DIMENSION	INSTRUCTIONS	SCALE ANCHORS
How Much To Learn	There is more to learn about some categories than others. Starting from scratch, relatively how long do you think it would take to become an expert about the following categories?	*long time* to *short time*
Communicate	Some categories are a good basis for communication; people are likely to understand each other easily when they talk about kinds of plant or animal using category names. Other categories are less good for communication; it is hard to predict what other people have in mind when they talk about plants or animals using category names.	*likely to understand* to *unlikely to understand*

2.1.1.2 Materials Design and Procedure. Each participant rated 160 concepts on just one of the eight dimensions. The concepts were 40 items from each of four domains, namely: People (social categories); Recreations; Health; and (folk) Biology (plants and animals). The eight dimensions and their corresponding low and high scale labels are shown in Table 1. The definitions of the dimensions for Biology are shown. These were adapted for the other domains. Included in the instructions were one putatively good and one poor example of the dimension (not included in the subsequent question list). For example, in addition to the text in Table 1, the instructions for *Induction* for the Biology domain continued:

> If three sharks were found to have enzyme X, it may seem very likely that another shark will also have enzyme X. For other categories you would be very likely to dismiss this as a coincidence. For example, if three weeds were found to have enzyme X, it may seem very unlikely that another weed will also have enzyme X. Other categories may fall in between these extremes.

Participants were told that if they had no idea what a word meant, they should leave that question blank. The participant was given as much time as they needed to understand the

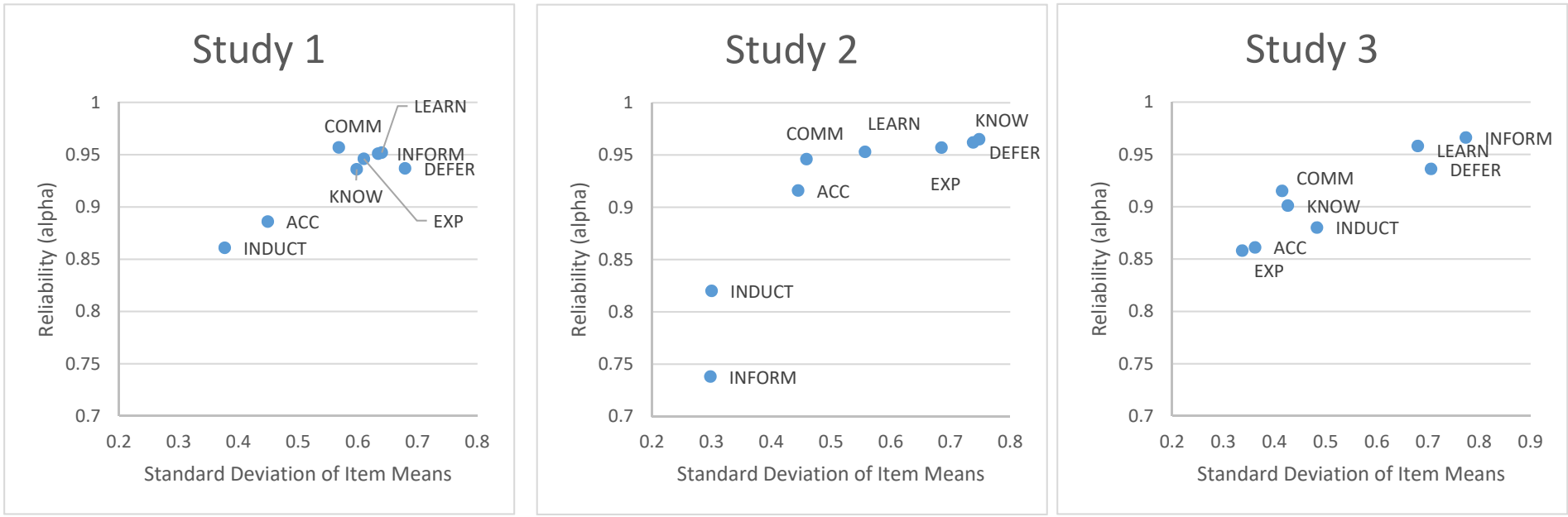
instruction for their assigned dimension. They were then asked to rate 160 concepts on that dimension. All scales ran from 1 to 5, anchored at the extremes with the scale anchors shown in Table 1. The four category domains were presented in a different random order for each participant and the 40 items for each category domain were presented randomly ordered on two pages. All words, with their mean ratings, are listed in the Supplementary Materials (Appendix A). Participants completed the questionnaire online, programmed in Qualtrics.

2.1.2 Results

2.1.2.1 Inter-subjective Agreement. Our first question was whether, for a given dimension, different people would appraise a wide range of concepts in the same way. To measure the degree of consensus we calculated Cronbach's α, a measure of reliability with a maximum of 1. Figure 1 shows the reliabilities for the dimensions. Alpha was greater than .86 for all eight dimensions, indicating very high levels of consensus and good reliability, given the relatively small sample sizes. Means for each item within each domain were calculated for each of the eight dimensions, using all non-missing data (2.4% of responses were missing[3]). Figure 1 shows how reliability varied as a function of the spread of means (SD) within the list. Where variance in means of the sample of items is low, one can expect measured reliability to be an underestimate. This effect can be seen for *Accuracy* and *Induction* in Study 1.

[3] Since participants were instructed to leave an item blank if they had no idea what the word meant, we did not force a response to each item in Qualtrics.

Figure 1. Inter-subject agreement in Studies 1-3. The graphs show reliability of dimensions as a function of the amount of variance (SD) in the means for items. Lower variance reduces measured reliability.



Key:

ACC: Accuracy

KNOW: How Much Do You Know

EXP: Explain

INDUCT: Induction

INFORM: Informativeness

DEFER: Deference

LEARN: How Much To Learn

COMM: Communicate

2.1.2.2 Factor Structure Analysis. Having established that all of our eight dimensions show strong between-subject agreement across concepts, our second question related to whether people keep track of eight separate dimensions of appraisal, or whether the dimensions collapse into a smaller number of factors. This analysis was based on the mean scale values for each of the 160 concepts on each of the eight dimensions. First, correlations between the dimensions across the 160 category items were calculated (see Table 2). High positive correlations were seen between the first four dimensions, and between the last four dimensions in the Table, indicating the probable existence of two underlying factors at this level of analysis.[4] (The order of dimensions has been changed from that in Table 1 to show the clustering into the two components. All further Tables now adopt this order.)

A principal component analysis (PCA) was performed with the aim of explaining variance of the eight dimensions across concepts in terms of variation in a smaller number of underlying components (each component being a weighted sum of the dimensions). PCA reduced the eight dimensions to just two underlying uncorrelated components, capturing 77% of the variance. Once two components had been "extracted", they were then "rotated" with Varimax rotation (Tabachnik & Fidell, 2007). The aim of rotation is to allocate each of the eight dimensions clearly to just one component, thus simplifying the interpretation of the analysis without loss of information. Each component is a weighted sum of the eight dimensions. The weight or 'loading' of a dimension on a component, Component 1 say, indicates how much of the variance in people's ratings along that dimension can be explained by variation in Component 1. These loading patterns are shown in Figure 2.[5]

[4] The Kaiser-Meyer-Olin (KMO) measure of sampling adequacy was 0.69, indicating the data were suitable for this type of factor analysis (Tabachnick and Fidell, 2007). Conventionally, a minimum value of 0.5-0.6 for KMO is required.

[5] An additional PCA was conducted with data transformed into z-scores within each domain, so that correlations were not influenced by mean domain differences on dimensions. The same two-component pattern was seen.

Table 2. Correlations (*r*) Between Dimensions, Across 160 Concepts for Study 1.

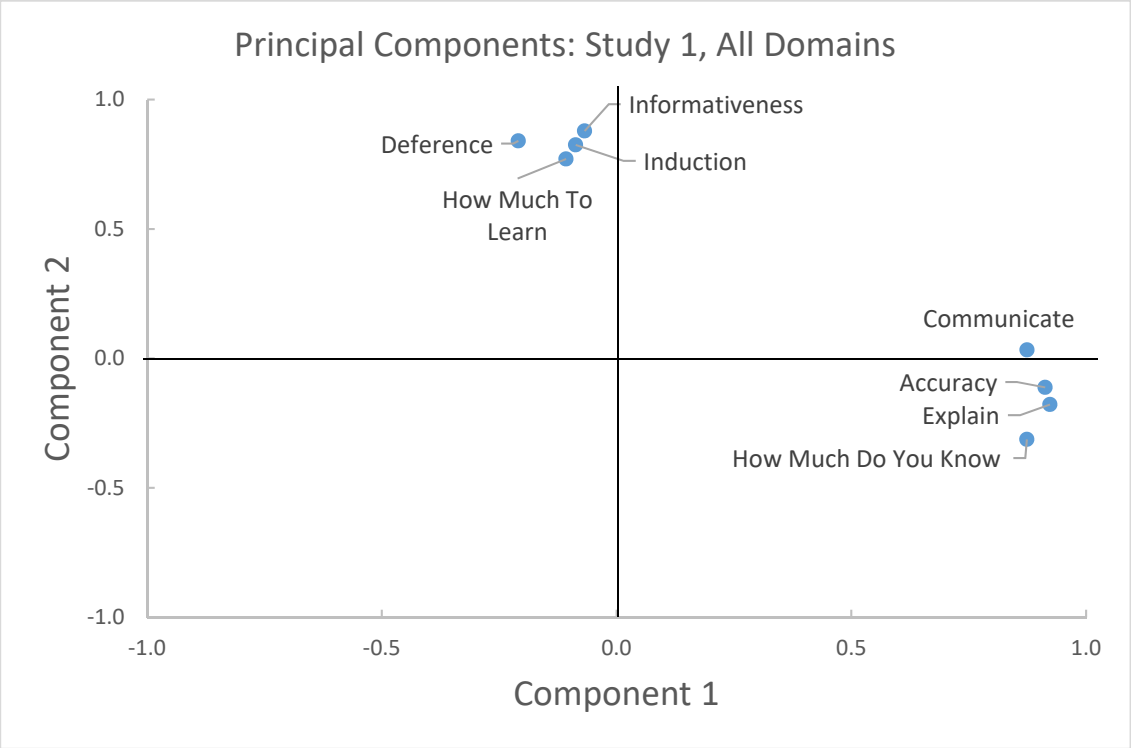
		Accuracy	How Much You Know	Explain	Communicate	Induction	Informative	Deference
1	Accuracy							
2	How Much You Know	.86**						
3	Explain	.79**	.82**					
4	Communicate	.68**	.64**	.79**				
5	Induction	-.14	-.28**	-.27**	-.10			
6	Informative	-.17	-.36**	-.22**	-.03	.76**		
7	Deference	-.26**	-.44**	-.34**	-.15	.64**	.59**	
8	How Much To Learn	-.21**	-.31**	-.18	-.10	.40**	.59**	.67**

KEY

> .8	> .6	> .4	> .2	0	< -.2	< -.4	< -.6	< -.8

Note: * p < .01, ** p < .001. (Bonferroni corrected alpha for family-wise Type I Error of .05 = .0018, r_{crit} = .25.)

Figure 2: Component loadings for the eight dimensions in Study 1, across all domains of concepts.



As seen in Figure 2, variation in the eight dimensions of appraisal could be explained by variation in two underlying components, each combining four of the eight dimensions. Component 1 combines the dimensions ***C****ommunicate*, ***A****ccuracy*, *How Much You* ***K****now* and ***E****xplain*. In order to separate the operational measures from the psychological construct, we use the shorthand 'CAKE' for the four dimensions and, for reasons explained below, the term 'sense of understanding' or 'SoUnd' for the psychological construct (Component 1). Component 2 combines the dimensions *How Much To* ***L****earn*, ***I****nduction*, ***D****eference* and ***I****nformativeness*. 'LIDI' is shorthand for these four dimensions. Since these dimensions do not cluster consistently in our subsequent experiments we do not introduce a term for a

psychological variable corresponding to Component 2.[6]

To probe whether these two components represent a general underlying structure, we examined whether this two-component structure is maintained within each conceptual domain. A separate PCA was performed for each of the four domains. If there are indeed just two underlying forms of appraisal, then the eight dimensions should continue to cluster into two components within each domain. Alternatively, given the very different kinds of concept found in social, health or biological domains (Dahlgren, 1985), differences in how appraisals are structured may appear. It should also be noted that when considering individual domains our data is at the limit of the recommended minimum ratio of cases to variables for PCA (5:1).

Figure 3 shows the component structure that emerged from the four within-domain PCAs. People, Recreations and Health showed a clear two-component structure, whereas the Biology domain produced a three-component structure. In all four domains, the four CAKE dimensions (*Communicate, Accuracy, How Much Do You Know, Explain*) formed a clear component. This suggests that there is a single factor (SoUnd) underlying assessments along these four dimensions. To give a sense of the significance of this factor, Figure 4 shows how the 40 social categories (the People domain) are distributed with respect to it.

By contrast, the structure of the other four dimensions (LIDI) varied across domains. This variability suggests that there is no corresponding single factor underlying appraisals of *How Much To Learn, Induction, Deference* and *Informativeness*. Although reliably rated, they come out as independent dimensions of epistemic appraisal.

---

[6] The four dimensions in Component 2 (LIDI) had low negative loadings on Component 1 (appearing to the left of zero on the horizontal axis), indicating a weak tendency for higher ratings on Component 2 to be correlated with lower ratings on Component 1.

Figure 3: Principal component plots for the four domains in Study 1. The grey ovals show the four dimensions of Sense of Understanding (SoUnd), consistently loading on Component 1.

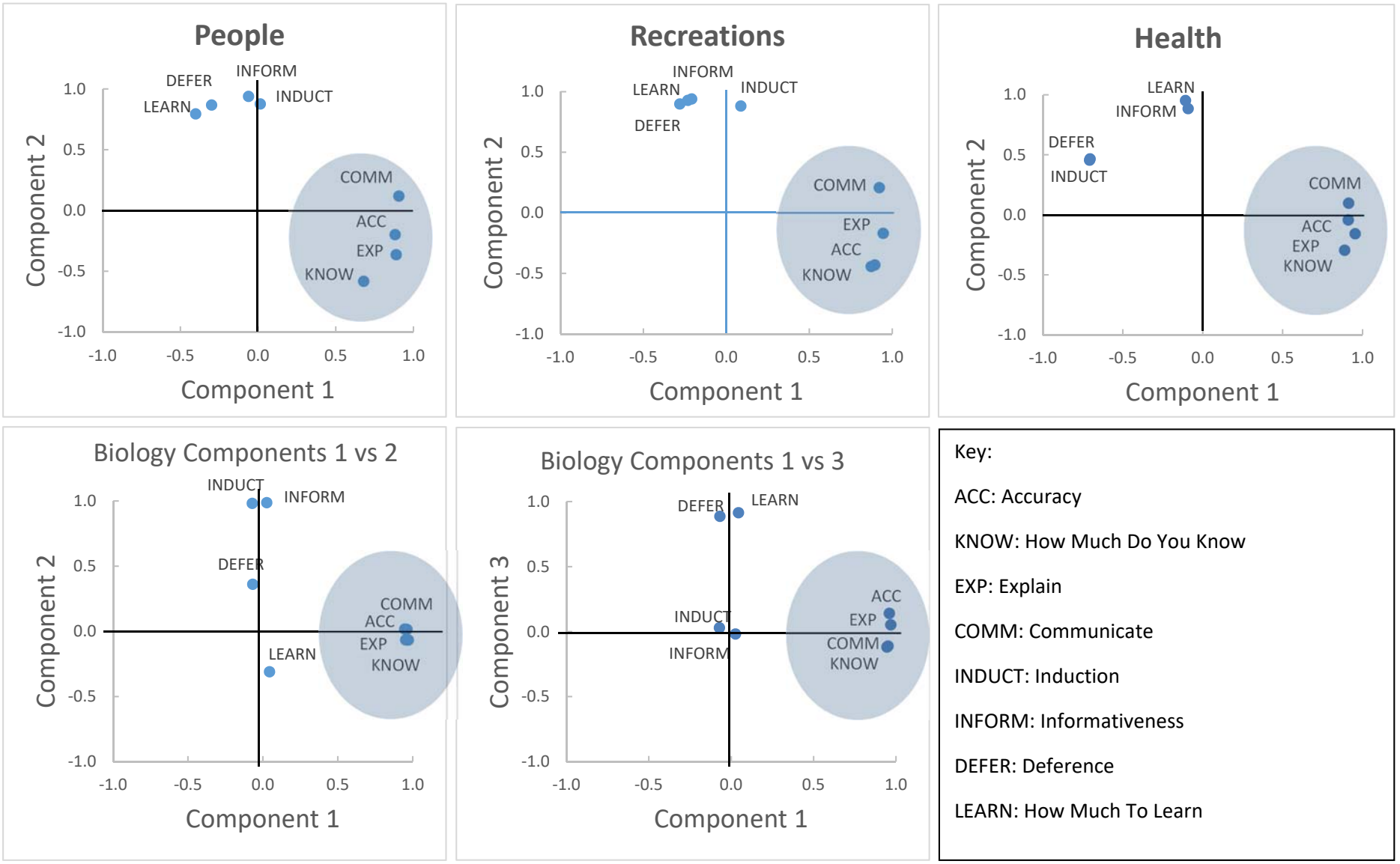
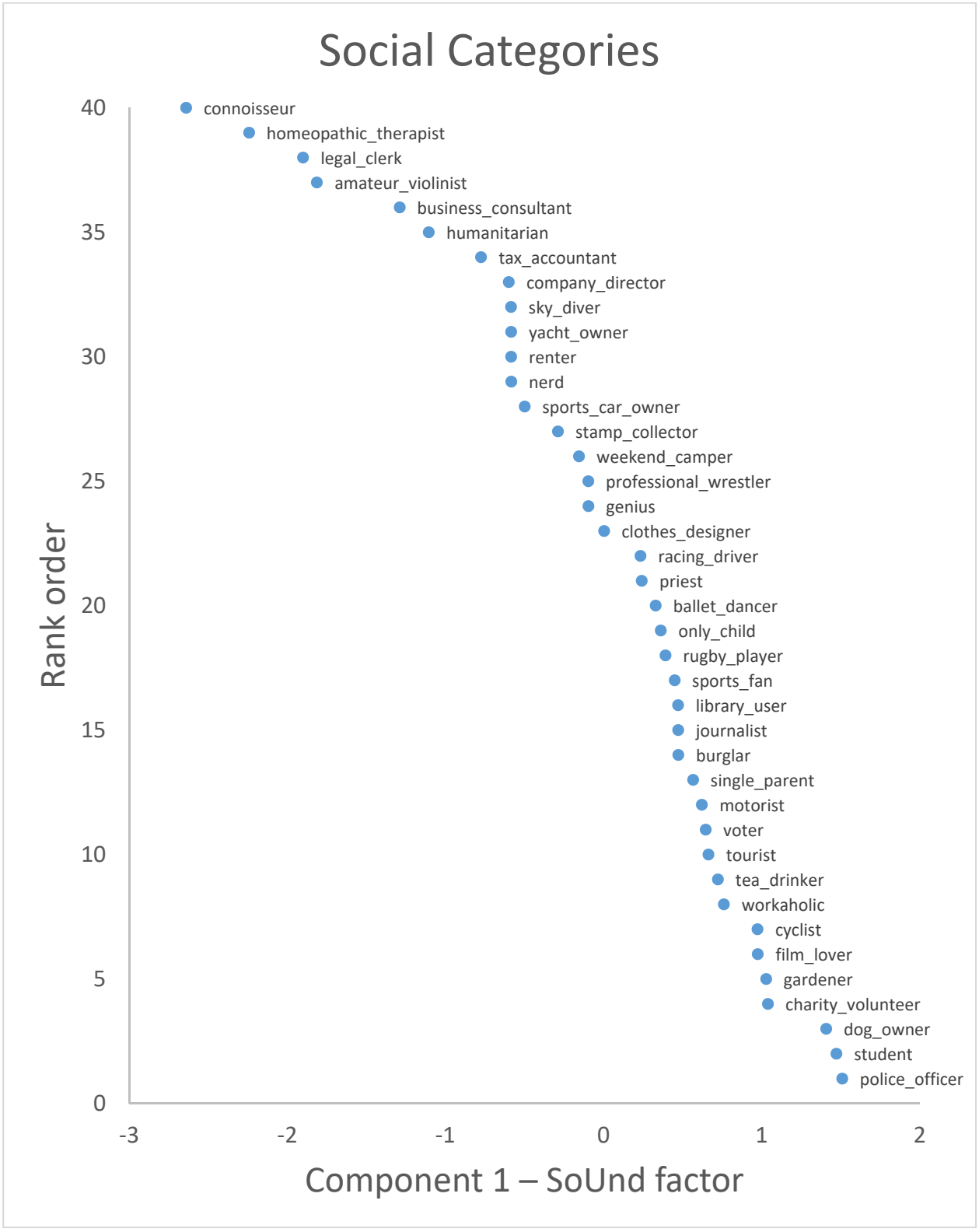


Figure 4. Plot of concepts in the People domain, showing distribution in the Sense of Understanding (SoUnd) factor discovered in Study 1.



2.1.3 Discussion

Study 1 was exploratory. Nevertheless, based on the philosophical literature, we expected that some or all of our dimensions would be reliably rated between subjects. We found that in fact all eight were. For each dimension, different people make similar judgements about how a range of concepts score on that dimension.

We went on to ask whether judgements about some of the different dimensions were driven by a single underlying factor. We uncovered evidence that a single factor captures four of the dimensions (*Communicate, Accuracy, How Much Do You Know* and *Explain* - CAKE), implying that a common intuition underlies these different judgements. This factor explained variance on these dimensions even within each conceptual domain. The other four dimensions did not relate to an underlying factor. *How Much To Learn, Deference, Induction* and *Informativeness* (LIDI) clustered differently in different domains, evidence that, although participants agree on how concepts should be scored on each of these dimensions, these appraisals vary independently.

The dimensions *Accuracy, How Much Do You Know* and *Explain* were drawn from our first broad family of epistemic appraisals, those concerned with the thinker's own understanding of a concept. The underlying factor we discovered (underlying the CAKE dimensions) includes a fourth dimension, *Communicate*, which we had thought could be in either broad family: connected to self-understanding, or alternatively a matter of assessing the value of a concept as a cognitive tool. Study 1 suggests it is the former. The psychological variable which underlies judgements about these four dimensions is a form of self-understanding or epistemic self-audit, so we label this factor 'sense of understanding' or 'SoUnd'. Study 1 provides preliminary evidence that SoUnd is a property of concepts.

2.2 Study 2: Replication with a Standard Sample of Concepts

Study 2 replicated the procedure and analysis of Study 1 with a new set of concepts. In

Study 1 we deliberately selected concepts which we believed would vary on our eight dimensions. A question remains as to what will emerge using a naturalistic sample of concepts. Will the dimensions still be reliable, and will the SoUnd factor still emerge consistently? Study 2 took concepts from taxonomic category norms used in previous research, expanding the domains to foodstuffs and artefacts.

2.2.1 Method

2.2.1.1 Participants. Using Prolific Academic, 419 participants were recruited (309 Female, 99 Male, 11 unspecified) for a small monetary reward (Age 17-73; M_{Age} = 35.8). Of these, 342 were native speakers, and the remainder fluent or competent in English. Fifteen participants failed to complete the questionnaire, leaving a final N of 404.

2.2.1.2 Materials, Design and Procedure. Participants completed the same questionnaire as in Study 1 with a new list of concepts. They were randomly allocated to one of eight groups, each group rating 160 items on just one of the dimensions of appraisal, as before. Concepts were drawn from social (Sports), biological (Fruit and Vegetables in a combined list), and two different artefact domains (Clothing and Furniture). In each domain, 40 concepts were selected at random from a standard list of taxonomic categories, Hampton and Gardiner (1983) (see Supplementary Materials, (Appendix B).

2.2.2 Results

2.2.2.1 Inter-subjective Agreement. As with Study 1, we first investigated which dimensions were rated reliably between subjects. Did different participants applying a given dimension rate the 160 concepts in a similar way on the five-point scale provided?

As in Study 1, to assess reliability of the scales across the full set of 160 categories, it was necessary for participants to have a complete set of responses. Missing data were handled as follows. Thirty-nine participants (9%) with 5 or more (out of 40) missing data points on any single domain were excluded from the reliability analysis. Where there were

fewer than 5 missing responses, the missing responses were replaced with the mean. Figure1 shows reliability (Cronbach's α) for the eight dimensions. In addition, the standard deviation for category means is shown, since low variance across the means will contribute to lower measured reliability. Alpha was above .9 for six of the dimensions. *Induction* (.820) and *Informativeness* (.738) had lower values, attributable to low variance (SD of 0.3 or lower, see Fig. 1). Given that the categories were not sampled with the aim of generating variability on the dimensions (unlike in Study 1), the low variance for some dimensions is not unexpected.

2.2.2.2 Factor Structure Analysis. For each dimension, the mean rating for each of the 160 concepts was calculated. Correlations between the dimensions were calculated, and the four dimensions of SoUnd seen in Study 1 again correlated strongly (CAKE). The remaining four dimensions (LIDI) showed no clear pattern of correlation.

The mean ratings for most dimensions showed substantial differences between domains (Sport, Fruit and Vegetables, Clothing, Furniture).[7] To focus just on within-domain correlations, dimension scores were transformed into z-scores so that each domain had a mean of 0 and SD of 1 on each dimension. The data were then submitted to PCA. Two components captured 74% of the variance. Figure 5 shows the way the eight dimensions load onto these components. Component 1 is a cluster of the four SoUnd dimensions. Component 2 is represented by *Deference* and *How Much To Learn*. As described above, *Informativeness* and *Induction* suffered from low reliability and low variance within domains and so were less well correlated with either component. Where a domain lacks variation on a dimension, there is limited scope to measure correlations of that dimension with others.

[7] Such differences in Study 1 were smaller and had no notable effect on the PCA solution (see footnote 5).

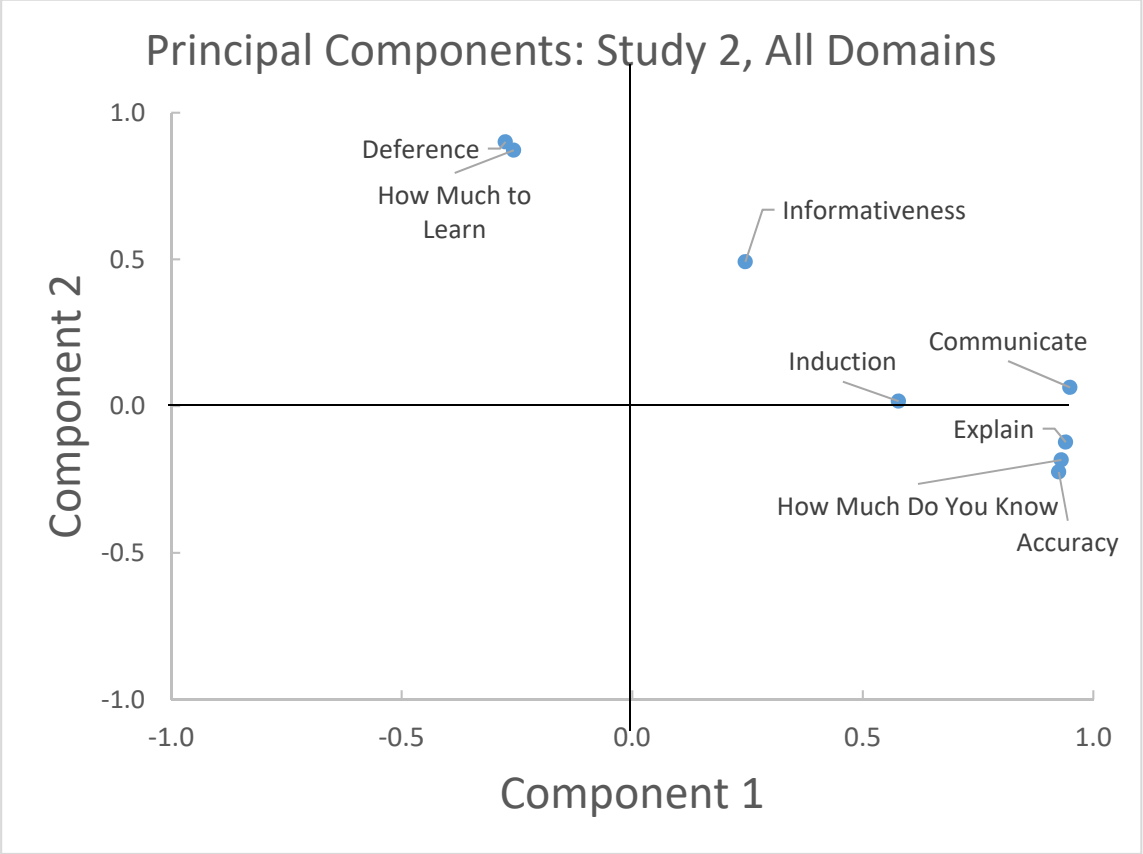
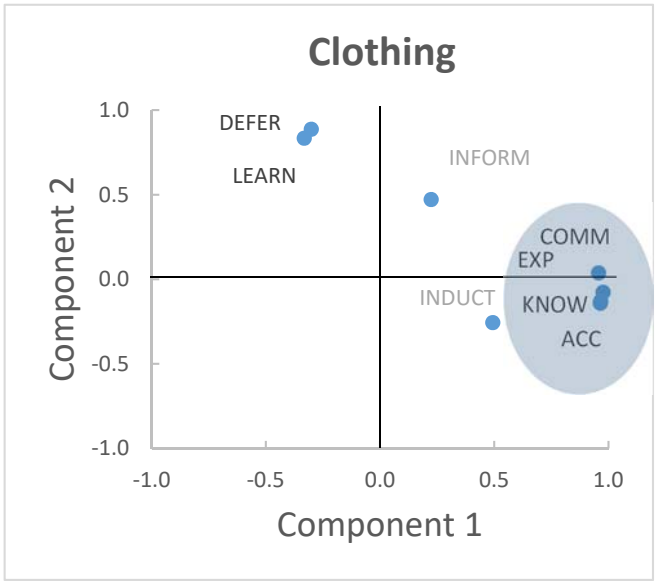
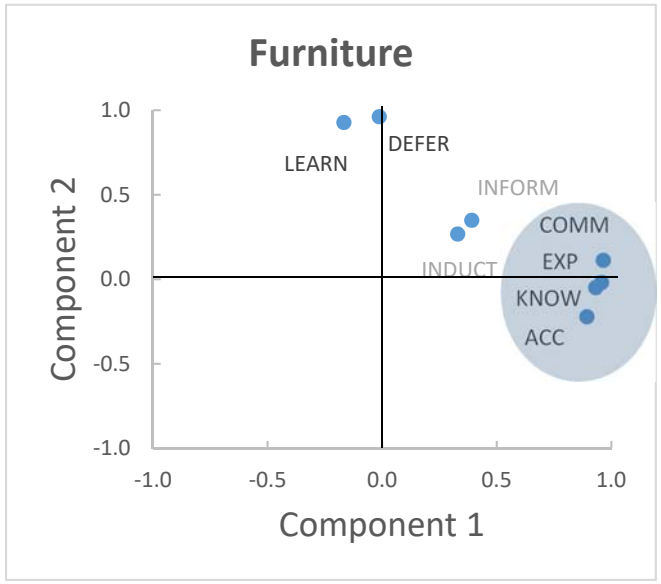
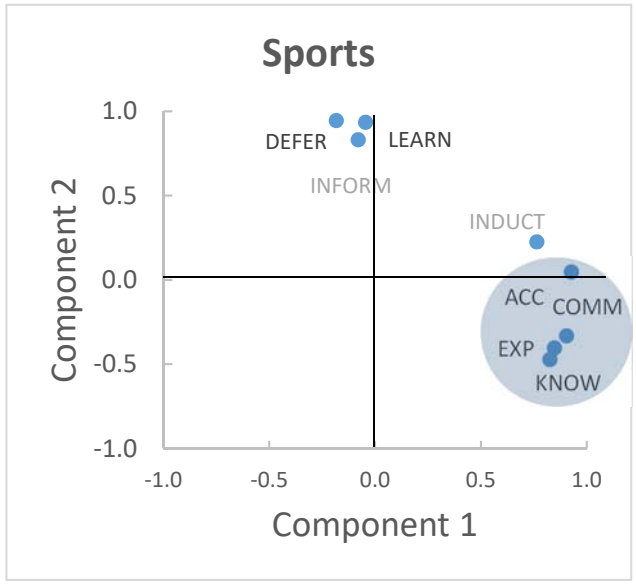


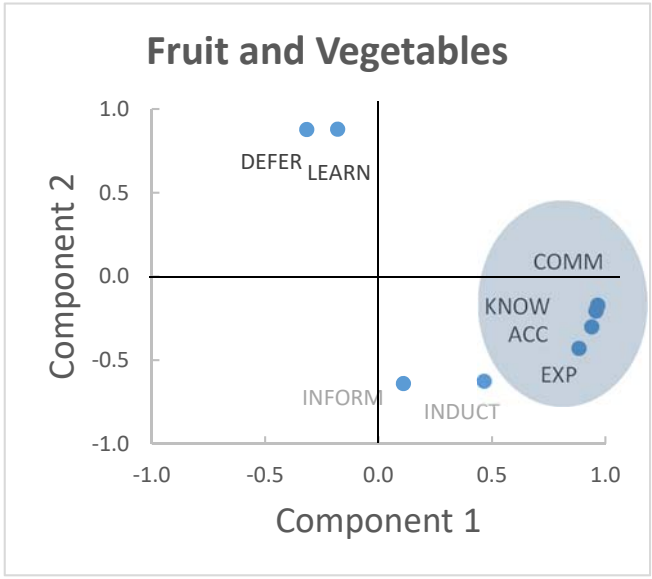
Figure 5. PCA loading plot for Study 2, all domains together with means superimposed.

As in Study 1, we performed a PCA with Varimax rotation within each domain. All four domains produced a two-component solution, as shown in Figure 6. Where reliability for a dimension dropped below .75, the label has been greyed out, to indicate that its position on the graph is unreliable. Component 1 captures the variance in the four SoUnd dimensions (*Communicate, Accuracy, How Much Do You Know* and *Explain*) in all four domains. *Deference* and *How Much To Learn* came out as strongly loaded on Component 2 in all domains. There was no clear cross-domain pattern for *Induction* and *Inference*, which in any event showed low reliability. Although *Deference* and *How Much To Learn* clustered together in the domains covered in Study 2, Study 1 found them to vary independently across other conceptual domains, suggesting they have no underlying common factor.









Key:

ACC: Accuracy

KNOW: How Do You Know

EXP: Explain

COMM: Communicate

INDUCT: Induction

INFORM: Informativeness

DEFER: Deference

LEARN: How Much To Learn

Figure 6. Loading plots for PCA within domains for Study 2.

Note: Greyed out dimensions had lower reliability (alpha < .85). Oval shaded areas contain the SoUnd dimensions.

2.2.3 Discussion

Replication with a naturalistic sample of concepts showed the same broad pattern seen in Study 1. For each of the four CAKE dimensions, ratings showed inter-subjective agreement. Ratings correlated across these dimensions and could be accounted for by a single underlying common factor, SoUnd.

Deference and *How Much To Learn* were also reliably rated between subjects for the 160 concepts used in Study 2. *Induction* and *Informativeness* were not, but that finding is hard to interpret since the domains of concepts we used generated little variation in judgements along these two dimensions.

Taken together, Studies 1 and 2 delivered the first evidence that some of the forms of epistemic appraisal of concepts that were suggested to us by the philosophical literature are indeed psychologically real: a factor SoUnd, a varying sense of how well thinkers understand different concepts, which drives judgements along the dimensions *Communicate, Accuracy, How Much Do You Know* and *Explain*; separate, independent appraisals of the dimensions *Deference* and *How Much To Learn*; and also some evidence of independent appraisal of *Induction* and *Informativeness* (in Study 1).

2.3 Study 3: Taxonomic Levels

Having discovered some forms of reliable concept appraisal, we next investigated whether these forms of appraisal are integrated with other aspects of the structure of concepts. In Study 3 we did that in relation to taxonomic structure.

Many concepts are organised hierarchically, for example MAMMAL-DOG-SPANIEL. There is considerable evidence that a representation of taxonomic level is part of the structure of concepts. In particular, the intermediate or "basic level" (e.g. DOG) has been found to be psychologically important. It acts as a preferred level for categorisation and other tasks. Basic

level categories show more within-category similarity than superordinate level concepts, while retaining considerable between-category dissimilarity, unlike subordinate level concepts (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). The basic level has been identified as the highest level at which category members are easily imageable and share the same parts (Corter & Gluck, 1992; Tversky & Hemenway, 1984).

We predicted that, if the forms of appraisal we had discovered were part of the structure of concepts, then they would predict aspects of taxonomic structure. Our method was to obtain ratings on the eight dimensions of appraisal for a new set of concepts, concepts selected in hierarchically-organised triples: superordinate-basic-superordinate. We expected that basic level concepts, being the preferred cognitive tool for many concept-involving tasks, would carry the highest sense of understanding. So we predicted that a high rating on the SoUnd factor (a weighted sum of the four CAKE dimensions) would correlate with a concept's being found at the middle level in these triplets, the basic level.

By contrast, as categories become increasingly specific, learning that an item belongs to the category continues to become more and more informative, and inductions about the category become increasingly reliable. A classic study by López, Atran, Coley, Medin, & Smith (1997) showed that U.S. students find induction based on folk-generic categories like ROBIN to be stronger than that based on the basic level categories identified by Rosch et al. (1976), e.g. BIRD, which are at a higher taxonomic level. Accordingly, we predicted that *Induction* and *Informativeness* would increase when moving down through the hierarchy from superordinate to basic and subordinate level concepts. We had no firm predictions about *How Much To Learn* and *Deference* although, as noted in the introduction, if people think that having more members (being more 'inclusive', Égré & O Madagáin, 2019) implies that there is more to learn about a category, then *How Much To Learn* would decrease down the hierarchy, as more specific categories have fewer members.

2.3.1 Method

2.3.1.1 Participants. Participants were 341 adults (235 Female, 105 Male, 1 unspecified), recruited through Prolific Academic for a small monetary reward (Age 18-72; M_{Age} = 35.2). Of these, 293 participants were native speakers of English and the remainder self-rated as fluent or competent. Nine participants failed to complete, leaving a final sample of 332.

2.3.1.2 Materials, Design and Procedure. The first step was to generate 53 suitable concept triplets, so as to give us 159 concepts to use in the dimension-rating task.[8] Initially, the five authors together came up with 65 hierarchically-organised triplets (e.g. MAMMAL-DOG-SPANIEL), with no overlaps. Pretesting was performed to select triplets that were suitable in terms of familiarity and hierarchical organisation. All words in a triplet had to be highly familiar (mean 4.5 or more on a 1 to 5 scale). Imageability ratings had to confirm that the middle term was indeed at the basic level: it was required to have an imageability rating of at least 4 out of 5, and to be at least as imageable as the two other terms. We also asked participants to judge whether each lower-level category did indeed belong in the class above, requiring that each class inclusion relation was confirmed by at least 95% of participants. Details of pretesting are in the Supplementary Materials (Appendix C).

A new set of participants were then allocated at random to one of eight groups, each group rating all 159 concepts, presented in a random order, on one of the eight dimensions. Participants completed the questionnaire online through Qualtrics.

2.3.2 Results

2.3.2.1 Inter-subjective Agreement. The data were scanned for missing values. Thirteen participants (4%) with missing data were excluded from further analyses. The final

[8] Studies 1 and 2 used 160 concepts, but 160 is not divisible by 3.

number of participants for each level for each dimension ranged between 37 and 42. All dimensions had reliability (α) greater than 0.85. Values are shown in Figure 1, together with SD for the category means. Thus, with this new set of concepts, we replicated the first result of Studies 1 and 2: that people reliably judge concepts along these dimensions in the same way.

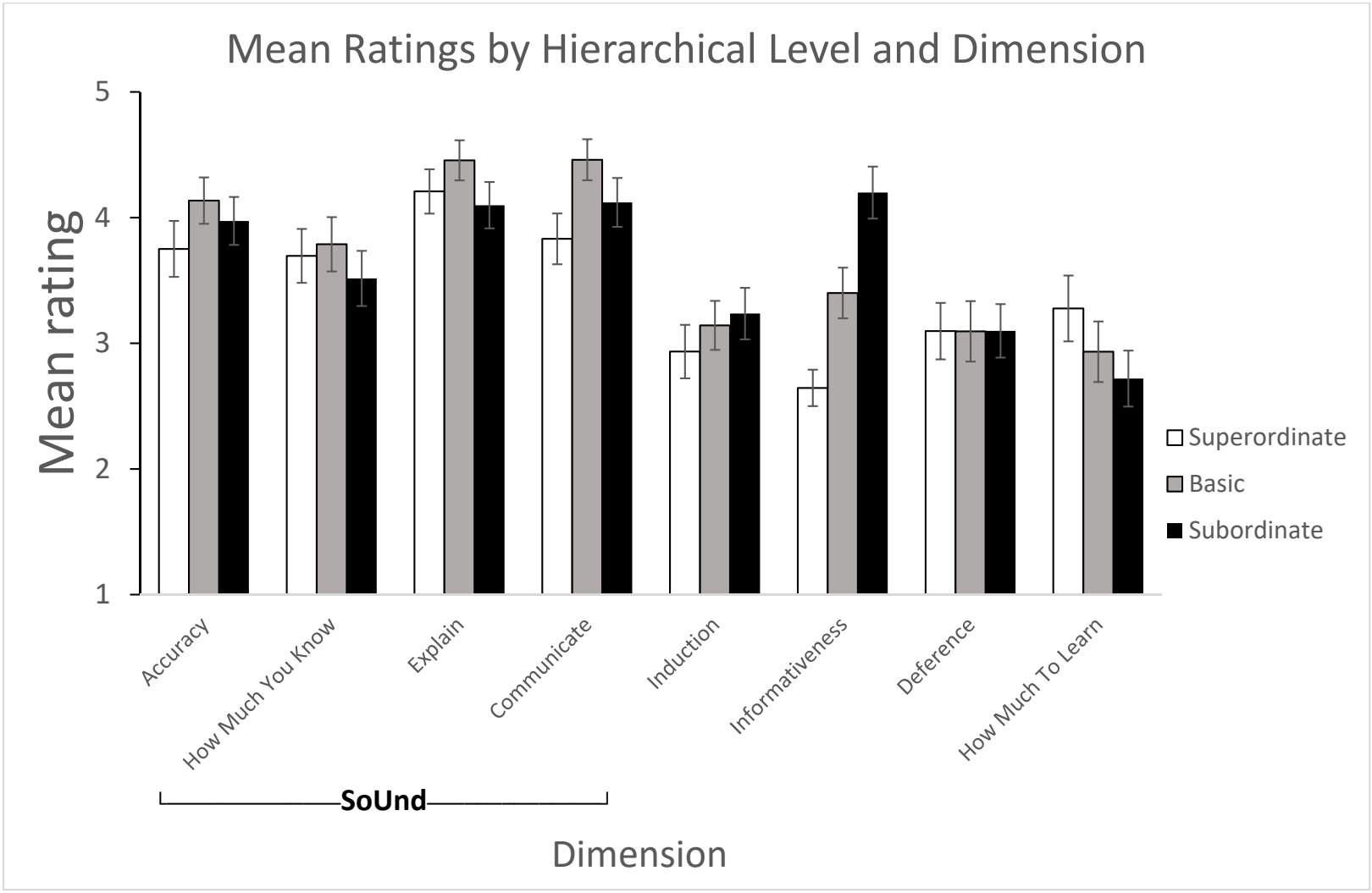
2.3.2.2 Differences Between Hierarchical Levels. For each dimension, the mean rating of the 53 items at each of the three levels was calculated (see Fig. 7). The means were submitted to a mixed 8x3 ANOVA with Dimension between-subjects with eight levels, and Level within-subjects with 3 levels. There were significant effects of both Dimension (F (7, 324) = 29.3, $p < .001$, $\eta_p^2 = .39$), and Level (F (2, 648) = 61.2, $p < .001$, $\eta_p^2 = .16$), and a significant interaction (F (14, 648) = 49.4, $p < .001$, $\eta_p^2 = .52$). The effects were as strong when treating items as a random effect ($\eta_p^2 = .56$, .59 and .44 respectively). There was therefore strong evidence that the dimensional means varied across levels, and in different ways. A breakdown analysis considered the Level factor for each dimension individually. All effects of Level were significant at $p < .001$, with the exception of a weaker effect for *Induction* ($p = .01$), and zero effect for *Deference (F < 1).* (With Bonferroni correction for 8 comparisons, α = .006, and *Induction* fails to reach significance.)

All four SoUnd dimensions showed the predicted pattern of higher ratings for the basic level, thus further validating the SoUnd factor constructed out of these four dimensions. Basic level concepts also tend to be more familiar, and although we only chose concepts with high familiarity, there was still some variation within the set of items. Entering familiarity as a covariate in the ANOVA showed it to be a significant predictor of ratings for all four dimensions. However, three of the dimensions (*Communicate, Accuracy* and *Explain*) still showed strong significant effects of level when controlling for familiarity.

For the dimensions of *Induction* and *Informativeness* there was, as predicted, a linear

increase from superordinate through basic to subordinate levels. As the category became more specific, so within-category similarity increased, and with it the perceived inductive potential and informativeness of the category. The effect however was much stronger for *Informativeness* ($\eta_p^2 = .87$) than for *Induction* ($\eta_p^2 = .09$). A trend in the reverse direction was seen for *How Much To Learn*. People think there is more to learn about a category if it has more members. Interestingly, for *Deference* the means were all equal and close to the midpoint of the scale.

Figure 7. Mean ratings for each dimension as a function of level for Study 3. Error bars show 95% CI. The four dimensions to the left relate to sense of understanding and show a consistent basic level advantage.



2.3.3 Discussion

The dimensions of *Communicate, Accuracy, How Much Do You Know* and *Explain* all showed the highest ratings for basic level concepts. Studies 1 and 2 provided evidence that judgements on these dimensions are driven by a common underlying factor, a sense of how well the thinker understands a given concept (SoUnd). The fact that they behave the same way in relation to taxonomic structure is further evidence that the four CAKE dimensions are driven by a common factor. Study 3 provides evidence that SoUnd is integrated with a much-studied aspect of conceptual structure, namely taxonomic organisation, and thus that it forms part of the structure of concepts.[9] The highest SoUnd attaches to basic level concepts, people's preferred way of categorising the world (Rosch, 1977). Post hoc analysis confirmed that the effect was not driven solely by differences in familiarity.

In keeping with the earlier studies, we found that the dimensions *How Much To Learn, Induction, Deference* and *Informativeness* (LIDI) worked in different ways. We predicted that subordinate categories would be rated as higher on the *Induction* and *Informativeness* dimensions. The higher within-category similarity found at the subordinate level should make a concept a better basis for induction. Further differentiation from the basic level should make the concept more informative. This prediction was borne out. On the other hand, superordinate categories were rated more highly on *How Much To Learn*. This result would make sense if people envisage a wider diversity of members to learn about. We observed no variation in *Deference* across levels.

2.4 Study 4: Induction with Single Categories

Studies 4 and 5 continue to explore the hypothesis that concept appraisal forms part of

[9] Our method did not test whether SoUnd is more predictive of basic level categories than other measures (e.g. Corter & Gluck, 1992). We did not set out to test whether SoUnd is the basis on which concepts are taxonomically organised.

the structure of concepts. Taking the dimensions of concept appraisal validated in the first three studies, we now ask whether they are integrated with another central feature of concepts, namely their role in category-based induction.

As noted in the introduction, a category provides a better basis for induction if members of the category reliably share many properties (Égré & O Madagáin, 2019; Millikan, 2000). Previous studies have shown that beliefs about within-category similarity (category coherence) correlate with judgements in inductive reasoning tasks (Patalano, Chin-Parker, & Ross, 2006; Patalano, Wengrovitz, & Sharpes, 2009), and are related to the entitativity scale in the Haslam et al. (2000) study of essentialism. Based on this research, we predicted that there would be dimensions of concept appraisal that predict people's inductive judgements. Our dimensions of *Induction* and *Informativeness* ask explicitly about two aspects of category coherence – how reliably are properties shared between members, and how many properties are shared? The study was exploratory with respect to the other dimensions of appraisal, although we expected that concepts about which people have a high sense of understanding, as measured by the SoUnd construct, might have an inductive advantage, as might those about which people showed *Deference*, being more kind-like or scientific; similarly for high ratings in *How Much To Learn*.

In this initial study we did not aim to establish whether dimensions of concept appraisal can act as better predictors of inductive judgements than other measures. We simply wanted to investigate how, if at all, concept appraisals are integrated with the inductive structure of concepts. The method employed was correlational. The dependent variable was the judgement that a novel property, found in three exemplars of a category, was likely to be found in another randomly selected category member. Using partial correlation, we aimed to identify whether the dimensional ratings collected in Study 1 would predict inductive judgements. Since class size could have an effect on inductive judgements (Nelson & Miller,

1995), we also collected ratings of frequency for each category, so as to ensure that beliefs about the size of a category were not a cofounding factor in the results.

2.4.1 Method

2.4.1.1 Participants. One hundred and twenty-five participants (85 Female, 38 Male, 2 unspecified) were recruited (Age 18-68; M_{Age} = 34.44). Of these, 103 participants were native speakers, and the remainder were fluent or competent in English. Two participants were excluded for failing to complete the study.[10]

2.4.1.2 Materials, Design and Procedure. Participants were randomly assigned to one of three groups each completing 28 induction problems in one of three domains: People (N = 42), Recreations (N = 41), and Health (N = 40). All problems followed an identical structure. The premise stated that three members of a category (e.g., "police officers", "people who swim") were found to have a particular characteristic (e.g., personality type X, or an increase in levels of X). Participants were then asked to indicate the probability of another member of the category showing the same characteristic on a sliding scale of 0 ('no more than chance') to 100 ('a lot above chance'). The following is an example from the People domain:

> *A psychologist studying personality has noticed that people belong to one of two different personality types: Personality type X and personality type Y. In her study, she noticed that three people who were library users had personality type X. How likely is it that the next library user she tests will also have personality type X?*

Using the means from Study 1, we selected categories to maximize differences along the eight dimensions across the 28 problems in each domain, while avoiding strong correlations between the cluster of four SoUnd dimensions and the other four. To this end, pairs of categories were selected to be maximally contrasted on either the SoUnd factor (a

[10] Sample size was dictated by consideration of previously reported effect sizes.

weighted sum of the CAKE dimensions), or on the sum of the other four dimensions (LIDI).[11]

We selected 16 pairs of categories in each of the three domains, eight matched on one set of dimensions and differing maximally on the other, and eight with the reverse contrast. Categories were selected in matched pairs as we needed materials in this form for use in Study 5, where categories were pitted against each other directly. Because of some overlap between pairs, additional concepts were randomly selected to bring the total number of concepts in each domain to 28. All the induction problems and concepts used are provided in the Supplementary Materials (Appendix D).

Each domain used two scenarios, each with 14 categories to judge. Participants were instructed that there was no right or wrong answer. The scenarios were presented in a fixed order, and the order of categories was counterbalanced. Upon completion, participants were able to provide comments on how they made their judgements. Finally, participants judged how many people were in each category in the context of the problem scenario, using a Likert scale of 1 (very few) to 5 (very many).

2.4.2 Results

2.4.2.1 Exclusions and Transformation. Participants with 3 or more of the 28 induction responses missing (N = 2) or who gave the same rating to all questions (N = 7) were excluded,[12] leaving a final sample of 114. Otherwise, missing data were replaced with the participant mean. Similar exclusion criteria were applied to the class size questions, leaving a sample of 108 (of 125) for the measure. Reliability checks on the dependent variable showed alpha between .43 and .81 across domains. Mean inductive responses were calculated for each category on the scale from 0 to 100. Within domains, mean and (SD) were

[11] For the materials selected, the four LIDI dimensions were positively correlated with each other within each of the three concept domains.

[12] Including these participants did not contribute to any differences between scenarios.

as follows: People, M = 36% (6%), Health, M = 46% (8%), and Recreations, M = 57% (3%). Distributions did not deviate from normal.

2.4.2.2 Induction Responses. The starting point was the concept appraisals collected in Study 1: the mean rating of each concept on each dimension. To see whether these appraisals predicted inductive judgements, correlations were first calculated within each of the domains between the inductive response and each of the eight dimensions, and then averaged over domains. The four dimensions comprising the SoUnd factor all had mean correlations close to zero (< 0.1), showing no relation between people's average ratings on these dimensions (CAKE) and inductive judgements. The four remaining dimensions all had positive correlations, with *Deference* not statistically significant ($r = .311$, $p = .11$), but *Induction, Informativeness* and *How Much To Learn* all significantly above zero ($r = .489$, .482, and .483, $p < .001$).

Analysing the data across all 84 problems, two controls were introduced. First, we controlled for the strong differences between mean inductive responses across domains. Second, given that the *Induction* question in Study 1 is closely related to the question about induction asked in this study, we chose to hold the *Induction* dimension constant to discover whether the other seven dimensions were playing an additional role in the judgements in this study. Partial correlations were calculated between the inductive response and each of the seven dimensions, holding constant *Induction*, and with dummy-coded binary variables to equate domain differences. Results are shown in Figure 8.

Figure 8. Partial correlations of seven dimensions with the induction judgements provided in Study 4, holding constant domain differences and the *Induction* dimension. The dotted line shows the critical value of r for .05 significance with Bonferroni correction (r(79) = .31).

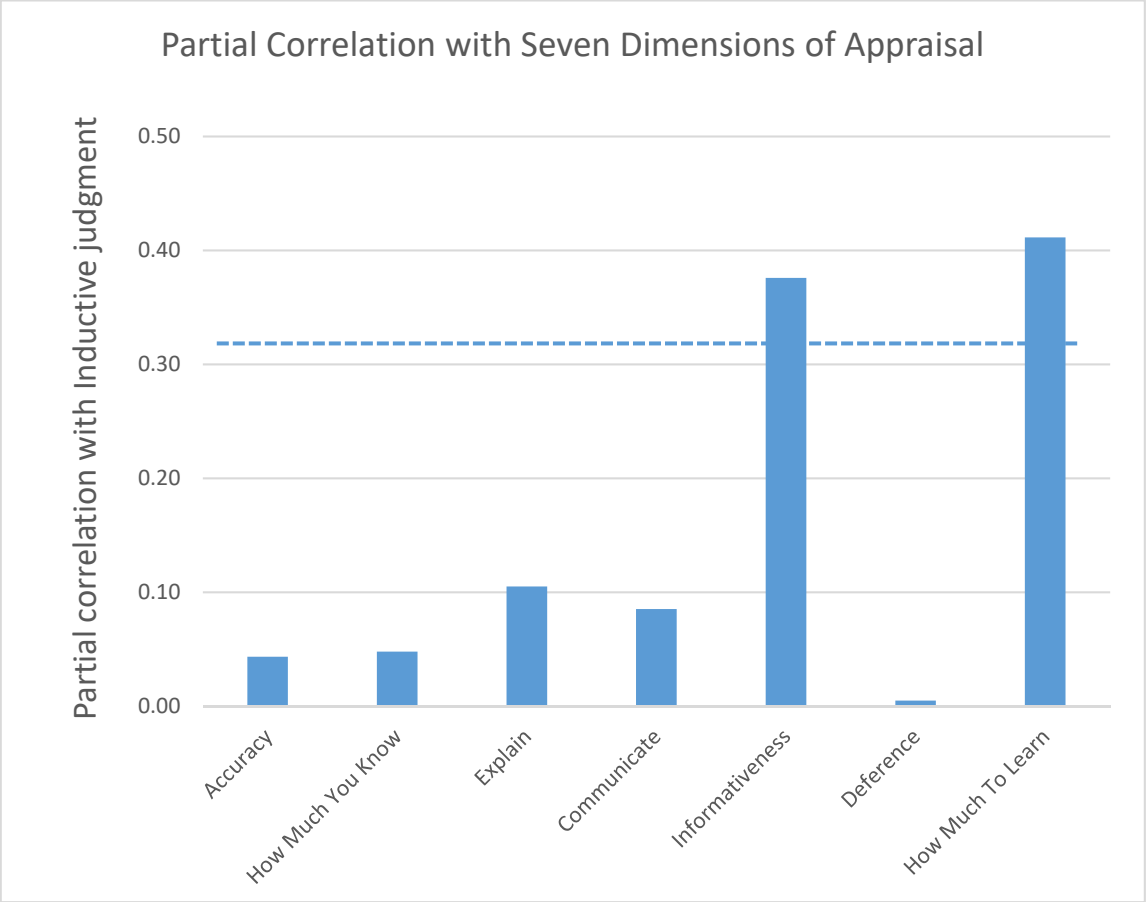


Figure 8 reveals that two dimensions predicted how much people felt that a novel property would also be found in other category members, over and above the *Induction* ratings in Study 1. These were *Informativeness* and *How Much To Learn.* None of the four dimensions comprising the SoUnd factor played a significant role in the judgement. We can conclude that the *Informativeness* and *How Much To Learn* dimensions of appraisal, in addition to *Induction*, are integrated with the aspects of conceptual structure that drive the way concepts are relied on as a tool for induction. Class size was also examined as a possible predictor of inductive judgements, but had no significant effect (partial correlation holding

domain constant = -.006, $p > .05$).

2.4.3 Discussion

The results further supported the hypothesis that the dimensions of stable concept appraisal examined in Studies 1 and 2 are integrated with the structure of concepts. In addition to *Induction*, the way a concept is rated for *Informativeness* and *How Much To Learn* were found to predict the extent to which the corresponding category would be relied on inductive inference. Category-based induction is one of the most important ways concepts are used as a cognitive tool (Millikan, 2000). Study 4 shows that appraisals of *Informativeness* and *How Much To Learn* are part of, or closely related to, those aspects of conceptual structure that are relied on in performing inductive inference.

2.5 Study 5: Induction with Two Competing Categories

In this final study we sought to probe more deeply into the relationship between concept appraisal and inductive judgements. Using a procedure from Patalano et al. (2006; see also Nelson & Miller, 1995), we examined whether dimensions of concept appraisal predict people's inductive judgements when two concepts are pitted against one another. For example, people might feel that ANOREXIA is a stronger basis for induction than SNEEZING. From Study 4, one would expect being highly rated for *Informativeness* and *How Much To Learn* should make a category a preferential basis for induction; also *Induction*, if the experiment is a valid test of inductive preference. Furthermore, by giving participants a forced-choice test, there would be a stronger chance of detecting any influence of the SoUnd-related dimensions on inductive judgements.

In this study two concepts appeared in each question. A scenario was designed in which members of one category were found to have a certain property and members of another category were found to have a contrary property. Presented with an individual who

belonged to both categories, participants were asked which of the two contrary properties would transfer to this individual.

Participants also provided confidence ratings. Previous research on metacognition has shown that feelings of confidence are often indicative of metacognitive processes in different kinds of judgment (Fleming & Daw, 2017; Koriat, 2011; 2015; Proust, 2012). In the present context, we wanted to know whether inductive preference predicted by dimensions of concept appraisal would be accompanied by an increase in confidence, suggesting a link between concept appraisal and metacognition.

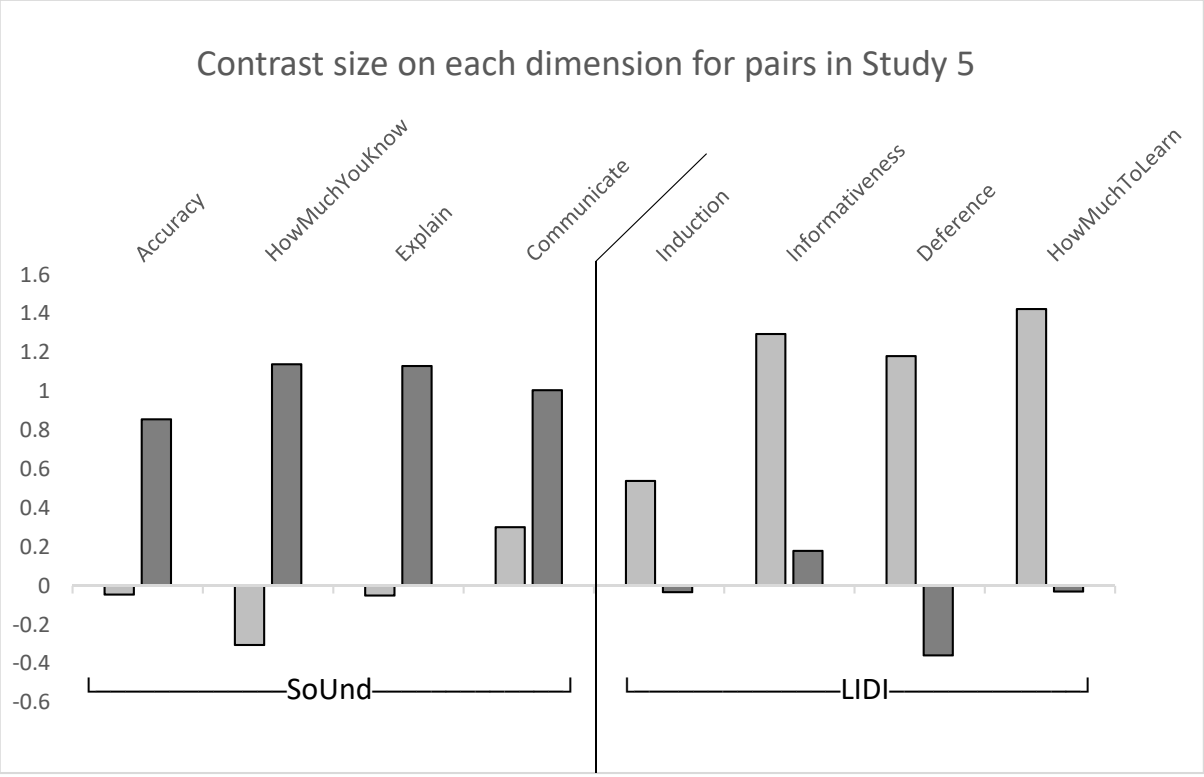
2.5.1 Method

2.5.1.1 Participants. Participants were 121 adults (88 Female, 31 Male, 2 unspecified) recruited as before (Age 18-80; M_{Age} = 35.18). Of these, 105 were native speakers and the remainder were fluent enough in English to take part.

2.5.1.2 Materials, Design and Procedure.

We used the same list of concepts as in Study 4. As described above, the materials were designed to consist of pairs of concepts that differed strongly in the appraisal ratings given by the participants in Study 1. Without trying to de-confound all eight dimensions, we selected pairs that either contrasted maximally on the four SoUnd dimensions (CAKE), while being broadly matched on the four LIDI dimensions, or the converse. For each contrast, there were eight pairs in each of the three conceptual domains (People, Recreations, Health). A new set of participants were randomly assigned to one of two groups, each group given pairs with only one type of contrast. Figure 9 shows the mean dimensional differences within pairs in the SoUnd contrast problems (dark bars, N = 61) and in the LIDI contrast problems (grey bars, N = 60).

Figure 9. Differences on each dimension between the high and low items in each pair for SoUnd (dark bars) and LIDI (light bars) problems. The four SoUnd dimensions are to the left, and LIDI dimensions to the right.



Order of concepts within pairs was counterbalanced, and each participant completed eight induction problems in three different domains (People, Recreations, and Health) for a total of 24 problems. Each problem stated that 80% of people in a given category (e.g., police officers) were found to have a particular characteristic (e.g., personality type X) whereas 80% of people in another category (e.g. library users) had the opposite characteristic (e.g., personality type Y). Participants chose which of the two characteristics an individual who belonged to both categories was likely to have. Participants also rated their confidence on a scale from 0 (simply guessing) to 100 (complete confidence). The following problem is an example of the question for People:

A psychologist studying personality has noticed that people belong to one of two different personality types: personality type X and personality type Y. In her study, she noticed that 80% of people who were POLICE OFFICERS *had personality type X, whilst 80% of people who were* LIBRARY USERS *had personality type Y. Sam is both a* POLICE OFFICER *and a* LIBRARY USER*. Which personality type is Sam more likely to have?*

In each domain, two different scenarios were used, with four problems in each scenario. All materials can be seen in the Supplementary Materials (Appendix E). Order of domains was randomized across participants, and order of problems within domain was constant. Upon completion participants were asked to describe what type of information they had used to make their judgements. As with Study 4, participants subsequently judged relative class sizes on a scale from 1 (more people in Category A) to 5 (more people in Category B).

Participants completed the questionnaire online through Qualtrics. Participants were told there were no right or wrong answers.

2.5.2 Results

2.5.2.1 Exclusions. Participants who consistently chose the same category (left or right on the page) for each question (N = 16) were excluded, leaving the final number of participants at 105: 53 for SoUnd problems and 52 for LIDI problems.[13] Further, 31 participants did not complete the class size questions and were excluded from just those analyses.

2.5.2.2 Induction Responses. Figure 10 shows the percentage of times each participant picked the higher-rated category (the category higher on the SoUnd dimensions in SoUnd problems and the category higher on the LIDI dimensions in LIDI problems).

[13] All analyses produced the same conclusions when all participants were included.

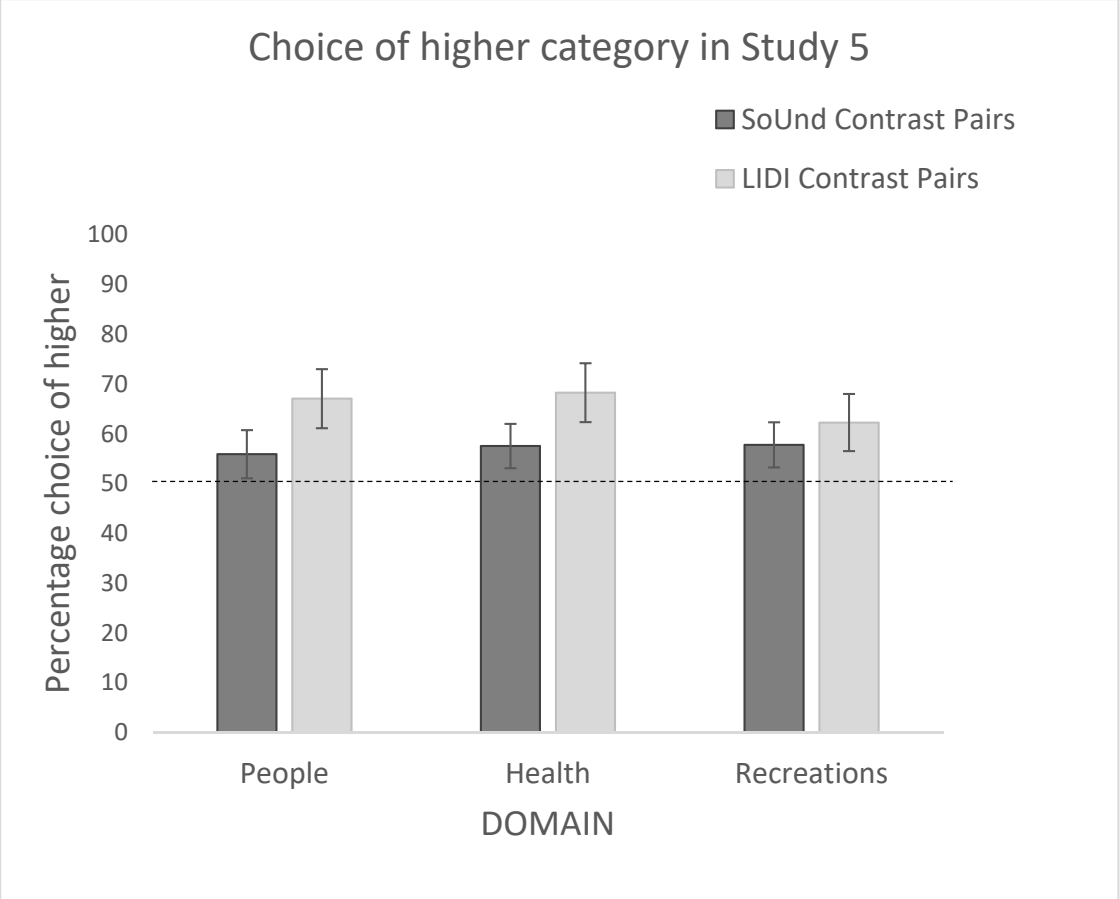


Figure 10. Percentage choice of the category higher in SoUnd or in the LIDI dimensions in Study 5. Dotted line shows chance level of 50%. Error bars show 95% CI.

The higher-rated category was chosen significantly above chance in both types of problem: SoUnd (M = 57%, SD = 11%; t (52) = 4.59, p <.001) and LIDI (M = 66%, SD = 15%; one-sample t (51) = 7.85, p <.001). A mixed ANOVA with domain within-subjects and problem type between-subjects showed that LIDI problems produced larger effects than SoUnd problems (F (1,103) = 12.02, p < .001, η_p^2 = .10).

To confirm that the effect for LIDI was not just being driven by the contrast on the *Induction* dimension, a partial correlation analysis across the 48 problems was run between frequency of selecting the higher category and the size of the contrast between members of a

pair on each dimension, holding the contrast for *Induction* constant. Significant partial correlations were found for *Informativeness* ($r(45) = .332$, $p = .023$) and *How Much To Learn* ($r(45) = .391$, $p = .007$). These results confirm similar findings in Study 4, although the p value for *Informativeness* here is only marginal to a Bonferroni corrected significance threshold of .017.

Thus, in this forced-choice setting, in addition to the effect of the dimensions *Induction, Informativeness* and *How Much To Learn* seen in Study 4, we also find that people's sense of understanding of a concept predicts whether it is likely to be seen as a reliable basis for making inductive choices.

2.5.2.3 Confidence Judgements. Participants indicated a higher level of confidence when they chose the higher category in the pairs contrasting in the LIDI dimensions (mean 43.8 versus 39.2, $t(50) = 3.60$, $p = .001$, $d = .22$), but there was no significant difference in confidence when they selected the higher category in SoUnd problems (mean 43.8 versus 44.2, $t(52) = -.45$, $p = .66$). The interaction was significant in a mixed ANOVA, $F(1,102) = 9.95$, $p < .01$, $\eta_p^2 = .09$). Thus, whilst there was evidence that the SoUnd appraisal of a concept predicts inductive choices, the collection of the other four dimensions, LIDI, had a stronger role, and was the only factor affecting confidence.

2.5.2.4 Class Size. To measure whether the category chosen as the basis for induction was influenced by class size, point biserial correlations between binary responses and prevalence ratings across the 24 problems were computed for each individual. If a participant was making a choice based in part on how many members a category has, there should be a correlation between the inductive judgement and the class size judgement.

For LIDI problems there was a small positive mean correlation between the judgements (mean = .11, $SD = .22$; $t(27) = 2.60$, $p = .02$). The chosen category tended to be

judged as having more members. For SoUnd problems, the mean correlation did not differ significantly from zero (mean = .01, *SD* = .25; *t* (52) = .27, *p* = .79). The two means did not differ on an independent t-test (*t* (27) = 1.88, *p* = .07). Thus, there was some weak evidence that participants preferred the larger category.

2.5.3 Discussion

Our final study confirmed the finding of Study 4, that concepts rated highly on the dimensions *How Much To Learn* and *Inference*, as well as *Induction*, are seen as a better basis for inductive generalisation. We also found evidence that the SoUnd factor, discovered in Studies 1 and 2, and operative in taxonomic structure in Study 3, also predicts inductive judgements when two categories are pitted against one another. The forced choice design of this study did not allow us to de-confound the relative importance of the eight dimensions (LIDI in one set of problems, CAKE in the other). Study 4 did, and found no evidence for an effect of the *Deference* dimension. Taken together, Studies 4 and 5 thus offer evidence that the following forms of concept appraisal are integrated with the inductive structure of concepts: *Induction*, *Inference, How Much To Learn* and SoUnd.

Interestingly, participants did not feel more confident when choosing the concept higher in SoUnd, whereas the contrast along the LIDI dimensions did produce a significant difference in reported confidence. Finally, class size did have some small effect on inductive choice, with the preferred category for induction tending to be judged as being the larger class (as found by Patalano et al., 2006).

3. General Discussion

Our research began with the idea that the study of concepts to date has been unnecessarily limited, focused overwhelming on the way concepts encode features or properties of objects. We wanted to investigate concepts from a new and different

perspective. In this discussion we start by summarising what we take our results to show about the way concept appraisal works (§3.1). We go on to highlight some directions for future research (§3.2).

3.1 Forms of Epistemic Appraisal of Concepts

As the first study on the topic, our aim was to establish whether or not the putative phenomenon of epistemic concept appraisal is interesting enough to merit further investigation. We had two questions: is concept appraisal psychologically real; and is it psychologically important? To answer the first question, we examined a number of ways that people might plausibly assess the epistemic merits of a concept. For each, we tested whether different people make that assessment in the same way – across a range of concepts drawn from diverse domains. If so, that would make it plausible that certain epistemic judgements about concepts – dimensions of concept appraisal – are a real feature of people's cognitive apparatus. Even judged reliably, however, these assessments might be *sui generis* and irrelevant to other cognitive functions. Our second aim was to examine their potential psychological importance by asking whether they are integrated with other central aspects of the structure of concepts.

Philosophical considerations suggested we look for forms of epistemic appraisal in two broad families: how much we know about a concept, and how much it enables us to know. Our experiments were necessarily exploratory, with many open questions rather than strong prior predictions. Nevertheless, the philosophical literature led us to expect that the factor structure might follow the division between these two broad families, with the dimensions *Accuracy, How Much Do You Know* and *Explain* clustering together, as elements of individual understanding – as indeed we found. The status of *Communicate* was an open question, since it could arguably fall within either family. In the event we found *Communicate* to be a consistent element of the sense of understanding factor. And indeed it

makes sense that the more you know about a concept, the more ready you will be to communicate with it.

We had expected *How Much To Learn, Induction* and *Informativeness* to cluster together, since they are all drawn from the second broad family – how much does the concept enable us to know? That did not turn out to be the case. People do indeed keep track of, and agree about, ways in which a category is good or bad as a cognitive tool – how well properties generalise to new instances, how many properties are shared by category members – but they do so separately, rather than running together these assessments in a more general epistemic rating, as with the sense of understanding.

We also expected that *Deference* might correlate negatively with the SoUnd cluster, since the less you know, the more you might rely on others. In the event, readiness to defer to experts was not related to how well the concept was understood. It was however reliably rated between subjects, suggesting that deference is not just a context-sensitive social phenomenon. When asked about different concepts, people agree that there are experts for some categories whereas for others everyone is entitled to their own opinion. *Deference* appeared to indicate that a term was natural-kind-like, making it behave in some respects like *Induction, Informativeness* and *How Much To Learn*, so it may be that concept-users are inclined to defer to experts about more inductively rich and kind-like categories, irrespective of whether they take themselves to understand the concept well or not. We did not, however, find evidence that *Deference* is integrated with conceptual structure (or at least, with the two aspects probed here: taxonomic structure and induction).

In Study 3 we saw that the other seven of our eight dimensions do predict the hierarchical structure of taxonomic concepts. Studies 4 and 5 turned to category-based induction. Would forms of concept appraisal predict whether a concept is seen as a dependable basis for projecting a property observed in some members of a category to a new

instance (Millikan, 2000)? We were particularly interested in that question for the family of appraisals that concern the epistemic utility of the category itself (LIDI). Philosophical considerations and previous experiments (Patalano et al., 2006; Patalano et al., 2009) suggested to us that *Informativeness* and *How Much To Learn* might predict the degree to which a concept would be relied on for category-based induction. That the *Induction* dimension itself is predictive in this respect would test the validity of that question. Furthermore, we considered that high ratings in SoUnd might make an independent contribution to inductive preference, although we made no strong prior prediction in this respect.

In the event we found that *Induction*, *Inference* and *How Much To Learn* all predict inductive preference, as does the SoUnd factor in the forced-choice task in Study 5. This result is further evidence that some forms of concept appraisal are integrated with key aspects of the wider structure of concepts: the SoUnd factor (measured by the four CAKE dimensions), the *Informativeness* dimension and the *How Much To Learn* dimension are three independent forms of stable concept appraisal. Studies 4 and 5 showed that these appraisals are integrated with the aspects of conceptual structure that underpin induction – category-based induction being one of the main things that makes a concept such a useful cognitive tool.

But what does 'integrated' mean here? A concept encodes information about category members: features used to categorise objects (small, round, hard, brown, shiny, rattles → HAZELNUT) and further properties which those objects are expected to have (e.g. a certain flavour). It encodes relationships between categories (WHALE → MAMMAL), in taxonomic hierarchies and thematic groups (e.g. items on the dinner table). Many concepts are structured so that people can reliably rate how typical an instance is; some also store individual exemplars, or carry causal knowledge of how some properties cause or enable other

properties (a mini theory). All these kinds of information are at the 'object level': they concern the object that falls under a concept, and its properties.

Concepts are such a vital cognitive tool that we hypothesized that they would also encode information of a rather different sort, information that is more 'meta level' – about the concept itself, and about its worth as a cognitive tool. That hypothesis prompted our investigation into whether epistemic appraisals form part of the structure of concepts. The hypothesis comes in two versions. In both versions information that is integrated in the structure of concepts is responsible for people's epistemic appraisals. In the strong version, concept appraisals are meta-data stored with a concept: further items of information that are encoded in the structure of a concept. As well as representing properties of objects, a concept represents how good it itself is epistemically. On this version, some representational device functions to encode epistemic utility, a bit like the way one can use a highlighter to represent which passages in a book are most important.

In the weaker version of the hypothesis, concept appraisals are not stored as meta-data but generated out of stored object-level data. On this version, concept appraisals are generated from other features of the structure of a concept rather than being encoded directly. *Induction* may be generated by checking if one has a mini theory, *Informativeness* by counting how much information one could store for objects of this general kind. It is less apparent that a unified sense of understanding could be generated in this way. But perhaps the sense of understanding is generated from episodic memories about occasions when one has used the concept: did I get my facts right, did other people understand me, how well did I explain what I meant? If so, SoUnd would not itself be a piece of information encoded with a concept, but would somehow be a stable result of operating on information that is stored. Our results to date do not decide between these two versions of the hypothesis. Hence we adopt the neutral term 'integrated'. What we take our results to have shown is that forms of concept

appraisal are integrated with key aspects of the structure of a concept (taxonomic and inductive structure), in the sense that they are either encoded by the concept directly, or generated in a stable way out of information encoded by the concept. Our experiments provide evidence that SoUnd, *Induction*, *Informativeness* and *How Much To Learn* represent forms of stable concept appraisal that are integrated with the wider structure of concepts.

Our experiments asked people what they think about their concepts. We could also ask whether their assessments are accurate (cf. metacognitive accuracy). For concept appraisals to be useful they must track, with some fidelity, the cognitive attributes they are concerned with. Given this important function, how can we assess the accuracy of the forms of concept appraisal we have uncovered? SoUnd concerns the thinker's evaluation of the quantity of accurate knowledge a concept carries for them (*Accuracy, How Much Do You Know, Explain* and *Communicate*). This could include all the information encoded with a concept such as its defining properties (if any) and prototypical and ideal features, together with implicit knowledge and sensorimotor expectations, and explicit beliefs about the class in question. Each piece of information may be accurately represented or it may be incorrect. That is hard to quantify in practice, but in principle the target of SoUnd is reasonably clear. By analogy with perceptual decision-making, we can define the sensitivity and bias of a thinker's sense of understanding (Fleming & Lau, 2014). SoUnd will show *metacognitive sensitivity* if variations in the SoUnd rating for a concept track variations in the amount (and quality) of correct information stored with it; and it will be *metacognitively unbiased* (or well-calibrated) if it allows the thinker to make an accurate estimate of how much correct information they will be able to retrieve relying on the concept. We could define in a similar way what it would be for the thinker to be metacognitively accurate in their appraisals of *Induction*, *Informativeness* and *How Much To Learn*.

3.2 Future Directions

In this section we identify some directions for future research. First, our methodology was partly inspired by existing research on natural kinds and essentialism. An obvious next step would be to ask our integration question about this aspect of the structure of concepts. Are some forms of concept appraisal integrated with the conceptual structure that takes a category to have an underlying essence (Newman & Knobe, 2019), causal structure (Murphy & Medin, 1985; Rehder & Hastie, 2004) or some other kind of category coherence between the features encoded by the concept (Patalano et al., 2006)? We can also look at coherence between concepts: some concepts encode information that coheres with lots of other things the thinker believes (Davidson, 1986; Quine, 1961). Is this reflected in *Induction*, *Inference* or *How Much To Know*? Answering these questions experimentally will give us a better appreciation of how concept appraisal integrates with the wider body of knowledge and expectations encoded in a concept.

Second, we would expect concept appraisal to be involved when people rely on concepts when talking to others. Ratings on the *Communicate* dimension show considerable inter-subjective agreement about which concepts provide a good basis for communication and which do not (Studies 1-3). We found that this dimension clusters with *Accuracy*, *How Much Do You Know* and *Explain* into a single factor that different people rate in a similar way across a diverse range of concepts. If we want to get an insight into why concepts come into use in our social groups, or go out of use, the way people select which concepts to communicate with may be the most important factor. People can signal this dimension directly in various ways, for example by putting scare quotes around a term, or using it with a rising, questioning intonation pattern. Finding that *Communicate* clusters as it does suggests that there are at least three other ways, in addition to indicating directly that it is not a good basis for communication, in which we could decrease a group's propensity to use a given

concept C: by giving feedback which shows that the information encoded by C is inaccurate, by asking questions about C which the thinker cannot answer (eroding the *How Much Do You Know* rating), or by tasking the thinker with explaining the category with the aim of dispelling the illusion of explanatory depth (Rozenblit & Keil, 2002). Experimental work is needed to investigate whether people's sense of understanding a concept, hence their propensity to communicate using it, can indeed be modulated in these ways.

These insights will be crucial if the project of conceptual engineering is to succeed (Cappelen, 2018; Machery, 2017; Thomasson, 2017). One way of addressing various social injustices is by changing the way we conceive of the relevant social categories (like WOMAN or RACE). However, to stop people using a familiar concept may take more than exhortation. We will need to find ways to act on the psychological factors that make people select some concepts, and leave others aside, when they engage in communication and reasoning. At the moment we have little idea what those factors are. They may be implicit as well as explicit. The findings we report here are a first step in the direction of gaining that insight. We found that the sense of understanding a concept, and appraisals of concepts along the *Induction, Informativeness* and *How Much Do You Know* dimensions, affect which concepts are relied on in two central conceptual tasks. That gives us an indication of how aspects of the structure of a concept may affect whether or not it is selected for various uses, communication amongst them. It thus points the way towards a potential place to intervene if we want people to stop using certain commonplace but problematic concepts.

The role concept appraisal can play in conceptual engineering points toward an even broader role. Social epistemology has highlighted how creating knowledge is a social process achieved collectively by a group of people (Goldman, 1999). Constructing concepts looks to be social in the same way. New concepts are introduced and tested in argument, reasoning and action (Dunbar, 1997). Some survive and become widely adopted, others flourish briefly

and fade, and many never get taken up at all. Concept appraisals are likely to be at work here. Concepts that are initially relied on tentatively will survive if they prove to be useful. Ratings along the dimensions of *Induction, Informativeness* and *How Much To Learn* may be part of what convinces people that a new concept is scientifically / epistemically useful, paving the way for its adoption. Correspondingly, a low sense of understanding for a concept which is taken to be important on these other dimensions may motivate an individual to inquire further into the category, learning more information and enriching their understanding of the concept. These suggestions are only speculative at this stage, but they do show that a deeper understanding of concept appraisal could tell us a lot about how humans collectively develop their conceptual schemes.

Finally, it is worth noting that the dimensions of concept appraisal we have discovered could potentially be illuminating in a very wide range of developmental research. Developmental transitions that have an obvious relevance include those in psychological essentialism (Keil & Batterman, 1984), causal complexity (Kominsky, Zamm, & Keil, 2018), categorization (Sloutsky, 2010), the illusion of explanatory depth (Mills & Keil, 2004) and conceptual change (Carey, 2009). There is not space here to explore these rich connections. In a recent review paper two of the authors have looked at the last case in more detail (Smortchkova & Shea, 2020).

4. Conclusion

Academics from all disciplines have strong views on the utility of many of the concepts and theoretical terms used in their own and other fields. We debate the validity and utility of concepts like MINDFULNESS, PERSONALITY, INTELLIGENCE and STRESS. Many of our concerns are broadly epistemic. We are asking how good the concepts are as a tool for scientific inquiry. Outside of scientific concepts, people debate the content and utility of socially

significant concepts like RACE and NATIONALISM. Again, the issues are partly epistemic.

Those epistemic assessments take place reflectively and explicitly. Until now it has been unclear whether ordinary concept users make similar assessments of everyday concepts, appraisals that may be less reflective and less obvious. This paper reports the first set of experiments to examine that question. Our results indicate that the putative phenomenon of concept appraisal is real, interesting, and merits further investigation.

References

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests *Handbook of Categorization in Cognitive Science (Second Edition)* (pp. 157-188): Elsevier.

Boyd, R. 1991. Realism, Anti-Foundationism and the Enthusiasm for Natural Kinds. *Philosophical Studies*, *61*, 127-148.

Cappelen, H. (2018). *Fixing Language.* Oxford: Oxford University Press.

Carey, S. (2009). *The Origin of Concepts.* Oxford: O.U.P.

Carr, J. R. (2017). Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research*, *95(3),* 511-534.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*(2), 291.

Dahlgren, K. (1985). The cognitive structure of social categories. *Cognitive Science, 9*(3), 379-398.

Davidson, D. (1986). A coherence theory of knowledge and truth. *Truth and interpretation* (pp. 307-319).

Davies, M. (2015). Knowledge–Explicit, implicit and tacit: Philosophical aspects. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 74-90).

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461-493). Washington DC: American Psychological Association Press.

Égré, P., & Bonnay, D. (2012). Metacognitive perspectives on unawareness and uncertainty. In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 321-342). Oxford: OUP.

Égré, P., & O Madagáin, C. (2019). Concept Utility. *Journal of Philosophy, 116(10),* 525-554.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91-114.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8,* 443.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*(5998), 1541-1543. doi:10.1126/science.1191883

Goldman, A. I. (1978). Epistemics: The Regulative Theory of Cognition. *The Journal of Philosophy*, *75(10)*, 509-523.

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford / New York: OUP.

Goldman, A. I., Beddor, B. (2015). Reliabilist Epistemology. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2016/entries/reliabilism/>.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology, 74*(4), 491-516.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39(1)*, 113-127.

Haslanger, S. (2000). Gender and race:(What) are they?(What) do we want them to be? *Nous, 34*(1), 31-55.

Hookway, C. (1994). Cognitive virtues and epistemic evaluations. *International Journal of Philosophical Studies*, 2(2), 211-227.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1441-1451.

Kalish, C. W. (2015). Normative Concepts. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts*, (pp.519-540). Cambridge, MA: MIT Press.

Keil, F. C. (1989). *Concepts, kinds, and conceptual development.* MA: MIT Press.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 221-236.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition, 127*(2), 242-257.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, 76(6), 1261-1277.

Kominsky, J. F., Zamm, A. P., & Keil, F. C. (2018). Knowing when help is needed: A developing sense of causal complexity. *Cognitive Science*, 42(2), 491-523.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge Handbook of Consciousness*, (pp. 289-325). Cambridge: Cambridge University Press.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental psychology: General, 140*, 117-139. doi:10.1037/a0022171

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*, 80-113. doi:10.1037/a0025648

Koriat, A. (2015). Metacognition: Decision making Processes in Self-monitoring and Self-regulation. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (Vol. 1, pp. 356-379). Oxford / Malden MA: John Wiley & Sons, Ltd.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. Chicago/London: University of Chicago.

López, A., Atran, S., Coley, J. D., Medin, D., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32,* 251-295.

Machery, E. (2017). *Philosophy Within Its Proper Bounds*: Oxford: OUP.

Millikan, R. G. (2000). *On Clear and Confused Ideas*. Cambridge: Cambridge University Press.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87(1),* 1-32.

Murphy, G. L. (2002). *The big book of concepts.* MA: MIT Press.

Murphy, G., & Medin, D. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92*(3), 289-316.

Nelson, L. J., & Miller, D. T. (1995). The distinctiveness effect in social categorization: You are what makes you unusual. *Psychological Science, 6*(4), 246-249.

Newman, G. E., & Knobe, J. (2019). The essence of essentialism. *Mind and Language, 34(5),* 585-605. DOI: 10.1111/mila.12226

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language, 54*(3), 407-424.

Patalano, A. L., Wengrovitz, S. M., & Sharpes, K. M. (2009). The influence of category coherence on inference about cross-classified entities. *Memory & Cognition, 37*(1), 21-28.

Proust, J. (2012). Metacognition and mindreading: one or two functions? In M. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 234-251). Oxford: OUP.

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*: Oxford University Press.

Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy*, *70(19),* 699-711.

Quine, W. V. (1961). Two Dogmas of Empiricism *From a Logical Point of View*. Cambridge, Mass: Harvard University Press.

Quine, W. V. (1970). Natural Kinds. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 1-23). Dordrecht: D. Reidel.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition, 91*(2), 113-153.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Memory and Language, 14*(6), 665.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology (vol. 1,* pp. 177–206). London: Academic Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8(3)*, 382-439.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26(5)*, 521-562.

Skorupski, J. (2010). *The Domain of Reasons:* Oxford: OUP.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science, 34(7)*, 1244-1286.

Smortchkova, J., & Shea, N. (2020). Metacognitive Development and Conceptual Change in Children. *Review of Philosophy and Psychology*. doi: 10.1007/s13164-020-00477-7

Sperber, D. (2010). The guru effect. *Review of Philosophy and Psychology*, *1(4),* 583-592.

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25(4)*, 359-393.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.

Thomasson, A. L. (2017). Metaphysics and Conceptual Negotiation. *Philosophical Issues, 27(1),* 364-382.

Treanor, N. (2013). The measure of knowledge. *Nous*, *47(3),* 577-601.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental psychology: General, 113*(2), 169.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect. *Memory & Cognition, 40,* 703-716.