Studying the existence and attributes of consensus on psychological concepts

by a cognitive psychometric model

Zita Oravecz
University of California, Irvine

Katherine Faust
University of California, Irvine

Daniel A. Levitis
University of Southern Denmark

William H. Batchelder
University of California, Irvine
Abstract

Psychological research can take a variety of directions while building on theoretical concepts that are commonly shared among the population of researchers. We investigate the question of how agreement or consensus on basic scientific concepts can be measured. Our approach to the problem is based on a state-of-the-art cognitive psychometric technique, implemented in the theoretical framework of Cultural Consensus Theory (CCT). With this approach, consensus-based answers for questions exploring shared knowledge can be derived while basic factors of the human decision making process are accounted for. An example of the approach is provided by examining the definition of behavior, based on responses from researchers and students. We conclude that the consensus definition of behavior is: “Behavior is a response by the whole individual to external and/or internal stimulus, influenced by the internal processes of the individual, and is typically not a developmental change”. The general goal of the paper is to demonstrate the utility of the CCT based approach as a method for investigating what current, working definitions of scientific concepts are.

Keywords: theoretical concept, behavior, Cultural Consensus Theory, cognitive psychometric model
Introduction

Researchers all over the world conduct investigations locally while sharing their findings and knowledge on the global scale. In the digital era, isolated research labs are rare and the flow of information is continuous. Scientific progress is incremental: it is based on the day-to-day work of a community of scientists who construct scientific knowledge. Ideally researchers interested in the same phenomenon should conduct their research using shared basic concepts, notions or definitions. But are these basic concepts really widely shared? Have some of them become obsolete? How can we study the consensus among researchers? One approach to the question is taken by sociologists who have been investigating the social construction and sociological explanation of scientific ideas (Berger & Luckmann, 1966; Kuhn, 1996). They describe a substantial amount of knowledge as an outcome of social processes. As a consequence, theories, ideas, and knowledge become more or less shared among the members of groups, cultures, or society at large.

We aim to study the extent of consensus and its influencing factors as related to definitions of theoretical concepts in psychology. Psychology generally deals with abstract concepts or constructs (i.e., attitude, intelligence, etc.) that need to be measured. While theoretical concepts in general are central in psychological investigations, there are complex ontological and epistemological questions with respect to them (e.g., Slaney & Racine, 2013).

Our focus is on introducing a cognitive psychometric method to study the existence and attributes of the current state of consensus on theoretical concepts among researchers. Within
this framework, we define consensus as general agreement on a certain content domain, or more specifically overlap in knowledge and opinion on a theoretical concept, shared by all members of a culture or group. The focus of our work is not the process of how consensus is achieved, but rather the current state of the consensus (if any).  

In addition to developing conceptual definitions, researchers also seek to operationalize concepts for empirical work. Therefore a definition can have several definiens or essential attributes, and the goal is to find a certain number of them that together make a definition simultaneously essential, operational, succinct and usable across contexts. The ultimate aim is to have a current summary of scientific expertise with respect to a concept in question, while reducing vagueness and ambiguity.

We propose the Cultural Consensus Theory (CCT, Batchelder, 2007; Romney & Batchelder, 1999; Romney, Weller, & Batchelder, 1986) framework as a cognitive psychometric approach to information aggregation for studying agreement in psychological concepts. In a typical CCT study members of a group respond to a set of questions tapping into some common knowledge pool. By using CCT models, answers that represent the consensus in the underlying shared knowledge can be extracted. When inferring the latent consensus from the raw item-by-person data, cognitive characteristics of the respondents are incorporated and these characteristics also represent informative outputs about the respondents themselves. Importantly, the respondents’ level of consensus knowledge is inferred from the correlation among the responses of all respondents, and the consensus answers are derived based on weighting the knowledge levels of individual respondents. More knowledgeable people
contribute more to the final consensus. Moreover, response style characteristics, such as willingness to guess if faced with uncertainty and guessing bias, are also accounted for. CCT based aggregation has proven to be superior to simple aggregation techniques, such as the majority rule (e.g., see Batchelder & Anders, 2012). CCT models have been used to study agreement within cultural entities in various settings. Weller (2007) summarizes many studies in cultural anthropology using CCT, examples include folk beliefs of the causes and treatments of various type of diseases (Weller, 1984b; Hruschka, Kalim, Edmonds, & Sibley, 2008; Baer et al., 2003; Weller, 1984a; Weller et al., 1999). In the field of psychology, CCT applications have concerned judgment of personality traits in social networks (Agrawal & Batchelder, 2012; Batchelder, Kumbasar, & Boyd, 1997), exploration of national consciousness (Yoshino, 1989) and extracting truth from eyewitness testimonies (Waubert de Puiseau, Åßfalg, Erdfelder, & Bernstein, 2012). Other applications examined properties and uses of plants (Hopkins, 2011) and ecological relations among species (Atran et al., 1999; Medin et al., 2006).

Merits of the CCT modeling approach are demonstrated here by analyzing judgments on the concept of behavior. Behavior provides a good example of a concept that is of central interest to psychology, though its definition is not straightforward. Generally speaking early behaviorists regarded behavior as anything that an organism can do that is measurable by some device (e.g., Watson, 1913). They were generally concerned with establishing psychology as a science rather than as speculative philosophy. The goal of behaviorism was to make psychology the scientific study of human and animal behavior. Today we can look at behavior in a more complex way that includes endocrine and neural elements as well.
A straightforward approach to defining the current state of the concept *behavior* is to study its attributes in the published literature, while determining the extent of consensus on how many attributes it has and exactly which ones. Levitis, Lidicker, and Freund (2009) followed this approach focusing on the field of behavioral biology (i.e., ethology). They found that researchers from societies publishing behavioral biology journals do not always agree on which features of behavior are fundamental. Also, while Levitis et al. (2009) used encyclopedias and other resources to derive fundamental features of behavior (items of their survey), one of the most striking aspects of their findings was that the scientific community rejected a large majority of these definitions.

In what follows we address the topic of studying consensus in scientific concepts in general and we augment it by contrasting expert and novice knowledge. The problem posed in Levitis et al. (2009) is used in this paper to demonstrate important aspects of the proposed consensus modeling. First, we examine how CCT in general can be used for extracting consensus definitions for basic concepts within the communities that are using the concept. Second, we provide an example of how cognitive model-based data aggregation, rather than simple data aggregation, can lead to rich conclusions by analyzing data from Levitis et al. (2009) and newly collected data from psychology students. Comparing the results in terms of consensus model parameters highlights the differences in expert and novice conceptualizations of behavior. On a more general level this work shows how state-of-the-art modeling techniques can contribute to studying prevailing consensus on definitions of psychological concepts.
Cultural Consensus Theory

Extracting consensus based definitions of scientific concepts can be achieved using Cultural Consensus Theory modeling. In cultural consensus studies a set of respondents answer various questions exploring shared knowledge on a certain topic. The cultural consensus approach is rooted in a formal model called the two high-threshold model from signal detection theory (e.g., Macmillan & Creelman, 2005) and belongs to the larger class of multinomial processing tree models (Batchelder & Riefer, 1999). In short, a two high-threshold model is a formal recognition measurement model that describes the probability that a previously learned item exceeds a memory threshold.

While different CCT approaches exist, we focus on the Extended Condorcet Model (ECM, Oravecz, Faust, & Batchelder, in press), which is an extended version of the General Condorcet Model (Batchelder & Romney, 1988). The name of the General Condorcet Model (GCM) recognizes Marquis de Condorcet (1743-1794), a French philosopher, mathematician and political scientist. He considered the problem of a jury where each member votes guilty or innocent, with all members having an independent equal probability of making the ‘correct’ decision. His main goal was to derive the probability of a correct decision as defined by the probability the majority would be on the side of the ‘correct’ decision. Similar to CCT applications, the ‘correct’ decision is not pre-established, instead it resides in the consensus judgments derived from the jurists who have access to the court proceedings. The GCM generalizes the setting to allow people to have different probabilities of making a ‘correct’ decision, and uses responses of each person to a series of binary choices to sharpen the estimation.
Consider a simple questionnaire created to collect information on the consensus on a certain scientific concept, with a group of researchers answering the questionnaire items pertaining to possible features of the concept. For example, one item in the questionnaire used by Levitis et al. (2009) states that “All behaviors are directly observable, recordable and measurable”. Figure 1 summarizes how the proposed model accounts for the decision process of a respondent for this item. The underlying assumption can be described as follows: The consensus answer on an item “All behaviors are directly observable, recordable and measurable” can be either ‘True’ or ‘False’. This consensus answer is estimated from the data. Sometimes respondents feel unsure about whether or not they know the answer, therefore not only ‘True’ and ‘False’, but also ‘Don’t know’ answers are allowed. In Figure 1 the possible responses appear in rectangles and the decision processes leading to them can be described by going down different branches of the tree.

*Insert Figure 1 about here*

In the ECM, each person can be described by their level of knowledge (as related to the consensus answer), their willingness to guess when they are uncertain about an answer, and their guessing bias (whether they would rather guess ‘True’ or ‘False’). As can be seen in Figure 1, there are two ways to arrive at the (‘correct’) consensus answer: either by knowing the answer (thick branches) or by guessing it correctly (thin branches leading to the ‘correct’ answer). One advantage of using a cognitive model for the decision process is that we can
In a questionnaire with many items it would be unrealistic to assume that all items are equally difficult to answer. In the ECM, the probability of knowing the consensus answer for an item is in fact decomposed into two parameters: the respondent’s ability and the item’s difficulty level, as is commonly done in psychometric test theory (De Boeck & Wilson, 2004). If the item difficulty level is higher than the person’s ability level, most likely he/she will not know the correct answer (thin branches on the tree). Items with difficulty levels lower than the person’s ability will most likely result in a correct answer (following the thick branches of the tree).

For the ECM it is assumed that the members of a group converge toward the same set of consensus answers, called the consensus answer key. This assumption (referred to as the single culture assumption) is tested using outputs from the model. However, the strength of this consensus or uncertainty around it can differ among items. The Bayesian modeling framework (Gelman, Carlin, Stern, & Rubin, 2004) that we apply for parameter estimation can quantify uncertainty in the consensus-based answer estimates, as parameters in the Bayesian framework are not fixed quantities (in contrast to their frequentist counterparts) but instead have probability distributions.

When there is strong agreement on an item, the consensus-based answer estimate has high certainty. However, it is possible that for certain item(s) the consensus estimate turns out to be very uncertain (‘noisy’) because of a lack of agreement among respondents. If there are several consensus-based answers with high uncertainty, and the single culture assumption of the model
does not seem to be supported (see posterior predictive model check later) one should consider an alternative model, which can account for multiple underlying cultures, as described in Anders and Batchelder (2012).

Method

Model Specifications

Consider a simple questionnaire developed to collect information on the consensus on a certain topic, with $N$ respondents answering $M$ items. The answer from a single respondent $i$ ($i = 1, \ldots, N$), for item $k$ ($k = 1, \ldots, M$) is denoted as $Y_{ik}$. On the data level, ‘True’, ‘False’ and ‘Don’t know’ (DK) answers are allowed. Figure 1 summarizes the decision tree of a single respondent $i$ for item $k$ based on the proposed cognitive model for the ECM. We introduce the following notations: The consensus answer is denoted as $Z_k$ for item $k$. The probability that respondent $i$ knows the consensus answer for item $k$ has probability $D_{ik}$. The probability that respondent $i$ ventures a guess is $b_i$. If the respondent is not willing to take a guess, it is assumed that they mark the ‘Don’t know’ option, with probability $1 - b_i$, in which case they would say ‘True’ with probability $g_i$.

The probability for each type of answer can be derived by simply going down the branches of the corresponding answers in Figure 1. Formally, the probabilities of the answer categories are written as:
\[ p(Y_{ik} = 'True' = 1) = Z_k[D_{ik} + (1 - D_{ik})b_i g_i] + (1 - Z_k)[(1 - D_{ik})b_i g_i] \] (1)

\[ p(Y_{ik} = 'False' = 2) = Z_k[(1 - D_{ik})b_i (1 - g_i)] + (1 - Z_k)[D_{ik} + (1 - D_{ik})b_i (1 - g_i)] \] (2)

\[ p(Y_{ik} = 'DK' = 3) = Z_k[(1 - D_{ik})(1 - b_i)] + (1 - Z_k)[(1 - D_{ik})(1 - b_i)]. \] (3)

Then the data \((Y_{ik})\) are assumed to come from a categorical distribution (denoted as Cat) with a 3 × 1 probability vector \(p_{ik}\) with elements corresponding to the three answer categories:

\[ Y_{ik} \sim \text{Cat}(p_{ik}), \] (4)

where the probability vector is \(p_{ik} = [p(Y_{ik} = 1) \ p(Y_{ik} = 2) \ p(Y_{ik} = 3)]\), with elements defined in Equations 1, 2 and 3.

The probability of knowing the correct answer for an item \((D_{ik})\) is a function of the respondent’s ability and the item’s difficulty level:

\[ D_{ik} = \frac{\theta_i(1 - \delta_k)}{\theta_i(1 - \delta_k) + \delta_k(1 - \theta_i)} \]

where \(\theta_i\) is the ability parameter belonging to respondent \(i\) and \(\delta_k\) denotes the item difficulty level of question \(k\), where \(\theta_i, \delta_k \in R\).

To summarize, the Extended Condorcet Model introduced above has \(3 \times N\) informant-specific parameters, namely the ability parameters \((\theta_i)\), the willingness to guess parameters \((b_i)\) and the guessing bias \((g_i)\) parameters. Also, it has \(2 \times M\) item-specific parameters: the consensus-based answer for each item \((Z_k)\) and the item-difficulty parameter for each item \((\delta_k)\).
Hierarchical extension

The ECM is formulated as a hierarchical (or multilevel) model (see Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The assumed hierarchical structure on the model parameters relaxes the interchangeability condition among them while it assumes that they share certain commonalities expressed by their population distributions. In the hierarchical CCT model information is pooled across participants, which enhances the recovery of these person level parameters. In the current application this is especially desirable as there are only a few items so the person level parameter estimates might have a lot of uncertainty. The hierarchical framework requires the selection of a population distribution for each type of parameter. The current modeling choice involves transforming the unit scale variables to the real line and assuming normal population distributions on this scale.

With respect to the covariate modeling we use a one-step analysis, which - in contrast to a two-step analysis - does not resort to first estimating person-specific parameters and then regressing them on predictors. In a one-step analysis uncertainty in the person-specific parameter estimates is directly accounted for when estimating regression weights. There are three person-specific parameters, namely ability ($\theta_i$), guessing bias ($g_i$) and willingness to guess ($b_i$), all of which can be made functions of predictors, in our case self-reported level of expertise and consistency in the judgments (as defined later). These predictors are standardized, and denoted by $x_{i,exp}$ and $x_{i,con}$ (for expertise and consistency respectively), for each respondent $i$. It is assumed that the logit-transformed person-specific parameters come from a joint multivariate normal population distribution (denoted as MVN), formally:
\[
\begin{bmatrix}
\logit(\theta_i) \\
\logit(g_i) \\
\logit(b_i)
\end{bmatrix}
\sim \text{MVN}
\left(
\begin{bmatrix}
\beta_{\theta0} + \beta_{\theta,\text{exp}} x_{i,\text{exp}} + \beta_{\theta,\text{con}} x_{i,\text{con}} \\
\beta_{g0} + \beta_{g,\text{exp}} x_{i,\text{exp}} + \beta_{g,\text{con}} x_{i,\text{con}} \\
\beta_{b0} + \beta_{b,\text{exp}} x_{i,\text{exp}} + \beta_{b,\text{con}} x_{i,\text{con}}
\end{bmatrix}, \Sigma
\right).
\]

The intercepts, \( \beta_{\theta0}, \beta_{g0}, \) and \( \beta_{b0} \) represent the baseline values across people and can be interpreted as population means when covariate scores are standardized (or omitted). The terms after the intercepts decompose person-specific variation around the intercepts into person-specific predictors (\( x_{i,\text{exp}} \) and \( x_{i,\text{con}} \)) and regression weights (\( \beta_{\theta,\text{exp}}, \beta_{\theta,\text{con}}, \beta_{g,\text{exp}}, \beta_{g,\text{con}}, \beta_{b,\text{exp}}, \) and \( \beta_{b,\text{con}} \)). This part allows for the explanation of inter-individual variation in the person-specific cognitive parameters. Residual variation is modeled through the covariance matrix \( \Sigma \), representing the unexplained variations with respect to ability, guessing bias and willingness to guess, along with possible covariation in those.

The logit transform of the item difficulties, \( \delta_k \), is also modeled by a normal distribution (denoted as \( N \)) as \( \logit(\delta_k) \sim N(0, \sigma^2) \), where the population mean is fixed to 0 to preserve model identification. The answer key, is assumed to be generated by a Bernoulli process, with probability parameter \( \pi \), formulated as \( Z_k \sim \text{Bern}(\pi) \), where values for \( \pi \) above 0.5 mean that the latent answer is more likely to be ‘True’, while values under 0.5 designate a higher chance for ‘False’.

**Bayesian data analysis**

Statistical inference for the HECM is developed in the Bayesian framework. Early
methods for statistical inference, as described for example in Weller (2007), relied partly on
factor analysis and partly on Bayesian posterior probability (see later) to derive consensus
estimates. The HECM builds upon this work and introduces a fully Bayesian treatment of CCT
models, while also extending the original model hierarchically and by incorporating a ‘Don’t
know’ response alternative.

Bayesian methods rely directly on the principles of probability theory and model
parameters in this framework can be described in terms of probability distributions, which offers
an intuitively appealing way of comprehending uncertainty about the parameters. Full
introductions to Bayesian statistical inference can be found in Gelman et al. (2004), Gill (2007)
and Kruschke (2011). Several researchers have advocated the usefulness of Bayesian methods in
psychology (e.g., Rouder & Lu, 2005; Gelman & Shalizi, 2013; Lee & Wagenmakers, 2005).
While Bayesian data analysis provides tools to deal with complex models, it is still not part of
the mainstream of statistical techniques in the social sciences. For some recent discussion of the
philosophical foundations of Bayesian statistics please consult Andrews and Baguley (2013).

Bayesian statistical inference focuses on the posterior density of the parameters. In
general, the posterior density represents the probability distribution of the parameters given the
data, namely \( p(\gamma | Y ) \propto p(Y | \gamma)p(\gamma) \), where \( \gamma \) denotes the vector of all model parameters and \( Y \)
represents the data. The posterior density is proportional to the product of the likelihood of the
data given the parameters and the prior distribution of the parameters. The latter represents our
prior knowledge about the model parameters. The more data acquired, the more influential the
likelihood function becomes on the posterior, therefore in turn dominating the prior. For the
HECM, non-informative distributions are chosen as priors (see details in Appendix A).

To avoid the computational difficulty of integrating over the multidimensional posterior, numerical integration methods are used. Specifically we implemented Markov chain Monte Carlo (MCMC) algorithms to generate samples from the full conditional distribution of a parameter or a block of parameters (see e.g., Robert & Casella, 2004). Several different types of posterior statistics can be calculated from the posterior samples, for example point estimates, posterior standard deviations (the Bayesian analogue of the standard error), and Bayesian p-values.

For carrying out parameter estimation for the HECM, the computer software JAGS (Plummer, 2011) was used, while the results were interpreted in MATLAB. A user-friendly graphical interface based software that can carry out the statistical inference for the ECM is in preparation, and an easy-to-use software package for dichotomous data (including manual and installation guidelines) is already freely available (for details, see Oravecz, Vandekerckhove, & Batchelder, in press). For brevity, we do not report extensive algorithm or model fit checks, except to note that the MCMC procedure converged rapidly, and that posterior predictive checks indicated satisfactory fit of the model to the data. For code used for the parameter estimation please e-mail the first author.
Study design about the concept of behavior

Participants and questionnaire

Findings based on the ECM are demonstrated through two data sets containing information on the definition of behavior. In addition to re-analyzing the original data from Levitis et al. (2009), new data were collected from students at a southwestern university using the original survey items. While in the original data set from Levitis et al. (2009) the respondents were behavioral biologists (members of three scientific societies producing behavioral biology journals), the new sample consisted of students who were majoring in psychology and also taking a history of psychology class. The final size was $N = 169$ with the following distribution between the societies: Animal Behavior Society (ABS; 78 professionals, 14 knowledgeable, 5 “layman”), International Society for Applied Ethology (ISAE; 41 professionals, 14 knowledgeable, 4 “layman”) and Society for Plant Neurobiology (PNB; 2 professionals, 4 knowledgeable, 7 “layman”). The psychology student sample ($N = 82$) consisted of undergraduate students who got credit for their participation. None of them marked their level as professional in behavioral psychology; however, 14 considered themselves knowledgeable. Comparing results from the two data sets involves comparing not only two different but related fields, but also an expert group already using the concept of behavior in their work and another one engaged in the process of understanding the concept. We expected to find differences in the consensus of what constitutes behavior, along with differences between the student and expert populations with respect to their level of knowledge about behavior.
As in Levitis et al. (2009), we seek a definition of behavior with the stated goal that it is not only essential (meaning what people think it should), operational (usable in deciding what is, or is not behavior) and succinct, but also widely applicable. The definition should be useable across contexts, and by researchers in several fields. In the data sets, there are 13 items containing different fundamental features of behavior, see their detailed description in Appendix B. These 13 items present concepts drawn from existing definitions of behavior and were chosen based on the criterion that they might be part of a consensus definition of behavior. This was judged based on their appearance in widely cited definitions or their use in several of the published definitions that were surveyed in the primary and secondary literature. For example, based on the definition of Tinbergen (1955), “the total movements made by the intact animal” three putative important characteristics of behavior were drawn. These were item (G) “Behavior always involves movement” item (I) “Behavior is something whole individuals do, not organs or parts that make up an individual” and item (K) “Behavior is something only animals (including people) do, but not other organisms.” This generated list of the items with the proposed features of what might or might not be considered to be behaviors. The list of potentially fundamental items was refined by testing it on biology students and revising confusing or ambiguous wording before finalizing the language of the survey. Moreover, based on these 13 proposed features of behavior, 20 examples of behavior were created (see their detailed description in Appendix C). Respondents had to decide whether these items were examples of behavior.
Overview of previous research

The analysis in Levitis et al. (2009) calculated summary statistics about whether the respondents regarded certain descriptions as fundamental features of behavior. The authors aimed to derive a consensus answer on each item the following way: if at least 64% of the respondents regarded an item as a fundamental feature of behavior or an example of behavior or if less than 24% considered the item that way, their consensus criterion was met. Their findings reveal several interesting problems. First, with their heuristic of 64% - 24% consensus was not found on 2 out of the 13 items on the proposed features of behavior. Second, significant connection was found between knowing the consensus responses and the self-reported knowledge level in behavioral biology. It was also noted that when the sample is divided into subgroups based on professional association membership, the consensus on the items tends to differ among these groups. Moreover, they found that there were internal inconsistencies when generalizing from example to definition. Although respondents endorsed a behavioral feature on an abstract level, they often gave contradictory responses on the concrete level by not agreeing that a phenomenon with the chosen feature was an example of behavior. These findings directed us towards specific issues that should be addressed within the CCT modeling framework (e.g., predictors to include in the model).

Overview of the ECM analysis

In the analysis of Levitis et al. (2009) each question was considered individually, whereas with the current approach the whole pattern of responses is explored while relying on a cognitive
model of decision making. Moreover, in the ECM approach a multilevel structure is assumed, that pools information across participants as well as items, while allowing for inter-individual and inter-item variation in terms of person-specific competency, guessing bias and willingness to guess, and item-specific item-difficulty parameters.

To further improve modeling accuracy, explanatory variables (covariates) are added, specifically the self-reported level of expertise and the number of consistent judgments. The number of consistent judgments can be calculated by counting how many times the respondents answered consistently to the features and the phenomena that exhibit those specific features. The example phenomena are associated with code numbers corresponding to the proposed features of behavior, see Appendix C. The calculation of consistent judgments is based on these codes.

Results

Proposed features of behavior

Responses to the 13 items on fundamental behavioral features were analyzed by the ECM. The outcome consists of information about the items in terms of consensus answer and difficulty level. Respondents are characterized in terms of three person-specific parameters capturing their consensus knowledge, guessing bias and willingness to guess. All three person specific parameters are regressed on the predictor variables of self-reported expertise and consistency.
Results for the items, based on their posterior probability distributions, are shown in Table 1. The first column contains keywords for each item. The following column displays the posterior median estimate for the consensus answer key, labeled as ‘True’ or ‘False’. The third column shows the posterior standard deviation (denoted as PSD), which measures the uncertainty in these estimates (it provides similar information about the estimate as the standard error in the classical statistical framework). The relatively low values of PSD for most answer key estimates in the behavioral biologist sample suggest that the model predicts ‘True’ or ‘False’ with great certainty for most items. There are two items highlighted in Table 1 (‘All behaviors are directly observable, recordable and measurable’ and ‘Behavior is something only animals (including people) do, but not other organisms’), for which the original analysis in Levitis et al. (2009) did not arrive at a definitive conclusion. In the current analysis there is relatively large uncertainty (see PSD) in the estimates of the consensus-based answers for these two items indicating that the agreement on these items is indeed not very strong among behavioral biologists. However, as a relatively large portion of the posterior samples fall on one side of the posterior probability distributions of these two answer key estimates, and the posterior predictive model check (see later) supports the assumption that there is a single culture generating responses in the data, we conclude that the consensus-answer on both of these items is ‘False’.

*Insert Table 1 about here *
The fourth column in Table 1 displays the item difficulty level. Values close to 1 indicate very difficult items. Not surprisingly, the items with larger uncertainty in their posterior consensus answer estimates are also the most difficult ones. The easiest item seems to be the one stating ‘Behavior always involves movement’, which is regarded as ‘False’ in the current analysis.

The right half of Table 1 displays the results based on the psychology student sample. Generally speaking the consensus pattern is very similar to the biologist sample, except for two items, namely the ‘Behavior is always in response to the external environment’ and ‘All behaviors are directly observable, recordable and measurable.’ However, the posterior standard deviation estimates show the largest degree of uncertainty for these items. In fact, these estimates have so much uncertainty in the student sample that the consensus model based analysis cannot provide a very reliable judgment about the consensus answers for these two behavioral features in the student population.

Except for the two uncertain items described above (‘Behavior is always in response to the external environment’ and ‘All behaviors are directly observable, recordable and measurable’), the rest of the consensus estimates in terms of True/False appeared to be identical to the behavioral biologist population. This suggests that the basic concepts of behavior are shared relatively well between the two fields. However, the psychology students are still in the process of acquiring the scientific concepts, therefore their consensus estimates are generally more uncertain when compared to the behavioral biologist sample.
Consensus in the three behavioral biologist societies was examined separately as well. Table 2 shows the results for the answer key estimates in the full sample as well as for the three societies. It can be seen that the consensus answer estimates in two large societies (Animal Behavior Society, International Society for Applied Ethology) are identical to each other and with that of the full sample. The difference between the two societies can be found when we look at the uncertainty in the estimates: while the ABS estimates show practically no uncertainty, the ISAE consensus estimates have large uncertainty for a couple of items.

*Insert Table 2 about here*

As discussed before, there were two items for which the original analysis on the full sample (Levitis et al., 2009), was not able to determine consensus-based answers, those are highlighted in Table 2. The estimates in the sub-samples suggest that the two large groups are in agreement with the consensus derived in the full sample, at least in terms of the final labeling of the posterior point estimates. However, we note again that ABS estimates show almost no uncertainty, unlike those of the ISAE, where several have high uncertainty. In contrast, the Society for Plant Neurobiology (small sample, N = 13) seems to disagree with them with respect to the ‘All behaviors are directly observable, recordable and measurable’ (full sample: ‘False’, PNB: ‘True’). As it turns out PNB estimates are in disagreement with respect to a couple of other items as well; however, these estimates for PNB have a lot of uncertainty.

Having estimated a consensus-based answer key, we can formulate a definition of
behavior. Based on the current study relying on modeling patterns in decision making and agreement with respect to the concept of behavior, we conclude that: “Behavior is a response by the whole individual to external and/or internal stimulus, influenced by the internal processes of the individual, and is typically not a developmental change.” This definition has implications. It entails that behavior results from responses to stimuli. However, in contrast to a traditional behaviorist view, behavior does include all measurable actions by an organism. Indeed, according to the consensus definition, behavior is not always observable. In particular, internal processes, that might not be observable, play an important role in forming behavior. The consensus definition also rules out developmental change as part of behavior, implying a more goal-oriented definition of behavior.

**Testing the one underlying culture assumption**

The question arises whether the responses can be considered to come from a single underlying culture. For the current data sets this assumption was tested by a so-called posterior predictive model checking technique (Gelman et al., 2004; Batchelder & Anders, 2012). In short, minimum residual factor-analysis (Comrey, 1962) was applied to calculate eigenvalues from the informant-by-informant correlation matrices for the full behavioral biologist sample and for the sub-samples. A one factor solution (one underlying culture) is indicated by a sharp decline after the first eigenvalue, with only minimal differences among the rest of the eigenvalues. This pattern was recovered in the original data sets, as well as in simulated data sets based on the HECM. Details on the analysis and graphical illustrations of the results can be found in
Appendix D. To conclude, the posterior predictive checks support the one-underlying culture assumption for the two full samples, as well as for the sub-samples of the behavioral biologist sample.

*Characteristics of the two cultures*

To compare behavioral biologist and psychology student populations even further, we look at population level characteristics of the two groups, as displayed in Table 3. All numbers are probabilities. The average ability is 0.54 in the behavioral biologist population, which means that on average respondents have a better-than-even probability of knowing the consensus answer to items with item difficulty level lower than 0.54 (see item difficulty values in the last column of Table 1). Not surprisingly the average ability is lower in the psychology student sample, contributing to the larger level of uncertainty in their answer key estimates. Looking at the guessing bias, while 0.5 would represent neutral guessing tendency, it turns out that both groups prefer to guess ‘False’, which is an interesting finding if we take into account that all items were derived from existing definitions of behavior. Finally, willingness to guess when faced with uncertainty is, on average, higher among the psychology students.

*Insert Table 3 about here*
Predicting latent cognitive variables

Results from regressing the predictors on the person-specific parameters are summarized and displayed in Table 4. The predictors consist of self-reported expertise and of consistency in responses (as described above). The model parameters are ability, willingness to guess, and guessing bias. The point estimates for the weights were calculated by taking the mean of the posterior samples, while the standard errors in the estimates were represented by the standard deviations in the posterior samples. The question of interest about regression weights concerns how likely it is that they have effects on the criterion variables, namely on the person-specific ECM parameters. To carry out this inference, Bayesian $p$-values were computed from the posterior samples. This statistic is equal to twice the parameter’s posterior sample mass below 0 if the posterior mean is positive, or twice the mass above 0 if it is negative. It can be interpreted as the smallest symmetric tail area that contains 0. The smaller the $p$-value, the stronger the data support the conclusion that the regression weight differs from 0. A $p$-value around or smaller than 0.05 indicates credible evidence that the explanatory variable has an effect on our person-specific parameters, in the direction of the posterior mean.

*Insert Table 4 about here*

First we consider the results on the ability parameter from Table 4 in the behavioral biologist sample. As can be seen in the first row concerning ability, the self-reported expertise level had a positive posterior mean estimate (0.1558), but the magnitude of this effect was low
and the corresponding $p$-value indicated no substantial evidence that the ability differs as a function of this predictor. In contrast, the $p$-value in the next row concerning consistency in judgment as a predictor had a credibly nonzero regression coefficient. This supports the hypothesis that the consistency in judgments (agreement between considering the phenomena as an example of behavior, or not, and accepting its features as fundamental for a behavioral definition, or not) positively affects the ability related to knowing the consensus answer. However, the same connection was not found in the psychology student sample.

Consistency in judgments was connected to response style as well: higher consistency was related to higher willingness to guess in both samples, and bias toward guessing ‘True’ as a consensus answer in the behavioral biologist sample. In the full sample there was a bias toward guessing ‘False’, however behavioral biologists who gave more consistent responses were more likely to guess ‘True’ than were those who gave less consistent responses.

**Conclusions**

The hierarchical extended Condorcet Model provides the researcher with a method to examine the existence, extent and attributes of consensus among a group of respondents in their knowledge about some content domain. The original problem of searching for definitions of behavior, as described by Levitis et al. (2009), represents a typical area where ECM can contribute to our understanding of concepts in psychology. The proposed features of behavior are used as items to summarize theories and knowledge about behavior. By taking advantage of modeling person- as well as item-specific characteristics of the data, we were able to infer
consensus knowledge based answers on two items for which simple aggregation techniques failed to provide a definitive conclusion in the original article.

By adding data from a sample of psychology students, and by applying modern Bayesian analysis to the assessment of consensus, we have acquired a more complete picture on the consensus about behavioral definitions. For most items in the original sample we were able to derive consensus-based answers and quantify uncertainty for all of the consensus estimates. In the student sample it was shown that the uncertainty level was larger, and although their consensus answers were different for two items, these estimates had so much uncertainty that we could not establish a substantial difference in the answer key estimates between the two groups. We conclude with the reassuring thought that even where we do not have an officially agreed upon definition of a concept, or full agreement on what qualifies as a certain concept, we can still examine consensus, and describe where and to what extent it exists. Using this approach we can potentially create solid definitions for complicated concepts such as behavior.

Finally, we believe that there are several areas in the field of psychology where our approach can be applied. In psychodiagnosis it could be used for conceptualizing the agreement on different aspects of psychopathological classifications. In personality psychology the most important features of different personality types could be derived based on the consensus answers from experts in the field. Generally speaking, cultural consensus theory can be used in any scientific field to examine consensus on the meaning of abstract concepts and to clarify the agreed-upon features of the concepts.
References


Methodology.


Appendix A

Prior Specifications

In the Bayesian framework prior information on the model parameters have to be defined. For the person and item specific parameters this prior information is provided by the population distributions described above. For the parameters of these population distributions a non-informative prior information is set. Formally, a diffuse normal distribution is defined on the regression parameters:

$$\beta_f \sim N_{J+1}(0, 10 \mathbf{I}_{J+1})$$

where $f$ can indicate any of the $F=3$ model parameters: $\theta$, $b$ and $g$ and where $\mathbf{I}$ stands for the identity matrix. As prior on the covariance matrix, a inverse-Wishart distribution (denoted as inv-W$^{-1}$) is chosen:

$$\Sigma \sim W^{-1}\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, df\right),$$

with $df=3$ degrees of freedom. In case of the population distribution of the $\delta_k$, a uniform (denoted as U) distribution is chosen on as a prior the standard deviation parameters:

$$\sigma_k \sim U(0.01, 10).$$

As prior on the Bernoulli probability for the answer key $Z_k$, we assign a uniform (denoted as U) prior distribution:
In the analyses of the small subsample of PNB and in the psychology student sample, in both of which the knowledge level was rather low, $\pi$ was fixed to 0.5 to facilitate convergence. $\pi = 0.5$ designates equal chances of ‘True’ and ‘False’ consensus answers.
Appendix B

**Items of proposed features of behavior**

A. A developmental change is usually not a behavior.

B. Behavior is always in response to the external environment.

C. A behavior is always an action, rather than a lack of action.

D. All behaviors are directly observable, recordable and measurable.

E. People can all tell what is and isn’t behavior, just by looking at it.

F. Behavior is always influenced by the internal processes of the individual.

G. Behavior always involves movement.

H. Behaviors are always the actions of individuals, not groups.

I. Behavior is something whole individuals do, not organs or parts that make up an individual.

J. A behavior is always in response to a stimulus or set of stimuli, but the stimulus can be either internal or external.

K. Behavior is something only animals (including people) do, but not other organisms.

L. In humans, anything that is not under conscious control is not behavior.

M. Behavior is always executed through muscular activity.
Appendix C

Items of phenomena that might be judged as behaviors

1. Ants that are physiologically capable of laying eggs don’t do so because they aren’t queens (C, G).

2. A sponge pumps water to gather food (B, M).

3. A spider builds a web.


5. A plant’s stomata (respiration pores) close to conserve water (I, K, M).

6. A plant bends its leaves towards a light source (K, M).

7. A person’s heart beats harder after a nightmare (B, I, L).

8. A person sweats in response to hot air (G, I, L, M).

9. A beetle is swept away by a strong current (F, M).

10. A rat has a dislike for salty food (B, C, G, J, M).

11. A person decides not to do anything tomorrow if it rains (B, C, G, J, M).


13. A mouse floats in zero gravity in outer space (E, F, G, M).

14. A group of unicellular algae swim towards water with a higher concentration of nutrients (F, H, K, M).

15. A frog orbits the Sun along with the rest of the Earth (F, M).
16. Flocks of geese fly in V formations (H).

17. A dog salivates in anticipation of feeding time (B, G, I, M).

18. Herds of zebras break up during the breeding season and reform afterwards (H).

19. A chameleon changes color in response to sunlight (G, M).

20. A cat produces insulin because of excess sugar in her blood (B, G, I, M).
Appendix D

Posterior predictive model checking of the one-culture assumption

A basic assumption of the HECM applied in the paper is that respondents represent one culture. In the Bayesian context this assumption can be tested by taking advantage of the posterior predictive model checking technique (see for example, Gelman et al., 2004). As a first step we select a test statistic that reflects the one underlying culture property. A good measure for it stems from minimum residual factor-analysis (Comrey, 1962), and it works with the eigenvalues calculated from the informant-by-informant correlation matrix. The first eigenvalue is supposed to be multiple times of the second one, with only minimal differences among the rest of the eigenvalues if there is one factor – one culture – underlying the data (for details, as well as for models with multiple cultures, see Anders & Batchelder, 2012).

When carrying out posterior predictive checking (PPC) data are generated based on the HECM and the conditional posterior distributions of the parameters. In the HECM, the existence of one underlying culture is a basic assumption. Therefore when data are generated from the HECM, the one culture assumption is automatically met. In turn, the eigenvalues that are calculated from these generated data sets represent a sequence of typical eigenvalues from data with one underlying culture.

As part of the PPC we generate 300 data sets from the HECM and the conditional posterior distributions of the parameters. Then we calculate the first, second and third
eigenvalues and display their pattern graphically. Our expectation is a large drop between the first and second eigenvalues and only a slight slope between the second and the third, indicating one underlying consensus answer key. Now we calculate the eigenvalues from the ‘real’ data and plot their pattern. If the real data set stem from one culture, the eigenvalue ratio based on the real data set should fall inside the area designated by the pattern from the generated data sets.

The five panels of Figure 2 show the results from five such graphical tests: from Panel (a)-(e), the results correspond to the full biologists sample, psychology student sample, ABS, ISAE, and PNB, respectively. The grey lines correspond to the eigenvalue graphs based on the simulated data sets, while the black lines display the same graph based on the real data set. As can be seen in Panel (a) (behavioral biologists), the real data falls inside the grey area, suggesting that the one culture assumption is met; however, the evidence is not overwhelmingly convincing, as the black line in the middle of the graph (at the second eigenvalue) approaches the top of the grey area quite well. When Panel (a) (behavioral biologists) is compared to Panel (b) (psychology students), the black line in the latter one lies almost exactly in the middle of the grey area, strongly supporting the one culture assumption. Finally, Panel (c)-(e) shows that the behavioral biologist societies meet the one culture assumption quite satisfactorily.
Acknowledgements

We would like to thank to an anonymous reviewer and to Kate Slaney for their useful comments.
Appendix F

Funding

This material is based in part on work supported by the U.S. Army Research Office under Grant Number FA9550-09-1-0510.

This material is based in part on work supported by the Oak Ridge Institute for Science and Education (ORISE).

The work of the third author was supported by the Max-Planck Odense Center on the Biodemography of Aging, a collaboration between the Max Planck Society and the University of Southern Denmark.
Author Note

Correspondence concerning this article may be addressed to: Zita Oravecz, UCI, Department of Cognitive Sciences, 3213 Social & Behavioral Sciences Gateway Building, Irvine, CA 92697-5100; phone: +19498245984; email: zoravecz@uci.edu
Footnotes

1 Our use of ‘consensus’ differs from the use of consensus as ‘coming to agreement’, which plays out in science in the process by which researchers come to agree about the meanings of theoretical concepts. Rather, we use ‘consensus’ to refer to the extent of overlap in the criteria people attribute to specific theoretical terms.

2 From the original sample those respondents who did not answer any of the behavioral definition items were removed. Then a further two respondents were removed as they did not reply to the question concerning their knowledge level in behavioral biology.

3 The respondents had three answer options according to the original coding: Agree, Do Not Agree, or Not sure, with respect to whether the statement describes a feature of the behavior or not. The original wording is somewhat confusing when it comes to interpreting the CCT results as an important feature of the model is the agreement/disagreement among the respondents. Therefore we simplify by relabeling the answers as ‘True’, ‘False’ or ‘Don’t know’, in terms of whether it is a feature of or example of behavior.
Table 1: Posterior point estimates on the 13 items of the proposed features of behavior.

<table>
<thead>
<tr>
<th>Item keywords</th>
<th>Behavioral biologist</th>
<th>Psychology students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Answer key</td>
<td>Item difficulty</td>
</tr>
<tr>
<td></td>
<td>label</td>
<td>PSD</td>
</tr>
<tr>
<td>‘not developmental’</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘always response to external’</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘always action’</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘observable, recordable, measurable’</td>
<td>False</td>
<td>0.38</td>
</tr>
<tr>
<td>‘people can tell’</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘influenced by internal’</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘always movement’</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘action of individuals’</td>
<td>False</td>
<td>0.26</td>
</tr>
<tr>
<td>‘whole individual’</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘response to stimuli’</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘only animals’</td>
<td>False</td>
<td>0.34</td>
</tr>
<tr>
<td>‘always conscious’</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘always muscular activity’</td>
<td>False</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Table 2: Posterior point estimates of the answer key on 13 items of the proposed features of behavior in the full behavioral biologist sample and in ABS, ISAE and PNB.

<table>
<thead>
<tr>
<th>Item</th>
<th>All (N=169)</th>
<th>ABS (N=96)</th>
<th>ISAE (N=60)</th>
<th>PNB (N=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>keywords</td>
<td>Label</td>
<td>PSD</td>
<td>Label</td>
<td>PSD</td>
</tr>
<tr>
<td>‘not developmental’</td>
<td>True</td>
<td>0.00</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘a. response to external’</td>
<td>False</td>
<td>0.00</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘a. action’</td>
<td>False</td>
<td>0.00</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘observable,recordable,measurable’</td>
<td>False</td>
<td>0.38</td>
<td>False</td>
<td>0.04</td>
</tr>
<tr>
<td>‘people can tell’</td>
<td>False</td>
<td>0.00</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘influenced by internal’</td>
<td>True</td>
<td>0.00</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘always movement’</td>
<td>False</td>
<td>0.00</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘action of individuals’</td>
<td>False</td>
<td>0.26</td>
<td>False</td>
<td>0.02</td>
</tr>
<tr>
<td>‘whole individual’</td>
<td>True</td>
<td>0.00</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘response to stimuli’</td>
<td>True</td>
<td>0.00</td>
<td>True</td>
<td>0.00</td>
</tr>
<tr>
<td>‘only animals’</td>
<td>False</td>
<td>0.34</td>
<td>False</td>
<td>0.01</td>
</tr>
<tr>
<td>‘a. conscious’</td>
<td>False</td>
<td>0.00</td>
<td>False</td>
<td>0.00</td>
</tr>
<tr>
<td>‘a. muscular activity’</td>
<td>False</td>
<td>0.02</td>
<td>False</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 3: ECM population level results based on the behavioral biologist and the psychology student samples.

<table>
<thead>
<tr>
<th>Samples</th>
<th>Average ability</th>
<th>Guessing bias</th>
<th>Willingness to guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral biologists</td>
<td>0.54</td>
<td>0.45</td>
<td>0.87</td>
</tr>
<tr>
<td>Psychology students</td>
<td>0.40</td>
<td>0.29</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Table 4: Results on the covariates based on the data of the 13 items describing the proposed features of behavior from behavioral biologists and psychology students.

<table>
<thead>
<tr>
<th>Model parameter</th>
<th>Explanatory variable</th>
<th>Behavioral biologist</th>
<th>Psychological students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Posterior mean</td>
<td>Posterior std</td>
</tr>
<tr>
<td>Ability</td>
<td>Expertise</td>
<td>0.1558</td>
<td>0.1044</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>0.2004</td>
<td>0.1109</td>
</tr>
<tr>
<td>Willingness to guess</td>
<td>Expertise</td>
<td>0.0288</td>
<td>0.1148</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>0.6045</td>
<td>0.1275</td>
</tr>
<tr>
<td>Guessing bias</td>
<td>Expertise</td>
<td>-0.0704</td>
<td>0.1391</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>0.4326</td>
<td>0.1489</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1. Decision making process based on the Extended Condorcet Model

Figure 2. The five panels display the results from five graphical posterior predictive checks based on eigenvalues. The grey lines correspond to the eigenvalue graphs based on the simulated data sets, while the black line displays the same graph based on the original data set. Panel (a)-(e), corresponds to the full biologists sample, psychology student sample, ABS, ISAE, and PNB, respectively.
“All behaviors are directly observable, recordable and measurable”

Consensus answer is ‘True’ ($Z_k$)

Knows ($D_i$)

- ‘True’

Willing to guess ($b_i$)

- ‘Don’t know’

Guessing ‘True’ ($g_i$)

- ‘True’
  - ‘False’

Consensus answer is ‘False’ ($Z_k$)

Knows ($D_i$)

- ‘False’

Willing to guess ($b_i$)

- ‘Don’t know’

Guessing ‘True’ ($g_i$)

- ‘True’
  - ‘False’