Apples and oranges: improving meta-analysis through conceptual review

Annemarie van Stee[a]

[a]Radboud University
Department of Philosophy
PO Box 9103
6500 HD Nijmegen
The Netherlands

a.vanstee@ftr.ru.nl | www.avanstee.nl

Word count (incl. abstract, keywords, bibliography): 6180

Apples and oranges: improving meta-analysis through conceptual review[1]

Abstract

As in all empirical sciences, meta-analysis is a vital tool for cognitive neuroscience. Meta-analyses should include data from studies that are comparable in terms of the mental processes involved. Experimental contrasts determine whether studies are thus comparable, yet many meta-analyses, especially informatics-driven ones, select studies based on labels (e.g. 'self-reflection', 'working memory'). Unfortunately, labels are an unreliable proxy for experimental contrasts: different contrasts may receive the same label and the same contrast may receive different labels. In other words, meta-analyses in cognitive neuroscience regularly compare apples and oranges and/or fail to include all apples. I call this the comparability problem and propose conceptual review as a method to address it. A conceptual review analyzes the conceptual implications of methodological choices in a cognitive neuroscience subfield. It sheds light on the comparability of studies and the construct validity of the operationalizations used. Among other benefits, conceptual review has the potential to improve the quality of meta-analyses, which is crucial for progress in cognitive neuroscience.

Keywords: meta-analysis; tasks; experimental contrasts; labels; cognitive ontology; cognitive neuroscience; informatics; philosophy of neuroscience; self-reflection; love

[1] An earlier version of this argument was published in Dutch (van Stee, 2019).

As in all empirical sciences, meta-analysis is a crucial research tool in cognitive neuroscience (CNS). For meta-analysis to succeed, it is important that studies that are grouped together investigate the same mental process and that studies that investigate different mental processes are grouped apart. After all, it is no use comparing apples and oranges. But what if we do not realize we are comparing apples and oranges, as we have labelled both of them 'apple' or 'fruit'? Here I introduce *conceptual review* as a way to systematically analyze the implications of methodological choices for the mental concept under study.

1. The comparability problem

The number of neuroimaging studies has grown exponentially over the past decades. To be able to see the overall patterns in the results of all these studies, large scale meta-analysis is necessary. The replicability crisis has brought home the realization that single studies tell little; that meta-analysis is crucial to determine which results are robust and which are not (Maxwell et al., 2015). Some have calculated that replication success of single studies in cognitive neuroscience is bound to be even lower than in psychology (Szucs & Ioannidis, 2017), which makes meta-analysis even more important. Online databases, such as Neurosynth (neurosynth.org) or BrainMap (brainmap.org), employ informatics to enable large scale meta-analyses of neuroimaging studies. It may seem obvious that these meta-analyses should only compare studies that are in fact comparable. In practice, however, this requirement is often not met.

Databases currently annotate imaging data with *labels* for mental processes. Researchers enter these labels manually (BrainMap) or they are extracted automatically from research articles (Neurosynth). Other types of meta-analysis, e.g. in review articles, also tend to select and group studies on the basis of the mental labels researchers have applied. In such cases, labels are a proxy for the tasks used to obtain data, or more specifically, for the experimental contrast implemented through a task. Experimental contrasts determine what mental processes are involved and are thereby the crucial factor for comparability between studies. Yet labels may be applied inconsistently to contrasts, for example when two different contrasts receive the same mental label or, vice versa, when the same contrast is employed in different CNS subfields and labelled differently accordingly (Figdor, 2011; Francken & Slors, 2014; Poldrack et al., 2011; Poldrack & Yarkoni, 2016).

For example, 'working memory' is a mental label applied to neuroimaging studies employing the N-back task, but also to studies employing the Wisconsin card-sorting task, or the Sternberg task, or others. What is more, the Sternberg task allows for different experimental contrasts, such as high load versus low load, or probe match versus probe mismatch (Poldrack et al., 2011). An automated meta-analysis on 'working memory' will thus include a wide range of different experimental contrasts. Vice versa, the Wisconsin card-sorting task is not just used in 'working memory' studies, but also in 'task-switching' studies (Francken & Slors, 2014). Given these different labels, meta-analyses will not group together Wisconsin card-sorting studies, even when they are entirely comparable in terms of their experimental contrasts. In other words, we are comparing apples and oranges and failing to compare apples with apples. I will call this the *comparability problem.*

People have attempted to improve the comparability of studies through setting up cognitive ontologies. In the next section I explain what they are and argue that as long as cognitive ontologies are not built from the methods up, they are bound to fail. I then present conceptual review as a method to analyze the implications of methodological choices for the comparability of studies, and show its benefits. After that, I explain how conceptual review's results could be integrated with cognitive ontologies to facilitate automated meta-analysis in cognitive neuroscience.

2. Cognitive ontologies

A cognitive ontology is a standardized terminology, specifically a set of standardized cognitive terms. That is, in this case 'ontology' refers to a nomenclature that researchers intend to use to facilitate mutual understanding and automated aggregation of research findings (Janssen et al., 2017; Poldrack et al., 2011). This sense of 'ontology' stems from bioinformatics.

The most famous cognitive ontology is probably Cognitive Atlas (cognitiveatlas.org; Poldrack et al., 2011). Yet Cognitive Atlas suffers from several problems.[2] A fundamental theoretical problem with Cognitive Atlas is that it is built up around mental concepts and their definitions and thereby, around what researchers disagree about most.

First, Cognitive Atlas prioritizes mental concepts over tasks. In principle, it asks for definitions of both, but in practice definitions of concepts are easier to lift from research articles than definitions of tasks, as tasks are not generally defined, but simply given the practical details of. The level of practical detail one can fill out for tasks in Cognitive Atlas is limited. What is more, the relation between concepts and tasks is defined as 'measured-by', i.e. concepts are measured by tasks, putting concepts first – as opposed to, say, an 'operationalizes'-relation would have done. Yet starting from the concepts instead of from the tasks is the wrong way around if the aim is to build a standardized terminology that will be generally accepted. The facts about the methodology of a study are a better starting point as they are harder to disagree with. And crucially, what mental concept is actually investigated depends on the experimental contrast resulting from the methodological details.

Secondly, Cognitive Atlas does not contain a principled way to resolve disagreements. It offers wiki-style discussion pages and where discussion fails to resolve an issue, the option of forking a mental concept. A disambiguation page then lists several usages of the concept, each with its own concept page and 'measured-by' relations to tasks. Poldrack and colleagues also speak of the need for curation. They state that curation decisions "are expected to reflect the consensus of the community" but also that "a guiding principle is that curatorial decisions should be 'evidence-based' to ensure that questions about terminology, concepts, and relationships are not determined only by popular fiat" (Poldrack et al., 2011, p. 7).

As for the consensus option, consensus is hard to obtain where it matters most. The issue cognitive ontologies are meant to solve is the inconsistency in how mental concepts are defined and

[2] An important practical problem is the so-called 'data entry problem' that plagues many initiatives in neuroinformatics (Laird et al., 2005). Cognitive Atlas is set up as a wiki-site and thus depends on the input of many, yet researchers lack incentives to participate.

how mental concepts are linked to tasks. Entering disagreements into Cognitive Atlas through forking concepts repeats that problem, showing once more that the comparability problem exists, without solving it.[3] This problem is even worse considering that Cognitive Atlas is not the only cognitive ontology around. An overview article on cognitive ontologies reports that "[a] search from 'memory' in BioPortal returns over 150 terms from 30 ontologies" (Hastings et al., 2014, p. 1).

As for the evidence-based option, some have suggested that cognitive neuroscience itself may provide empirical evidence for the (in)appropriateness of labels (Lenartowicz et al., 2010; Price & Friston, 2005; Turner & Laird, 2012). For example, Agatha Lenartowicz and colleagues performed a classifier analysis on all data in the BrainMap database labelled 'cognitive control'. They tested whether components of cognitive control that are proposed in the literature actually have specific neural correlates associated with them. This was the case for 'response selection'. Lenartowicz and colleagues therefore argue that 'response selection' should be part of a cognitive ontology. The use of other terms could not be predicted on the basis of neural activity and Lenartowicz and colleagues suggest that these terms (e.g. 'response inhibition' and 'task/set switching') "are therefore manifest only in the minds of cognitive scientists" (Lenartowicz et al., 2010, p. 690).

However, to perform a classifier analysis on existing neural data, Lenartowicz and colleagues take the ways in which that data has been obtained and labelled at face value. Yet when a classifier cannot distinguish the neural activity correlating with a particular mental concept, it may be the ways in which neural data have been obtained and analyzed that are the problem, and not the mental concept itself. The comparability problem may be responsible as a matter of fact: if 'response selection' has been operationalized in the same way in different studies, yet 'response inhibition' has been operationalized in different ways, 'response selection' is more likely to have a specific neural

[3] Also, even where consensus is to be had, it may reflect mistaken assumptions of the research community and systematic task analysis remains necessary. For example, it has been questioned whether the N-back task, though widely used as a working memory task, is a valid operationalization of working memory at all (Jaeggi et al., 2010; Kane et al., 2007).

correlate associated with it than 'response inhibition'.[4] Again, the methodological choices made in CNS studies require critical scrutiny.

3. Conceptual review

A conceptual review is a systematic analysis of the methodological choices made in a CNS literature and their implications for the mental concept under study. It provides the building blocks for a cognitive ontology, starting with the factual details of methods about which researchers do not disagree and connecting these to mental concepts based on argument. It elucidates the comparability of different studies. The results of conceptual review thus provide a necessary condition for overcoming the comparability problem and they facilitate automated meta-analysis.

Although conceptual review is mainly intended to address the comparability problem, its relevance is broader than that. All cognitive neuroscientists know that neuroimaging results can only be interpreted in light of the experimental contrast used, and in light of the methodological details resulting in that experimental contrast. When setting up their studies or reading others' work, however, researchers mostly only check whether the employed task is a common one and whether the control condition is appropriate. Occasionally, neuroscientists working in a particular subfield notice how different tasks receive the same label and how unclear data patterns may result from that (e.g., Conway et al., 2005; Richards et al., 2013). Yet they tend to draw few consequences from this for how to set up meta-analyses, or how to communicate results to a wider audience. Conceptual review can be seen as an explication and systematization of the informal rules of reflection that many neuroscientists already apply, such that they cannot be overlooked by other neuroscientists. Plus, conceptual review analyzes methods beyond determining face validity and the appropriateness of the control condition. Finally, it explicates the consequences that should be

[4] Another problem with their argument is that Lenartowicz and colleagues conflate the informatics sense of 'ontology' with the metaphysical notion of ontology, when they claim that the lack of instrumental usefulness in a standardized nomenclature for CNS also implies that the concept does not exist (cf. Figdor, 2011).

drawn from methodological choices, for communication of results, for further research, but also, and foremost, for meta-analysis.

3.1 Setting up an inventory of methodological choices

The first step in setting up a conceptual review is to build an inventory of the methodological choices researchers have made in a particular subfield of CNS. This involves noting all methodological details that impact the eventual results, either because they are responsible for the experimental contrast or because they influence the scope of the study and the reach of its conclusions. These will include characteristics of the participants; stimuli; instructions; response; control condition; and possibly pre- or post-scan questionnaires, where these are used in the analysis.

For example, CNS research into love was reviewed in this way (van Stee, 2017, ch. 1).[5] It resulted in a large table that showed that participants in studies into 'romantic love' had to be in love, often even 'truly, madly, deeply' so, per the text on the participant recruitment flyers. In many cases, interviews and/or the Passionate Love Scale were administered prior to scanning to ensure that this was truly the case. Participants to studies on 'maternal love' had to be mothers and were not subjected to checks beforehand. Participants to a study on 'unconditional love' were recruited among assistants in a residential community for people with intellectual disabilities. The director of that community was asked to select those participants with a very high capacity for unconditional love.

The stimuli, instructions and control condition in the love studies were rather comparable: participants looked at pictures of their beloveds and were told to relax. They did not have to provide a response. The control condition consisted in looking at pictures of others. Depending on the study, these others were acquaintances or were unknown to the participants. The study into 'unconditional

[5] The examples I use here stem from earlier research into CNS of love and CNS of self-reflection. These are generally considered part of social and affective neuroscience, not cognitive neuroscience narrowly defined. We already saw that the comparability problem is at play there too however, and in fact, the problem may be more pronounced there: cognitive processes like memory are also investigated in other animals and thus an even larger range of tasks are labelled 'memory'. This means that a method for task analysis, like conceptual review, is urgently needed there too.

love' used a somewhat different task: participants looked at pictures of people with intellectual disabilities that they did not know. They were instructed to self-generate a feeling of unconditional love. The control condition consisted in simply looking at those same pictures, without self-generating love for the depicted people.

3.2 Unearthing built-in conceptual implications

Methodological choices have implications for the mental concept that is being investigated. Theoretical assumptions about the concept are built in. To be clear, neuroscientists need not explicitly hold these assumptions about the concept at play; researchers often choose the details of their tasks and participant selection procedures pragmatically. Yet these choices have conceptual implications, which in turn influence the comparability of studies.

Clearly, the studies in CNS of love pertain to different types of beloveds: romantic partners, children and unknown people with intellectual disabilities, respectively. Most readers of the 'romantic love' studies will also notice that it is in fact research into the peak experience of infatuation that we associate with the beginnings of relationships, not with the romantic love that can accompany a marriage of twenty years. Participant selection practices also show different assumptions at play: unconditional love is conceived of as hard, requiring expertise, whereas maternal love is assumed to be easy and inescapable for mothers.

Task characteristics have subtle implications: love is operationalized as an experience in these studies. Simply looking at pictures of the beloved means that studies aim to tap into feelings of love rather than love behavior or love attitudes. That participants are instructed to relax also implies that loving feelings are assumed to arise automatically when one sees a picture of the beloved.

The study on 'unconditional love' has a diametrically opposed built-in implication, however. That study assumes unconditional love needs to be actively self-generated. Its control condition consists in passively looking at a picture. This means that in this study it is assumed that love will not arise automatically when seeing a picture of a lovable person, whereas the others studies assume that it will. A meta-analysis exists that includes all these studies, as they all employ the label 'love'

(Ortigue et al., 2010). That meta-analysis notes that different types of love are involved in these studies, i.e. romantic love, maternal love and unconditional love. It does not register, however, that the studies contain diametrically opposed assumptions about the mental processes at play in love experiences. To discover this, one needs to analyze the methods beyond checking face validity and appropriateness of the control condition.

So how does one draw out the conceptual implications from methodological choices? There are several ways to facilitate this, and they run parallel to the ways in which philosophers draw out implicit assumptions underlying theoretical views, applying these to CNS' research methods instead. The two main ones are investigating the history of a method and comparing different methods to each other.

First, one can investigate how a task was developed and what motivated researchers to develop the task in the way that they did. For example, in CNS of self-reflection, the task that is used most often requires participants to judge quickly whether personality adjectives ('stubborn', 'friendly') apply to them or not. This task was originally developed in memory research, to demonstrate the so-called 'self-reference effect'. This explains why the first self-reflection studies still involved a memory test at the end, and also, why they often used labels like 'self-referential processing' or 'self-referential reflective activity' rather than 'self-reflection' (e.g. Kelley et al., 2002). The history of the task also explains why participants regularly only have 2 seconds to reflect on whether they are stubborn, whereas in ordinary language, the term 'self-reflection' is not often used to refer to such quick judgments.

The adjectives themselves have an entire history too. Current studies refer back to early studies in CNS of self-reflection (mainly Craik et al., 1999), which refer to a study by Norman Anderson (1968) as the source of their adjectives. Anderson in turn selected adjectives from lists compiled by Gordon Allport and Henry Odbert (1936). Allport and Odbert went through the entire Webster's Unabridged New English Dictionary of 1925 and jotted down 17953 terms that could possibly describe personality. Anderson selected a subset using several criteria, one of which was to exclude adjectives that denote physical characteristics. ‘Strong-minded’ thus made it onto his list, whereas ‘strong’ did not. That is, built into this stimulus set is the conceptual assumption that personality is of the mind and not of the body. Anderson obtained likeableness ratings for all selected adjectives and these are still used by researchers today to ensure that the valence of adjectives in the experimental and control conditions is counterbalanced. The likeableness ratings were obtained from 1960s university students, however, and it is questionable whether so-called blue collar workers would be as positive about ‘intelligent’, ‘open-minded’, ‘earnest’, ‘interesting’, ‘broad-minded’, ‘well-spoken’, ‘educated’, ‘clever’, ‘quick-witted’ and ‘brilliant’, which are all in the top 50 of most likeable personality descriptors.

A second way to gain a clearer view of the implications of methodological choices is by comparing different tasks to each other. We saw an example of this above, in the comparison of the commonly used passive love task to the active task of self-inducing unconditional love. It can also be done for CNS of self-reflection, for a few studies employ a different task: participants are instructed to reflect freely on themselves, for two minutes or so. They are not presented with stimuli and do not have to provide responses, but are interviewed after the experiment about the contents of their self-reflection. Participants then report reflections such as “I am extrovert and talkative, while on the other hand I have sides only my friends know” (Kjaer et al., 2002, p. 1081). This stands in clear contrast to the first type of task in which participants are required to decide within two seconds whether they are extrovert, yes or no. Comparing the two tasks draws attention to the time that participants get and the content their reflections can possibly have. Implications of the methodological choices are that whereas the second type of task by and large corresponds to ‘self-reflection’ as the term is used in ordinary language, the first task involves something like accessing a rough and ready self-schema. It is quite likely that different mental processes are involved.

3.3 Drawing consequences

In order to overcome the comparability problem and facilitate meta-analysis in CNS, the main question that needs addressing is: to what extent are labels consistently associated with experimental contrasts? In CNS of self-reflection, roughly two types of tasks are employed. Yet at least ten different labels are used to refer to these two types of tasks. Besides 'self-reflection', these are 'self-referential reflective activity', 'self-referential processing', 'self processing', 'thinking about selves', 'self-knowledge', 'reflective self-awareness', 'metacognitive evaluation of the self', 'judgments of self', and also simply 'the self'. The four studies employing the free self-reflection task receive three different labels: 'self-reflection', 'reflective self-awareness' and 'self-referential reflective activity'. The first of these, 'self-reflection', is also often used to refer to studies employing the task of quickly accessing a rough and ready self-schema (van Stee, 2017, ch. 2). That is, different labels refer to the same task, and the same label refers to different tasks. The comparability problem is impeding meta-analysis here, particularly automated versions of meta-analysis. It would be better if the two types of tasks consistently receive two different labels, e.g. 'self-reflection' on the one hand, and 'accessing one's personality schema' on the other hand.

The same holds for CNS of love. Different types of love are grouped together during meta-analysis as we already saw, even though in one case, different assumptions are made regarding the mental processes responsible for love experiences. Vice versa, studies can be found that employ the same methods as in maternal love studies, but these studies do not use the label 'love', but instead speak of 'positive mood in mothers viewing pictures of their newborn infants' (Nitschke et al., 2004) or simply of 'mothers' neural activation in response to pictures of their children' (Leibenluft et al., 2004). Even though they employ the same experimental contrasts and thereby study the same mental processes, they are not grouped together or even included in meta-analyses with the studies on maternal love. Again, the comparability problem shows up and meta-analysis is impeded.

Clearly, studies that employ the same experimental contrast, resulting from the same task and participant group, should be labelled the same. In this case, given that the control condition consists in looking at pictures of unknown children, it seems a stretch to claim we are studying neural correlates of love experiences instead of recognition processes or the like. It seems prudent to stick to the more factual 'mothers' neural activation in response to pictures of their children' for now. If researchers want to study maternal love, they need to control for familiarity. Also, they need to ensure participants actually experience love during scanning. One study included a post-scan questionnaire and found a participant who reported experiencing anger instead, as she could not help but mull over the row she had with her husband right before coming to the lab (Bartels & Zeki, 2004). This meant her data had to be excluded from analysis. It would be good if all studies on love would include such a check.

Besides facilitating meta-analysis, especially automated versions of meta-analysis, conceptual review is useful in other ways too. Unhelpful methodological choices will come to light. An unsuitable control condition in CNS of self-reflection for example, or the fact that CNS of love up until 2015 involved a total of 243 female participants versus 84 male participants (van Stee, 2017).

Furthermore, conceptual review provides a clear view on the scope of the conclusions we can draw on the basis of neuroimaging studies. For example, several studies concern the neural correlates of 'the experience of infatuation' more than those of 'romantic love' generally, let alone those of 'love'. It is vital that results are communicated bearing the scope and its limitations in mind, not just in research articles, but also to a wider audience of interested laypeople, and especially to those who would like to apply neuroscientific knowledge, e.g. in policy making.

Finally, conceptual review clearly shows the lacunae in research and thereby avenues for further research. CNS of self-reflection would benefit from studies investigating other aspects to selfhood than personality, for example. Also, given the hypotheses researchers are interested in when they investigate the neural activity in mothers watching pictures of their children, studies involving fathers are clearly needed.

4. Integrating conceptual review with cognitive ontologies

So far, I have argued that meta-analyses in cognitive neuroscience are often exposed to the comparability problem. Studies are grouped together on the basis of mental labels yet these are applied in inconsistent ways to experimental contrasts: studies with the same label use different experimental contrasts and studies using the same contrast are labelled differently. Cognitive ontologies such as Cognitive Atlas aim to overcome this comparability problem. However, Cognitive Atlas is built up around definitions of mental concepts, which is exactly what researchers disagree about most, and it does not incorporate a way to overcome disagreements. I argue we should start by looking at facts that researchers are unlikely to disagree about: the methodological choices made in a study. Analysis of these choices, such as happens in conceptual review, then explicates their conceptual implications. In turn, one can base arguments about comparability on these analyses. Let us look at each of these steps and how they can be integrated with informatics solutions such as cognitive ontologies.

To combine conceptual review with a cognitive ontology such as Cognitive Atlas would require building an ontology from the bottom up, starting with the methodological details that determine the experimental contrast and delineate the scope of the study. That 'task ontology', if you will, would contain the information gathered in steps I and II of conceptual review. An inventory containing all methodological details of different studies would be the first step. Employing informatics would facilitate this process and make automatic selection of studies on the basis of their methodological details possible.[6]

Step II can be incorporated too. Unearthing the conceptual implications of the methodological choices results in a description of the mental concept under study that tends to be more precise than currently, as in the earlier example of 'the experience of infatuation', rather than 'romantic love' or even 'love' generally. Experimental contrasts can be linked to these updated mental labels through an 'operationalizes' relation. All in all, this results in an ontology that starts with the methodological details and derives mental labels from there.

[6] The Cognitive Paradigm Ontology (CogPO for short (Turner & Laird, 2012), cogpo.org) aims to do this and originally, Cognitive Atlas even aimed to link to it. Given a few additions with respect to participant details, the informatics infrastructure for step I of conceptual review therefore exists. Unfortunately, CogPO appears to suffer from the data entry problem even more than Cognitive Atlas does. I address ways to overcome the data entry problem in the next footnote.

Having this ontology would in itself facilitate appropriate communication of results. It would also deliver the information on the basis of which we can build analyses of which studies are comparable and should be grouped together during meta-analysis, and which should be held apart. Drawing consequences like that would be part of step III of conceptual review. There is no way to automate this step; ultimately, it is based on argument. The criterion for comparability is clear, as it follows directly from CNS' aim to investigate the neural underpinnings of mental processes: if a methodological difference results in the involvement of different mental processes, studies should not be grouped together.

In general, current groupings of studies will often be too coarse. In many cases, it will be relatively straightforward to decide on the basis of a task ontology which studies are comparable and which are not, or, to return to the original metaphor, which studies have investigated apples and which oranges. My proposal to divide the self-reflection studies into two groups is an example. Step III of conceptual review thus results in a systematic basis for linking tasks and mental labels in a cognitive ontology and in imaging databases.

Tests of convergent validity may play a role in argument building too, as they test whether two different tasks supposedly measuring the same mental construct actually lead to converging results. For example, in working memory research, results of the N-back task correlate only weakly with another measure of working memory, a verbal working memory span task and this is one of the reasons some question the validity of the N-back task as a task involving working memory (Kane et al., 2007). This is an example of a reflective move performed by neuroscientists that is crucial for meta-analysis and thereby for progress in cognitive neuroscience. Yet such reflection on tasks is not done often enough, and these results and their consequences are too easily ignored. What is more, convergent validity does not yet ensure that either task is a valid operationalization of the concept of interest, as in, that it truly elicits the mental process it is meant to elicit. Relatedly, it does not in itself point to what mental label is appropriate for either task. Conceptual review may contribute crucial information on those points.

In general, building up a cognitive ontology on the basis of conceptual review ensures that labels for tasks have a more principled basis than they currently have, stemming from the methodological details. It ensures that studies that receive the same label are more likely to be comparable. Also, it ensures that enough factual details about the methodology remain present in the ontology for future researchers to build arguments on if they happen to disagree with the way an experimental contrast has been labelled. All of this is improvement over current practice in Cognitive Atlas.[7]

5. Conclusion

We are currently amassing enormous amounts of neuroimaging data. To gain reliable, replicable knowledge from that pile of information, meta-analysis is crucial. Automated versions of meta-analysis are an important tool, but they currently rely on mental labels that are applied inconsistently to experimental contrasts. This impedes meta-analysis and thereby progress in cognitive neuroscience. We need to pay more attention to the methodological details of studies to gain clarity regarding the comparability of studies. Conceptual review is a systematic way to analyze the implications of methodological choices for the mental concept under study. Its results facilitate precise communication of results and point out avenues for further research. Most importantly, its results elucidate the comparability of different studies. Conceptual review can stand on its own, yet combining it with a cognitive ontology may be beneficial: cognitive ontologies provide an informatics infrastructure that facilitates conceptual review and the linking of its results to databases such as BrainMap or Neurosynth. In turn, if cognitive ontologies were built up on the basis of the information

[7] To make a cognitive ontology based on conceptual review successful, the data-entry problem needs to be addressed as well. For that, we can draw on the experience of BrainMap's initiators. They eventually turned to graduate students to enter database submissions, for course credit in a course on meta-analysis, or also as a paid research assistant (Laird et al., 2005). Doing the same for conceptual review would not only be a pragmatic solution, but also offer graduate students the opportunity to become fully aware of different methodological strategies and their impact, before they embark on their own research projects. Other possible incentives could be to include conceptual review as a type of publishable research article; and eventually to make annotation with a label drawn from the ontology a prerequisite for publication of empirical research and inclusion of data in databases like BrainMap or Neurosynth.

conceptual review provides, they stand a much larger chance of succeeding at providing a standardized nomenclature for CNS than they do currently. In sum, conceptual review addresses the comparability problem. It thereby has the potential to improve the quality of meta-analyses, which is crucial for progress in cognitive neuroscience.

Bibliography

Allport, G. W., & Odbert, H. S. (1936). *Trait-names: A psycho-lexical study* (Vol. 47). Psychological Review Company.

Anderson, N. H. (1968). Likableness Ratings of 555 Personality-Trait Words. *Journal of Personality and Social Psychology*, *9*(3), 272–279. https://doi.org/10.1037/H0025907

Bartels, A., & Zeki, S. (2004). The neural correlates of maternal and romantic love. *Neuroimage*, *21*(3), 1155–1166. https://doi.org/10.1016/j.neuroimage.2003.11.003

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Craik, F. I. M., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., & Kapur, S. (1999). In Search of the Self: A Positron Emission Tomography Study. *Psychological Science*, *10*(1), 26–34. https://doi.org/10.1111/1467-9280.00102

Figdor, C. (2011). Semantics and Metaphysics in Informatics: Toward an Ontology of Tasks. *Topics in Cognitive Science*, *3*(2), 222–226. https://doi.org/10.1111/j.1756-8765.2011.01133.x

Francken, J. C., & Slors, M. (2014). From commonsense to science, and back: The use of cognitive concepts in neuroscience. *Consciousness and Cognition*, *29*, 248–258. https://doi.org/10.1016/j.concog.2014.08.019

Hastings, J., Frishkoff, G. A., Smith, B., Jensen, M., Poldrack, R. A., Lomax, J., Bandrowski, A., Imam, F., Turner, J. A., & Martone, M. E. (2014). Interdisciplinary perspectives on the development, integration, and application of cognitive ontologies. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00062

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412. https://doi.org/10.1080/09658211003702171

Janssen, A., Klein, C., & Slors, M. (2017). What is a cognitive ontology, anyway? *Philosophical Explorations*, *20*(2), 123–128. https://doi.org/10.1080/13869795.2017.1312496

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*(5), 785–794. https://doi.org/10.1162/08989290260138672

Kjaer, T. W., Nowak, M., & Lou, H. C. (2002). Reflective Self-Awareness and Conscious States: PET Evidence for a Common Midline Parietofrontal Core. *NeuroImage*, *17*(2), 1080–1086. https://doi.org/10.1006/nimg.2002.1230

Laird, A. R., Lancaster, J. L., & Fox, P. T. (2005). BrainMap: The social evolution of a human brain mapping database. *Neuroinformatics*, *3*(1), 65–78. https://doi.org/10.1385/ni:3:1:065

Leibenluft, E., Gobbini, M. I., Harrison, T., & Haxby, J. V. (2004). Mothers' neural activation in response to pictures of their children and other children. *Biological Psychiatry*, *56*(4), 225–232. https://doi.org/10.1016/j.biopsych.2004.05.017

Lenartowicz, A., Kalar, D. J., Congdon, E., & Poldrack, R. A. (2010). Towards an Ontology of Cognitive Control. *Topics in Cognitive Science*, *2*(4), 678–692. https://doi.org/10.1111/j.1756-8765.2010.01100.x

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Nitschke, J. B., Nelson, E. E., Rusch, B. D., Fox, A. S., Oakes, T. R., & Davidson, R. J. (2004). Orbitofrontal cortex tracks positive mood in mothers viewing pictures of their newborn infants. *NeuroImage*, *21*(2), 583–592. https://doi.org/10.1016/j.neuroimage.2003.10.005

Ortigue, S., Bianchi-Demicheli, F., Patel, N., Frum, C., & Lewis, J. W. (2010). Neuroimaging of Love: FMRI Meta-Analysis Evidence toward New Perspectives in Sexual Medicine. *Journal of Sexual Medicine*, *7*(11), 3541–3552. https://doi.org/10.1111/j.1743-6109.2010.01999.x

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D. S., Sabb, F. W., & Bilder, R. M. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, *5*, 17. https://doi.org/10.3389/fninf.2011.00017

Poldrack, R. A., & Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annual Review of Psychology*, *67*(1), 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, *22*(3–4), 262–275. https://doi.org/10.1080/02643290442000095

Richards, J. M., Plate, R. C., & Ernst, M. (2013). A systematic review of fMRI reward paradigms used in studies of adolescents vs. adults: The impact of task design and implications for understanding neurodevelopment. *Neuroscience & Biobehavioral Reviews*, *37*(5), 976–991. https://doi.org/10.1016/j.neubiorev.2013.03.004

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3). https://doi.org/10.1371/journal.pbio.2000797

Turner, J. A., & Laird, A. R. (2012). The Cognitive Paradigm Ontology: Design and Application. *Neuroinformatics*, *10*(1), 57–66. https://doi.org/10.1007/s12021-011-9126-x

van Stee, A. (2017). *Understanding Existential Self-Understanding. Philosophy Meets Cognitive Neuroscience*. Leiden University.

van Stee, A. (2019). Appels, peren en fruit: Conceptual review als taakanalysemethode. *Algemeen Nederlands Tijdschrift Voor Wijsbegeerte*, *111*(3), 433–452. https://doi.org/10.5117/ANTW2019.3.008.VANS