DOI 10.3758/s13428-011-0138-0

CAML—Maximum likelihood consensus analysis

André Aßfalg · Edgar Erdfelder

Published online: 27 July 2011

Abstract Consensus analysis enables estimation of individual differences in competencies and response tendencies when answer keys to dichotomous forced-choice questions are unknown. CAML, a set of functions written in R, implements maximum likelihood estimation for the general Condorcet model that underlies consensus analysis. CAML avoids problems of alternative approaches that have often rendered consensus analysis impractical or unfeasible in the past. It provides (1) measures of model fit, (2) a measure of consensus, (3) point and interval estimates of competencies and response tendencies, and (4) an estimate of the unknown answer key. The present article describes the general Condorcet model, the CAML algorithms, and the handling of the software. In addition, the validity of CAML results is tested in a recognition memory study using selective experimental manipulations of the parameters. The results show that CAML works very well in practice and provides valid estimates of competencies, response tendencies, and answer keys.

Keywords Cultural consensus analysis · General Condorcet model · Maximum likelihood estimation

The measurement of individual differences in competencies is one of the core issues in psychological assessment. Typically, item-response models are used to derive estimates of competence on the basis of a set of items with a well-defined answer key. The answer key enables classification of participants' responses as correct or incorrect, separately for each item. Responses scored correct or incorrect are then used as the basic data that enter into the analysis.

Sometimes, however, the situation is more complicated. This is certainly the case when the answer key to a set of items is unknown to the assessor. Consensus analysis (Batchelder, Kumbasar, & Boyd, 1997; Batchelder & Romney, 1986, 1988; Romney, Batchelder, & Weller, 1987; Romney, Weller, & Batchelder, 1986) was designed to handle this difficult situation. Such situations arise, for example, in the field of anthropology for which consensus analysis was developed originally. An anthropologist exploring an unknown culture must rely on the reports of informants. Assuming that several independent informants respond to the same questions concerning their culture, they will probably agree for some of the questions; for others, they will not. The true state of affairs is unknown to the anthropologist studying the informants' reports. Hence, responses cannot simply be scored as "correct" or "incorrect" (Romney et al., 1987; Romney et al., 1986).

The basic problem structure addressed by consensus analysis is also found in other areas of research. For example, in eyewitness testimony, several eyewitnesses may report on the same set of questions concerning a critical event—for example, an accident, a robbery, or a murder. Again, there is no predefined answer key, and the eyewitnesses will typically agree on some details, but not on others. These are just two examples to illustrate problems of *test theory without an answer key* (Batchelder & Romney, 1988).

Individual differences in informant competencies are a major reason for discrepant responses to the same set of questions. The precision of eyewitness testimony, for

A. Aßfalg (✉) · E. Erdfelder
Lehrstuhl Psychologie III, Universität Mannheim,
68131 Mannheim, Germany
e-mail: asfalg@psychologie.uni-mannheim.de

example, depends on the knowledge of the witness concerning a certain event. Situational, cognitive, and motivational factors, such as visual perspective, distance to the event, focus of attention, forgetting, and motivational involvement, may affect this knowledge. In addition, possible response tendencies, such as the willingness to respond "yes," have to be taken into account. Thus, individual differences in both competencies and response tendencies play a crucial role in the assessment of responses when the answer key is unknown to the assessor. In such cases, consensus analysis provides a valuable method for the conjoint estimation of individual competencies and the answer key to a set of questions (Romney et al., 1986).

In the present article, we introduce CAML (*consensus analysis via maximum likelihood*), a new consensus analysis tool written for the R statistics software (R Development Core Team, 2011). R is freely available and is supported by an active community that continuously provides software solutions to various statistical problems (e.g., Bulté & Onghena, 2008; Grassie, Luccio, & Di Blas, 2010; Nimon, Lewis, Kane, & Haynes, 2008). CAML makes use of the maximum likelihood method of parameter estimation. Notably, CAML does not require simplifying assumptions underlying the estimation method used most often, if not exclusively, in consensus analysis—namely, factor analysis (Batchelder & Romney, 1988; Romney et al., 1986). By providing improved software for consensus analysis, we hope that future choices of estimation procedures will be influenced less by computational simplicity of the estimation method and more by the performance of the method. For noncommercial purposes, CAML is freely available at http://psycho3.uni-mannheim.de/index.php?n=Main.CAML, along with sample data and a quick-start guide.

The second purpose of this study was to test the validity of the *general Condorcet model* (GCM) that underlies consensus analysis as implemented in CAML. We conducted a recognition memory experiment that required participants to memorize a set of words for a later recognition test including both old words studied previously and new words randomly intermixed. Of course, the answer key (i.e., whether a test item is old or new) is known in this application. However, by applying consensus analysis to the recognition judgments, the answer key is ignored. This enables the comparison of CAML estimates with the estimates obtained when the actual answer key is used. Moreover, we experimentally manipulated independent variables that should affect GCM parameters selectively if the model is psychologically valid. Although the GCM has often been used in anthropology (e.g., Romney et al., 1987), social network analysis (Batchelder et al., 1997), and cognitive psychology (e.g., Ameel, Storms, Malt, & Sloman, 2005; Bailenson, Shum, Atran, Medin, & Coley, 2002; Barg et al., 2006; Godoy et al., 2008; Johnson, Mervis, & Boster, 1992; Majid, Boster, & Bowerman, 2008; Malt, Sloman, Gennari, Shi, & Wang, 1999; Medin et al., 2006; Shafto & Coley, 2003), this is, to our knowledge, the first test of the validity of the GCM by means of selective experimental manipulations of the model's parameters.

In the next section, we briefly introduce the GCM. This is followed by a short description of the de facto standard for consensus analysis so far, the factor-analytic approach. Next, we discuss maximum likelihood estimation for the GCM. We describe how to estimate model parameters and confidence intervals and how to test for consensus among informants within this approach. This is followed by a discussion of the advantages of CAML, as compared with existing software for maximum likelihood estimation in multinomial models. Subsequently, we explain CAML handling—that is, data input, specification of the analysis, and the output of CAML. In the final section, we present the validation study of the GCM sketched above.

The general Condorcet model

The following outline of the GCM makes use of the notation previously introduced in the literature (Hu & Batchelder, 1994; Karabatsos & Batchelder, 2003). Assume that N informants respond to a set of M questions with one of two possible responses—for example, "yes" or "no." The response matrix X_{ik} contains the responses of informants $i = 1, \ldots, N$ to items $k = 1, \ldots, M$, such that

$$X_{ik} = \begin{cases} 1, & \mathit{if\ informant\ i\ answers\ ''yes''\ to\ item\ k} \\ 0, & \mathit{if\ informant\ i\ answers\ ''no''\ to\ item\ k.} \end{cases}$$

The response matrix is the only input required to perform consensus analysis. Furthermore, the (unknown) answer key for item k is given by

$$Z_k = \begin{cases} 1, & \mathit{if\ the\ correct\ response\ to\ item\ k\ is\ ''yes''} \\ 0, & \mathit{if\ the\ correct\ response\ to\ item\ k\ is\ ''no''.} \end{cases}$$

Consensus analysis accounts for all observations in the N-dimensional contingency table—that is, for all observed item-specific response patterns $<X_{1k,} \ldots, X_{ik}, \ldots, X_{Nk}>$, $k = 1, \ldots, M$, across the N informants. Because some response patterns may occur more than once, there are $J \leq M$ nonredundant response patterns (with $J = M$ only when none of the patterns occurs at least twice). We use $j = 1, \ldots, J$ as an index representing nonredundant response patterns. Each of the J nonredundant response patterns corresponds to one of the 2^N cells of the N-dimensional contingency table.

The GCM explains the 2^N possible data categories using three types of parameters. P_Z is the probability that the correct response is "yes." By implication, the complement 1 - P_Z is the probability that the correct response is "no." Furthermore, each informant is characterized by two probabilities. In accordance with signal detection theory (Macmillan & Creelman, 2008), these probabilities are called *hit*, H_i = $P(X_i$ = "yes" | Z = "yes") (i.e., the probability of informant i responding "yes" given that the correct response is "yes") and *false alarm*, F_i = $P(X_i$ = "yes" | Z = "no") (i.e., the probability that i responds "yes" given that the correct response is "no").

If we denote the vector of model parameters by $\theta = < P_Z, H_{i=1}^N, F_{i=1}^N >$, then the conditional probability of response pattern j, given that the correct answer to an item is "yes," is

$$p_{j|1}(\theta) = \Pi_{i=1}^N H_i^{Xij}(1 - H_i)^{(1-Xij)} \tag{1}$$

By analogy, given that the correct answer is "no," the conditional probability of pattern j is

$$p_{j|0}(\theta) = \Pi_{i=1}^N F_i^{Xij}(1 - F_i)^{(1-Xij)} \tag{2}$$

Thus, the unconditional probability of response pattern j is

$$p_j(\theta) = P_Z p_{j|1}(\theta) + (1 - P_Z)p_{j|0}(\theta). \tag{3}$$

An important special case of the GCM is based on the assumption that hits and false alarms of the informants depend (1) on their competencies D_i to detect a target item or a lure item and (2) on their response tendencies g_i to respond "yes" in the same way as is assumed in the so-called two-high threshold (2HT) model of recognition (Snodgrass & Corwin, 1988). Hence, we call this version of the GCM *the two-high threshold general Condorcet model* (2HT-GCM). According to the 2HT-GCM, informant i knows the true answer key and responds correctly with probability D_i to each item. However, if the answer key is unknown with probability 1 - D_i, informant i guesses "yes" with probability g_i and "no" with probability 1 - g_i. For the sake of convenience, we will henceforth refer to D_i as the competence parameter and g_i as the response tendency parameter. The standard GCM and the 2HT-GCM are equivalent, provided that $H_i \geq F_i$ holds for all i = 1, …, N. In this case,

$$D_i = H_i - F_i \tag{4}$$

and

$$g_i = \frac{F_i}{1 - H_i + F_i} \tag{5}$$

The order restriction that false alarms must not exceed hits adds to the complexity of the 2HT-GCM, as compared with the unrestricted GCM. Therefore, the 2HT-GCM is computationally more intensive than the GCM. However, the competence and response tendency parameters of the 2HT-GCM are easier to interpret than hits and false alarms of the unconstrained GCM. Obviously, arguments in favor of each model variant exist.

The factor-analytic estimation procedure

The most widely used estimation procedure for the GCM is based on factor analysis (Batchelder & Romney, 1988). This procedure employs a restricted version of the GCM assuming g_i = .5 for each informant i = 1, …, N. In a first step, the response matrix X_{ik} is transformed into a so-called matching score for each possible pair of informants. The matching score for informants i and l is simply the proportion of matching responses in X_{ik} and X_{lk} across all items k = 1, …, M. Computation of matching scores for all possible pairs of informants results in a $N \times N$ matching score matrix. In the second step, factor analysis is performed on the matching score matrix. Because the main diagonal of the matching score matrix is of no interest in consensus analysis, Batchelder and Romney (1988) suggested the minimum residual method for factor analysis (Comrey, 1962), which ignores the main diagonal. Batchelder and Romney (1988) showed that the factor loadings of the first unrotated factor estimate the informants' competencies, $\widehat{D}_i$, provided that g_i = .5, for all informants i = 1, …, N.

Although the factor-analytic approach to consensus analysis is easily applied in practice and, therefore, quite popular, there are three disadvantages to this procedure. First, the factor-analytic procedure relies on bivariate associations between informants as represented in the matching score matrix. It ignores possible higher-order associations between informants. For example, triple-order associations may occur when the degree of correspondence between the answers of two informants depends on the answers of a third informant. Second, while the assumption of homogeneous and neutral response tendencies (i.e., g_i = .5) may be adequate in many cases, it may be inadequate in others. There is no way to test this critical assumption in the factor analysis framework. Third, the factor-analytic approach relies on a very mild goodness-of-fit criterion. Fit is considered acceptable whenever factor analysis yields a one-factor solution, a criterion for which a formal test has not yet been developed.

The maximum likelihood estimation procedure

If we denote the observed frequency of response pattern j across items by c_j, then the likelihood function of the data under the GCM is

$$L\left(c_{j=1}^{J};\theta\right) = M!\Pi_{j=1}^{J}\frac{p_j(\theta)^{c_j}}{c_j!}. \tag{6}$$

Hu and Batchelder (1994) developed a version of the EM algorithm (Dempster, Laird, & Rubin, 1977) that provides an estimate of θ corresponding to at least a local maximum in the likelihood function (6) for a given set of observed frequencies c_j, j = 1, …, J. CAML uses Hu and Batchelder's (1994) EM algorithm to estimate the parameters of the GCM.

We implemented the unrestricted GCM in CAML to avoid the computational complexity of the 2HT-GCM. Thus, CAML first estimates hit and false alarm rates according to the unrestricted GCM. Subsequently, competence and response tendency estimates are derived by inserting maximum likelihood estimates of hit and false alarm rates in Eqs. 4 and 5. Note that the resulting D_i estimates may exceed the interval $0 \leq D_i \leq 1$, i = 1, …, N. We return to this issue below.

The GCM as defined by Eqs. 1–3 is also known as the *latent class model* with two latent classes. However, in contrast to latent class models that assign respondents to latent classes (Goodman, 1974; Lazarsfeld & Henry, 1968), the GCM assigns items to latent classes. These item classes correspond to the two values of the answer key (Batchelder & Romney, 1988). Because latent class models are special cases of *multinomial processing tree* (MPT) models (Batchelder & Riefer, 1999; Erdfelder et al., 2009), the GCM can, of course, also be seen as a special form of MPT models. This perspective has many computational advantages. In fact, most of the methods used in CAML are based on the MPT modeling approach (Hu & Batchelder, 1994; Riefer & Batchelder, 1988).

Model fit and model selection

Goodness of fit of the GCM can be evaluated using power divergence statistics (Read & Cressie, 1988)—for example, the well-known likelihood ratio statistic G^2 or Pearson's X^2. Because of the models' complexity, however, assessing model fit is a nontrivial exercise. Recall that the GCM captures 2^N possible response patterns across N informants. Hence, the number of possible response patterns is typically much larger than M, the number of observations (i.e., items). The GCM with 30 informants, for example, comprises more than a billion possible response patterns. By implication, even if the data set includes several hundred items, many data categories will remain empty. Hence, the boundary conditions for using asymptotic chi-square distributions as reference distributions for the power divergence statistics are not met (Read & Cressie, 1988). One way to cope with this problem is estimating the exact distributions using the parametric bootstrap (Collins, Fidler, Wugalter, & Long, 1993; Langeheine, Pannekoek, & van de Pol, 1996). Parametric bootstrapping (Efron, 1982) involves Monte Carlo sampling from a population model using maximum likelihood estimates based on the observed sample as parameters. For each Monte Carlo sample, the GCM is fitted again. The empirical distribution of the Monte Carlo fit statistics can then be used as an approximation to the exact distribution under the null hypothesis.

As an additional measure of model fit, CAML provides $\Delta BIC = BIC(H_0) - BIC(H_1)$, where BIC($H_1$) is the Bayesian information criterion for the saturated model producing perfect fit and BIC(H_0) is the BIC value for the fitted model (e.g., Carlin & Louis, 1996). If J denotes the number of nonredundant response patterns, J - 1 parameters are required in the saturated model. The GCM, in contrast, includes $2N + 1$ parameters only. Hence,

$$\Delta BIC = G^2 + \log M[(2N+1)-(J-1)].$$

The smaller ΔBIC is, the better the fit of the GCM. Negative values of ΔBIC clearly support the GCM over the saturated model. As compared with power divergence statistics such as G^2, the ΔBIC fit criterion has the advantage of being less dependent on sample size and of controlling for the number of parameters in the model (Read & Cressie, 1988).

The likelihood ratio statistic G^2 can also be used to compare nested models—that is, pairs of models with one model resulting from one or more parameter constraints applied to the other model. Possible restrictions include equality constraints and parameter fixations. An equality constraint of special importance for the GCM is $H_\mathrm{i} = 1 - F_\mathrm{i}$ for i = 1, …, N. If applied to the 2HT-GCM, this constraint is equivalent to $g_i = .5$ for i = 1, …, N. This can be seen by replacing H_i with 1 - F_i in Eq. 5 and solving for g_i. If this restriction holds, informants favor neither of the two response options in cases of response uncertainty. As was outlined in the previous section, this assumption is implied by the factor-analytic approach to consensus analysis (Batchelder & Romney, 1988). By evaluating the difference $\Delta G^2 = G^2_{rGCM} - G^2_{\mathrm{uGCM}}$ of the likelihood ratio statistics for the restricted GCM assuming $H_i = 1 - F_\mathrm{i}$ for i = 1, …, N $\left(G^2_{\mathrm{rGCM}}\right)$ and the unrestricted GCM $\left(G^2_{\mathrm{uGCM}}\right)$, it is possible to test this crucial assumption of the factor-analytic approach statistically. Simulation studies indicate that, in contrast to the global model fit statistics G^2_{uGCM} and G^2_{rGCM}, the difference statistic ΔG^2 is approximately chi-square distributed even in cases of sparse data (Agresti & Yang, 1987).

Measures of consensus

Consensus as defined in consensus theory is characterized by two features (Romney et al., 1986). First, consensus requires that informants share the same answer key. Second, the informants' competence parameters must be clearly positive. The rationale behind the latter requirement is that informants who do not make use of the answer key are unable to share it. On the basis of this definition of consensus, a measure of consensus is easily derived using CAML.

Recall that CAML employs a two-step procedure. First, hit and false alarm rates are estimated by applying the EM algorithm (Hu & Batchelder, 1994) to the unrestricted GCM. We denote these maximum likelihood estimates by $\widehat{H}_i$ and $\widehat{F}_i$, respectively. Second, competence and response tendency parameter estimates are derived by replacing H_i and F_i in Eqs. 4 and 5 by $\widehat{H}_i$ and $\widehat{F}_i$, respectively. In other words, CAML does not enforce $\widehat{H}_i \geq \widehat{F}_i$, i = 1, …, N, in the 2HT-GCM. Hence, the competence parameter estimate of informant i will be negative whenever Step 1 of the CAML algorithm results in $\widehat{H}_i < \widehat{F}_i$.

Because consensus implies positive competence parameters, the occurrence of at least one negative competence parameter estimate indicates lack of consensus (Romney et al., 1986). In fact, negative competence parameter estimates have frequently been used as measures of lack of consensus in the relevant literature (e.g., Bailenson et al., 2002; Johnson et al., 1992; Majid et al., 2008; Malt et al., 1999; Medin et al., 2006; Shafto & Coley, 2003).

Confidence intervals

Within the MPT framework (Hu & Batchelder, 1994), the *Fisher information matrix* can be used to obtain confidence intervals of parameter estimates. The main diagonal of the inverted Fisher information matrix includes variances of the parameter estimates (Efron & Hinkley, 1978). Because the asymptotic distribution of maximum likelihood estimates is Gaussian, these variance estimates can be used to construct confidence intervals. Hu and Batchelder (see also Moshagen, 2010) derived a closed form solution of the observed Fisher information for MPT models that is used in CAML.

There are, however, some disadvantages to using the Fisher information to determine confidence intervals. Because of small numbers of items, the number of observations is often small. Therefore, the estimate of the Fisher information may not be a good approximation to the true Fisher information (Hu, 1999). Other possible shortcomings include singularity of the Fisher matrix, negative variance estimates, and confidence intervals outside the interval [0, 1] that often occur in case of sparse data (Moshagen, 2010).

To avoid these problems, CAML additionally provides confidence intervals using the parametric bootstrap procedure (Efron, 1982). The rationale behind the parametric bootstrap is that the unknown distribution of parameter estimates can be approximated by repeatedly drawing Monte Carlo samples from the population model on the basis of the observed maximum likelihood estimates. The parameter of interest is estimated for each of the Monte Carlo samples. The sampling distribution of these estimates can then be used to extract some measures of variability, such as the standard error or $(1-\alpha) \times 100\%$ confidence intervals. It is also possible to obtain confidence intervals for the competence and response tendency parameters of the 2HT-GCM. Keep in mind, however, that these intervals may exceed the unit interval because $\widehat{H}_i \geq \widehat{F}_i$ is not enforced by the estimation algorithm.

Answer keys

Once point estimates of the model parameters are available, it is easy to estimate answer keys using Eqs. 1 and 2. The answer key estimate for item k, given the nonredundant response pattern j, is $\widehat{Z}_k = 1$ if $p_{j|1}(\widehat{\theta})P_Z > p_{j|0}(\widehat{\theta})(1-P_Z)$ and $\widehat{Z}_k = 0$ otherwise.

Shortcomings of existing software in the context of consensus analysis

As was outlined above, the GCM can be seen as an MPT model, as a latent class model, and as factor-analytic model. We will therefore discuss software for these model classes as possible alternatives to CAML. However, note that the MPT and latent class analysis software discussed below was developed for a much broader range of applications than CAML. Consequently, the issues discussed here are limited to consensus analysis and do not necessarily generalize to other applications of this software.

Despite the fact that CAML uses the MPT framework and employs the same version of the EM algorithm (Hu & Batchelder, 1994) also used in standard MPT software (Hu & Phillips, 1999; Moshagen, 2010; Rothkegel, 1999; Stahl & Klauer, 2007), consensus analysis is often unfeasible for the latter programs. For example, existing MPT software requires that the probabilities of all possible 2^N response patterns must be represented in the form of model equations. This poses serious problems. As was outlined above, the GCM for 30 informants includes more than a billion model equations. Obviously, it is impossible to communicate these equations to standard MPT software. CAML solves this problem by making use of the fact that the EM algorithm of Hu and Batchelder (1994) ignores model equations of response patterns with zero frequency. By implication, only J model equations—one for each of the nonredundant response patterns—need to be considered (see Eq. 3). CAML makes use of this incomplete but computationally sufficient set of model equations. Alternative MPT software, however, requires the full set of equations, an approach that is feasible for small N only.

Another concern is the lack of procedures to deal with sparse data in most of the currently available programs for MPT models. In such cases, the large-sample chi-square approximation to the likelihood ratio statistic G^2 does not work appropriately (Collins et al., 1993; Langeheine et al., 1996), and the estimate of the Fisher information matrix lacks precision (Hu, 1999). Bootstrap procedures that remedy these problems are not implemented in most of the software mentioned above (but see Moshagen, 2010, for an exception).

Furthermore, standard latent class software does not provide patterns of parameter restrictions that reflect meaningful variants of the GCM. For example, the assumption that response tendencies are homogeneous is conceptually interesting and practically easy to accomplish in CAML, but not in poLCA (Linzer & Lewis, 2010).

A minor issue with latent class software arises from the fact that the GCM is not globally identified. The reason for this is that the labeling of the latent classes (i.e., the values of the answer key) does not affect the expected frequencies under the model. In other words, the interpretation $P_Z = P(Z = 1)$ predicts the same response patterns as $P_Z = P(Z = 0)$. By implication, the estimation process may yield parameter estimates that correspond to either of the two possible interpretations of P_Z (Batchelder & Romney, 1988). To obtain unique estimates, we can make use of the fact that the average hit probability should exceed the average false alarm probability. CAML automatically chooses the interpretation of P_Z that complies with this constraint. In contrast, latent class software may produce solutions that correspond to both interpretations and, thus, does not guarantee unique estimates.

In addition to the software discussed above, the program ANTHROPAC (Borgatti, 1996) is able to perform consensus analysis on the basis of the factor-analytic method. ANTHROPAC has a broader range of applications than does CAML and provides additional data analysis methods often used in anthropology. However, the above-mentioned disadvantages of the factor-analytic approach to consensus analysis also apply to ANTHROPAC. For example, this program requires the assumption of homogeneous and neutral response tendencies (i.e., $g_i = .5$), does not provide a test of this assumption and also does not provide a technically sound statistical test of the GCM. In sum, as compared with the software discussed in the present section, we think that the advantages of CAML warrant and necessitate the introduction of a new consensus analysis tool.

The software

CAML requires the response matrix X_{ik} containing the responses of each informant to each item as input. Table 1 depicts the responses of four informants to 16 old–new questions in a recognition experiment. We will turn to the explanation of the data set in more detail below. The lower half of Table 1 illustrates the corresponding R input to define the response matrix—henceforth, abbreviated as X.

To apply consensus analysis to X, we implemented the function CAML(X, g, eps, runs, max.iter, fisher, boot.runs,

Table 1 The answers of four informants (I1–I4) to 16 old–new questions (upper half) and the corresponding definition of a response matrix named "X" in R (lower half)

	Item															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
I1	new	new	old	new	new	old	new	old	new	new	old	new	old	old	old	old
I2	new	new	new	new	new	new	new	new	old	old	old	old	old	old	new	new
I3	old	new	new	new	new	old	new	old	old	old	old	old	old	old	old	old
I4	new	old	new	new	new	old	new	new	old	old	old	old	old	old	old	old
X <- rbind (																
	c(0,	0,	1,	0,	0,	1,	0,	1,	0,	0,	1,	0,	1,	1,	1,	1),
	c(0,	0,	0,	0,	0,	0,	0,	0,	1,	1,	1,	1,	1,	1,	0,	0),
	c(1,	0,	0,	0,	0,	1,	0,	1,	1,	1,	1,	1,	1,	1,	1,	1),
	c(0,	1,	0,	0,	0,	1,	0,	0,	1,	1,	1,	1,	1,	1,	1,	1)
)																

alpha). All function arguments except X have default values. Hence, it is sufficient to enter CAML(X) for basic consensus analysis. This will provide the most important information: estimates of informant competencies, response tendencies, the answer key, and a measure of consensus based on negative competence estimates. All function arguments, their meaning, their admissible values, and their default values are listed in Table 2.

To assess model fit, the output also contains the likelihood ratio and the information criteria AIC (Akaike, 1973) and BIC (Schwarz, 1978). In addition, the BIC difference between the saturated model and the fitted model (ΔBIC) is included in the output. As already explained in more detail above, due to the sparseness of data, the likelihood ratio statistic G^2 usually does not follow a chi-square distribution if the model holds. Hence, the exact distribution is estimated using the parametric bootstrap. The argument boot.runs specifies the number of bootstrap samples used to determine the p-value for the observed G^2 statistic. Large values of boot.runs will provide more precise estimates of p, especially if p is small, but also will increase computation time. The number of bootstrap samples necessary to achieve a predefined precision can be calculated with the procedure developed by Andrews and Buchinsky (2000).

As was outlined above, the factor-analytic procedure requires homogeneous and neutral guessing parameters; that is, all informants are equally inclined to respond "yes" or "no" if the correct response is unknown (Batchelder & Romney, 1988; Romney et al., 1986). The maximum likelihood approach to consensus analysis does not require this constraint. In CAML, it is possible to test the hypothesis that response tendencies are homogeneous using the argument g. This argument accepts Boolean values T or F, meaning *true* and *false*, respectively. If g = T, the EM algorithm is applied to both the general, unrestricted GCM and the restricted GCM assuming $g_i = .5$ for all informants $i = 1, \ldots, N$. Moreover, the output includes the likelihood ratio difference statistic ΔG^2 and its critical value based on the assumption that ΔG^2 is chi-square distributed. The Type I error probability is defined by the argument alpha. The necessary sample size to achieve a certain Type II error rate or the achieved Type II error rate, given a certain sample size, can be computed using power analysis software such as the chi-square procedure of G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009).

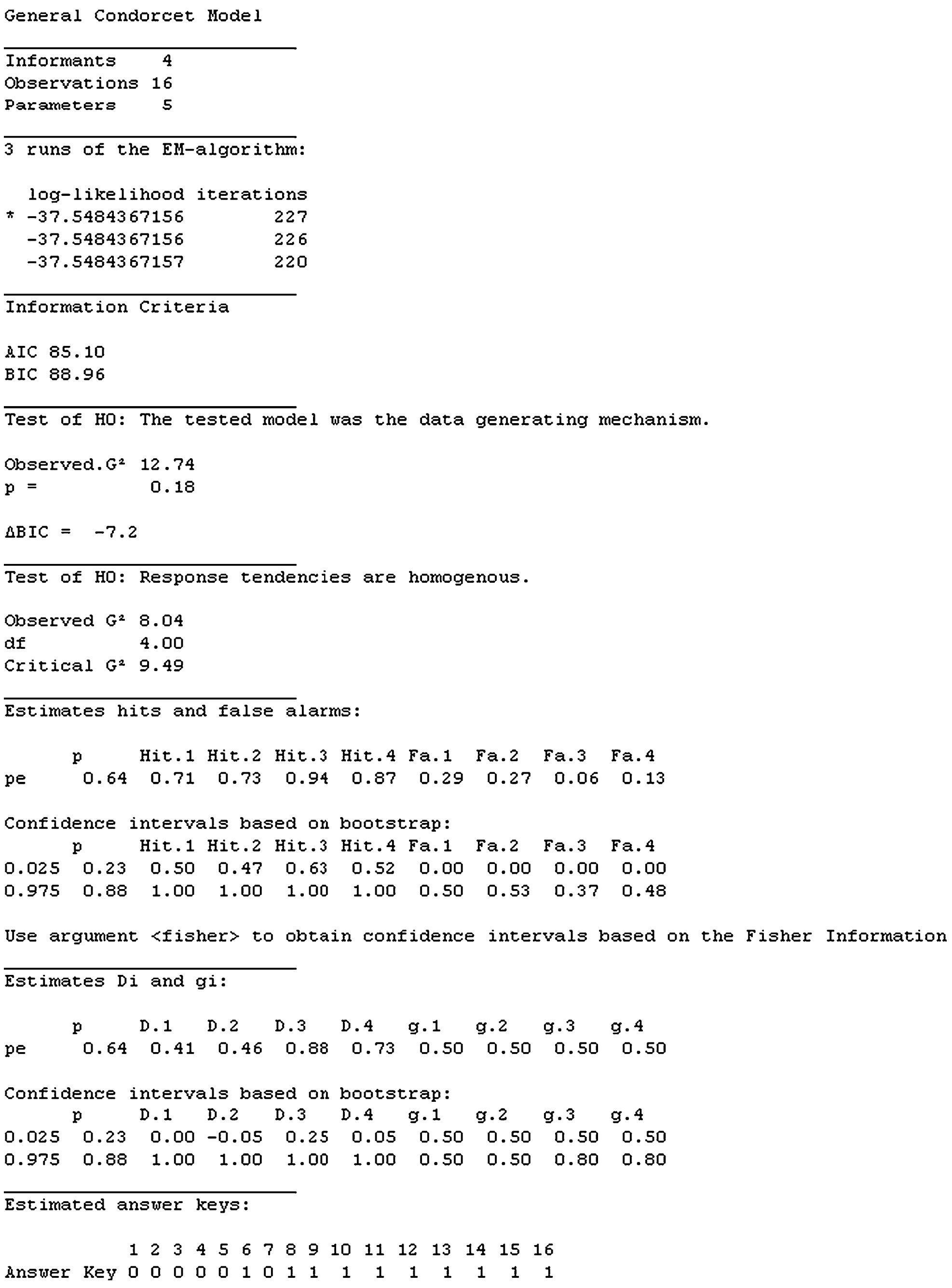
CAML users may choose between different features of the estimation algorithm, such as the criterion of convergence eps and the maximum number of iterations of the EM algorithm max.it. The argument runs allows changing the number of runs of the EM algorithm from different random start values. A lower number of runs decreases computation time but increases the risk that the algorithm may not converge to the global maximum of the likelihood function.

Two arguments define the type of confidence interval reported by CAML. The argument fisher is a Boolean variable. fisher = T provides confidence intervals based on the inverse of the observed Fisher information. As was mentioned earlier, this approach might fail if the Fisher information matrix is singular or if one of the variance estimates is negative. In this case, CAML will provide a warning message, and confidence intervals based on the Fisher information will not be provided. Alternatively, confidence intervals can be obtained by setting the argument boot.runs to a value larger than zero. CAML will conduct as many iterations of the parametric bootstrap procedure as is specified in the boot.runs argument to determine confidence intervals for the parameters. Again, both the precision of the confidence interval estimates (Andrews & Buchinsky, 2000) and the time necessary to complete the analysis increase with the number of iterations. Last but not least, the desired confidence level can be determined using the alpha argument. For example, the default value of alpha = .05 produces $(1 - \alpha) \times 100$ —that is, 95% confidence intervals.

To illustrate the use of CAML, we analyzed the response matrix X shown in Table 1. The output generated by CAML based on the command CAML(X, runs = 3, g = T, boot.runs = 100) is depicted in Fig. 1. The first section of the output lists

Table 2 All arguments for the function CAML, as well as their meaning, admissible values, and default values

Argument	Meaning	Admissible Values	Default Value
X	$N\times M$ response matrix	0, 1	–
g	Should CAML test if response tendencies are homogeneous? (F = False, T = True)	F, T	F
eps	Stopping rule for the EM algorithm	Positive reals	10^{-10}
runs	Number of independent repetitions of the EM algorithm with new starting values	Positive integers	5
max.iter	Maximum amount of iterations for the EM algorithm	Positive integers	1.000
fisher	Should confidence intervals based on the Fisher information matrix be computed? (F = false, T = true)	F, T	F
boot.runs	Number of bootstrap iterations to determine confidence intervals	Nonnegative integers	0
alpha	α for the likelihood ratio tests and the bootstrapped confidence intervals	(0, 1)	.05



```
General Condorcet Model
__________________________
Informants    4
Observations 16
Parameters    5
__________________________
3 runs of the EM-algorithm:

  log-likelihood iterations
* -37.5484367156        227
  -37.5484367156        226
  -37.5484367157        220
__________________________
Information Criteria

AIC 85.10
BIC 88.96
__________________________
Test of H0: The tested model was the data generating mechanism.

Observed.G² 12.74
p =          0.18

ΔBIC =  -7.2
__________________________
Test of H0: Response tendencies are homogenous.

Observed G² 8.04
df          4.00
Critical G² 9.49
__________________________
Estimates hits and false alarms:

      p     Hit.1 Hit.2 Hit.3 Hit.4 Fa.1  Fa.2  Fa.3  Fa.4
pe     0.64  0.71  0.73  0.94  0.87  0.29  0.27  0.06  0.13

Confidence intervals based on bootstrap:
      p     Hit.1 Hit.2 Hit.3 Hit.4 Fa.1  Fa.2  Fa.3  Fa.4
0.025  0.23  0.50  0.47  0.63  0.52  0.00  0.00  0.00  0.00
0.975  0.88  1.00  1.00  1.00  1.00  0.50  0.53  0.37  0.48

Use argument <fisher> to obtain confidence intervals based on the Fisher Information
__________________________
Estimates Di and gi:

      p     D.1   D.2   D.3   D.4   g.1   g.2   g.3   g.4
pe     0.64  0.41  0.46  0.88  0.73  0.50  0.50  0.50  0.50

Confidence intervals based on bootstrap:
      p     D.1   D.2   D.3   D.4   g.1   g.2   g.3   g.4
0.025  0.23  0.00 -0.05  0.25  0.05  0.50  0.50  0.50  0.50
0.975  0.88  1.00  1.00  1.00  1.00  0.50  0.50  0.80  0.80
__________________________
Estimated answer keys:

           1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Answer Key 0 0 0 0 0 0 1 0 1 1  1  1  1  1  1  1
```

Fig. 1 Output produced by CAML for the data matrix presented in Table 1 after typing CAML(X, runs = 3, g = T, boot.runs = 100) in the R console

the number of informants, the number of observations, and the number of model parameters.

Next, the output lists the log-likelihoods for the three runs of the EM algorithm defined by the argument runs = 3, along with the number of iterations required to reach the criterion of convergence. By definition, parameter estimates of the run with the largest log-likelihood (marked * in the output) are the maximum likelihood estimates.

Following the information criteria, the observed likelihood ratio statistic G^2 is listed along with its upper-tail probability under the null hypothesis. Because of the argument (g = T), the output also includes the likelihood ratio difference statistic ΔG^2 along with the critical value with respect to the null hypothesis $g_i = .5$, $i = 1, \ldots, N$.

The parameter estimates of P_Z (denoted p in the output), hits, false alarms, competencies (i.e., D), and response tendencies (i.e., g) are listed below the fit indices. The rows labeled “pe” include point estimates of the parameters. Because the default value of alpha is .05, the rows following the point estimates contain estimates of the .025 and .975 quantiles of the sampling distribution—that is, the 95% confidence interval.

The last part of the output comprises answer key estimates for each of the 16 items. The meanings of “0” and "1" are defined by the user when the dichotomous responses are coded. In the present case, 0 represents "new" and 1 represents "old" (see Table 1).

A validation study of the general Condorcet model

Model parameters should reflect the psychological variables they are supposed to measure. Consequently, the validity of a model hinges on the validity of all its parameters. In the context of MPT models, parameter validity is typically tested by specifying experimental manipulations that should affect specific psychological processes as captured by specific parameters (Batchelder & Riefer, 1999; Erdfelder et al., 2009). If the model is valid, each experimental manipulation should selectively affect the parameter of interest, but not others. This approach has found widespread use in cognitive psychology (Bayen, Murnane, & Erdfelder, 1996; Buchner, Erdfelder, Steffens, & Martensen, 1997; Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995; Chechile & Meyer, 1976; Erdfelder & Buchner, 1998; Klauer, Stahl, & Erdfelder, 2007; Klauer & Wegener, 1998; Nadarevic & Erdfelder, 2011; Steffens, Buchner, Martensen, & Erdfelder, 2000). To our knowledge, however, a test of the GCM's validity has not been attempted so far.

We tested the validity of the 2HT-GCM in a recognition memory experiment. The primary reason for choosing this paradigm is that the answer key (i.e., the old–new-status of each item) is known in this application and can thus be used as an additional criterion of validity. Furthermore, the data structure perfectly matches the requirements of consensus analysis; that is, each of *N* informants responds to a set of *M* dichotomous test items. We predicted that the study time provided for each item in the study phase should affect the competence parameters of the 2HT-GCM selectively: The more time participants have to study the items, the higher their subsequent recognition competence. Similarly, the proportion of old items in the recognition test should affect the response tendency parameters selectively: The higher the proportion of old items in the test, the larger the bias to respond "old." Last but not least, the answer key estimated by the 2HT-GCM should be very similar to the actual answer key underlying the data.

Despite the fact that applications of consensus analysis typically refer to field studies, rather than to controlled laboratory environments (e.g., Romney et al., 1987), the experimental validation study presented here is essential for all applications of consensus analysis. Since the GCM, like any model, includes assumptions, it cannot be taken for granted on a priori grounds that consensus analysis works as intended. Hence, it needs to be tested empirically whether consensus analysis is actually able to uncover answer keys, latent competences, and response tendencies that correspond to the true underlying answer keys, competences, and response tendencies. Only if the validity of the GCM can be established experimentally is its application to field data warranted.

Method

Participants Thirty-two students at the University of Mannheim, with a mean age of 22.25 years (*SD* = 2.64), participated as part of their study requirements. Of the participants, 57.5% were female.

Design We manipulated *study time* for each item (short vs. long) and the *base rate* of old items in the test phase of the experiment (low vs. high) in a between-subjects fashion.

Materials and procedure One hundred fifty German nouns randomly selected from the data base of Hager and Hasselhorn (1994) served as stimulus material. Ten words served as primacy and recency buffers and were not analyzed in the test phase. The remaining words were randomly divided in two lists of 70 words each. Both lists equally often served as study list and distractor list, respectively.

After providing informed consent, each participant was tested individually at a personal computer. First, the participants studied one of the word lists. Each study item appeared at the center of the screen for 4 s (long study condition) or 0.5 s (short study condition), with an interstimulus interval of 0.5 s. Subsequently, participants entered the test phase of the experiment. In the test phase, the proportion of old items was either 30% (i.e., 30 old items and 70 new items) in the low-base-rate condition or 70% (i.e. 70 old items and 30 new items) in the high-base-rate condition. The test items were presented sequentially at the center of the screen. Two response buttons labeled "old" and "new" appeared under each test item. After the participant clicked one of the buttons, the next test item appeared. The order of items in the study phase and test phase was randomized for each participant.

Results and discussion

We first estimated the parameters of the standard 2HT recognition model (Snodgrass & Corwin, 1988). This model makes use of the actual answer key. As can be seen in the upper half of Fig. 2, study time significantly affected competence estimates, $F(1, 28) = 23.48$, $p < .001$, $\eta^2 = .43$. As was expected, the competence estimates in the long study condition were significantly higher than those in the

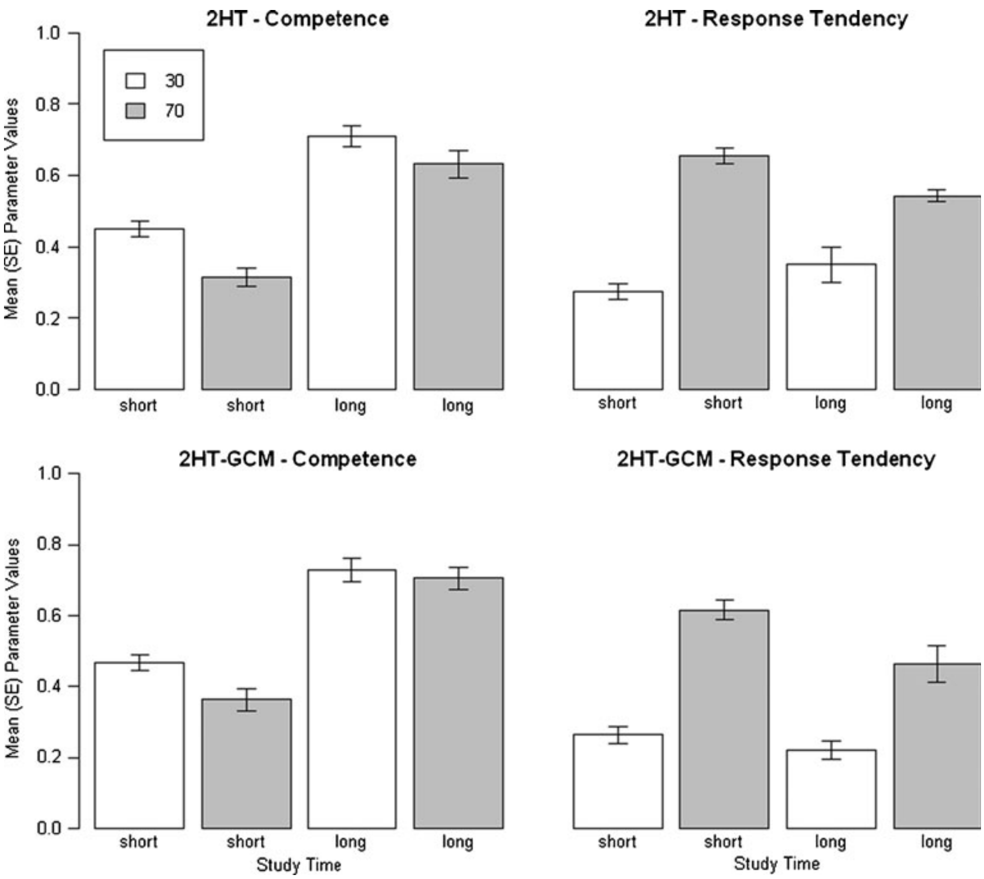


Fig. 2 Mean (*SE*) parameter estimates for the 2-HT model (upper half) and the 2HT-GCM (lower half) in a recognition experiment where the study duration (short vs. long) and the base rate of old items in the test phase (30% vs. 70%) were manipulated

short study condition. In contrast, the study time manipulation did not affect response tendencies, $F(1, 28) < 1$. In addition, the high base rate increased the informants' tendency to respond "old," in comparison with the low base rate, $F(1, 28) = 12.80$, $p = .001$, $\eta^2 = .29$. The base rate manipulation did not affect the competence estimates, $F(1, 28) = 3.14$, $p = .09$, $\eta^2 = .06$. The interaction of study time and base rate was not significant, either for the competence estimates, $F(1, 28) < 1$, or for the response tendency estimates, $F(1, 28) = 3.24$, $p = .09$, $\eta^2 = .07$. To summarize, the manipulations were successful and affected the target parameters of the 2HT recognition model selectively.

In the next step, we analyzed the data of the recognition experiment using CAML. For this purpose, the data had to be separated into four data sets, because the actual answer keys differed between study word lists and the two base rate conditions. Not separating these data would result in a lack of consensus, since informants would not agree on the correct response. We thus analyzed four groups of 8 participants each. Within each group, half of the participants were in the long study duration condition, and the other half in the short study duration condition. Of course, actual answer keys were ignored in the CAML analyses.

All four groups showed consensus, as indicated by positive competence estimates. To evaluate model fit, we generated 10,000 bootstrap samples [i.e., CAML(X, boot.runs = 10000)] for each data set to determine the *p*-values for the observed likelihood ratio statistics under the GCM null hypothesis. The model fitted three of the four data sets nicely ($p > .1$). Given a conventional significance level of .05, the model failed to fit one of the four data sets ($p = .02$). Upon further inspection, we discovered that one of the informants demonstrated perfect memory, resulting in extreme parameter estimates. Upon exclusion of this informant from the analysis, the model fitted the data well ($p = .33$). Thus, the overall model fit can be considered as good.

As can be seen in the lower half of Fig. 2, the effects of the experimental manipulations observed for the 2HT-GCM parameter estimates mirror the results previously reported for the standard 2HT recognition model. As was expected, the longer study duration produced higher competence estimates, as compared with the short study duration condition, $F(1, 28) = 27.09$, $p < .001$, $\eta^2 = .48$, but did not affect response tendencies, $F(1, 28) = 2.13$, $p = .16$, $\eta^2 = .04$. Moreover, informants showed higher tendencies to respond "old" in the condition with a large base rate of old items, as compared with the condition with a low base rate, $F(1, 28) = 19.64$, $p < .001$, $\eta^2 = .39$. In contrast, the base rate manipulation did not affect competence parameter estimates, $F(1, 28) = 1.16$, $p = .29$, $\eta^2 = .02$. Again, neither the interaction with respect to the competence estimates, $F(1, 28) < 1$, nor the interaction with respect to the response tendency

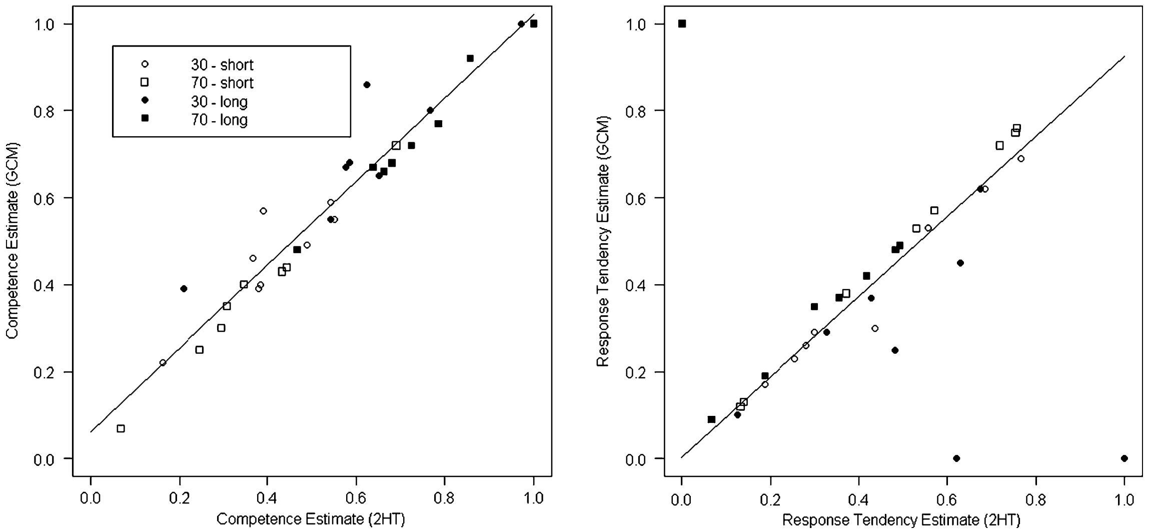


Fig. 3 Scatterplots depicting the relationship between parameter estimates for the 2-HT model and the 2HT-GCM in a recognition experiment where the study duration (short vs. long) and the base rate of old items in the test phase (30% vs. 70%) were manipulated. Competence estimates are depicted on the left, and response tendency estimates on the right. Regression lines exclude boundary estimates

estimates, $F(1, 28) < 1$, reached statistical significance. These results observed at the group means level match those previously reported for the standard 2HT recognition model (based on the actual answer keys) almost perfectly.

How well do the parameter estimates of the 2HT recognition model and the 2HT-GCM agree at the level of individual informants? There is an almost perfect linear relationship between the competence estimates of both models, $r = .96$, $p < .001$, as illustrated in the left side of Fig. 3. This strong correlation does not hold for the response tendency estimates, $r = .33$, $p = .06$. However, closer inspection of the scatterplot (see the right side of Fig. 3) revealed that there is a strong linear relationship between estimates from the two analyses. However, three outliers at the boundary of the parameter space obscure this relationship.[1] After the exclusion of the three boundary parameter estimates in Fig. 3, the correlation between the response tendency estimates of the two model versions is very high, $r = .96$, $p < .001$.

To evaluate the proportion of correctly estimated answer keys, we compared the answer key estimates resulting from CAML with the known true status of each item. In addition, we evaluated the performance of the majority rule. This model-free rule takes the majority responses as an estimate of the answer key, ignoring differences in competence between the informants. On average, CAML estimated 96.25% of the answer key correctly. In contrast, the majority rule was correct in only 92.25% of the cases. The high proportions of correctly estimated answer keys correspond to the results of other studies that also showed high rates of correspondence (Karabatsos & Batchelder, 2003; Romney et al., 1986).

General discussion

The goal of the present work was to introduce a program for consensus analysis based on maximum likelihood estimation and to test the validity of the model underlying consensus analysis. We employed experimental manipulations that selectively affect competence and response tendency parameters. In the 2HT-GCM, these parameters are estimated without any knowledge of the true answer keys. Nevertheless, the parameter estimates obtained reflected these manipulations nicely. Furthermore, the parameter estimates based on consensus analysis almost perfectly mirrored those of the standard 2HT recognition model using the actual answer keys. Finally, the answer key estimates of the GCM corresponded to the true answer keys. In sum, the GCM was able to recover individual competence parameters, as well as the true answer keys, with a very high precision despite the small data set in each condition (8 participants per condition).

Although the approach to consensus analysis presented here overcomes some shortcomings of the factor-analytic procedure, there is one limitation inherent in the CAML approach itself: The present version of CAML cannot

[1] Competence estimates close to 1 render response tendency estimates unreliable. To illustrate this point, consider a case where the hit rate is $H = .99$ and the false alarm rate is $F = .01$. Eqs. 4 and 5 indicate that the corresponding competence would be $D = .98$ and the response tendency $g = .5$. By introducing small estimation errors of .01 to both the hit rate and the false alarm rate, it can be seen that the response tendency estimate is much more affected than the competence estimate. More precisely, the response tendency estimate in this hypothetical case takes on values in [0, 1], whereas the competence estimate is in [.96, 1.00].

account for possible differences in item difficulties. Karabatsos and Batchelder (2003) presented a promising approach to consensus analysis based on the Markov chain Monte Carlo method that yields estimates of the item difficulties in addition to the 2HT-GCM parameters. Currently, we do not see how the estimation of item difficulties can be incorporated into the procedure presented above. However, in a simulation study, we found that the procedure implemented in CAML outperforms the Markov chain Monte Carlo approach with respect to tests of consensus and yields comparable estimates of answer keys and informants' competencies even when difficulties vary between items (Aßfalg & Erdfelder, 2011).

Another possible limitation of CAML is that it is currently available for the R software platform only. Although software written in R was very successful in recent years (e.g., Bulté & Onghena, 2008; Grassie et al., 2010; Nimon et al., 2008), the user interface of R might not appeal to everyone. We intend to address this limitation in future versions of CAML.

In a nutshell, CAML provides a powerful approach to consensus analysis that avoids limitations and simplifying assumptions of other procedures developed for the same purpose. We therefore believe that CAML is the most promising approach to consensus analysis for dichotomous response data currently available.

Author Note We are grateful to Morten Moshagen and Dennis Boywitt for helpful comments on earlier versions of the manuscript.

References

Agresti, A., & Yang, M.-C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis, 5,* 9–21.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language, 53,* 60–80.

Andrews, D. W. K., & Buchinsky, M. (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica, 68,* 23–51.

Aßfalg, A., & Erdfelder, E. (2011). *Consensus analysis: A comparison of methods*. Unpublished manuscript.

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition, 84,* 1–53.

Barg, F. K., Huss-Ashmore, R., Wittink, M. N., Murray, G. F., Bogner, H. R., & Gallo, J. J. (2006). A mixed-methods approach to understanding loneliness and depression in older adults. *Journals of Gerontology B, 61,* 329–339.

Batchelder, W. H., Kumbasar, E., & Boyd, J. P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology, 22,* 29–58.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103–112). Greenwich, CT: JAI Press.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53,* 71–92.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 22,* 197–215.

Borgatti, S. P. (1996). *ANTHROPAC 4.0*. Natick, MA: Analytic Technologies.

Buchner, A., Erdfelder, E., Steffens, M. C., & Martensen, H. (1997). The nature of memory processes underlying recognition judgments in the process dissociation procedure. *Memory & Cognition, 25,* 508–517.

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology. General, 124,* 137–160.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40,* 467–478.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall.

Chechile, R., & Meyer, D. L. (1976). Bayesian procedure for separately estimating storage and retrieval components of forgetting. *Journal of Mathematical Psychology, 13,* 269–295.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28,* 375–389.

Comrey, A. L. (1962). The minimum residual method of factor-analysis. *Psychological Reports, 11,* 15–18.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B, 39,* 1–38.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., & Hinkley, D. V. (1978). Assessing accuracy of maximum likelihood estimator—observed versus expected Fisher information. *Biometrika, 65,* 457–482.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Assfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie–Journal of Psychology 217,* 108–124.

Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 24,* 387–414.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160.

Godoy, R., Eisenberg, D. T. A., Reyes-Garcia, V., Huanca, T., Leonard, W. R., McDade, T. W., et al. (2008). Assortative mating and offspring well-being: Theory and empirical findings from a native Amazonian society in Bolivia. *Evolution and Human Behavior, 29,* 201–210.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215–231.

Grassie, M., Luccio, R., & Di Blas, L. (2010). CircE: An R implementation of Browne's circular stochastic process model. *Behavior Research Methods, 42,* 55–73.

Hager, W., & Hasselhorn, M. (Eds.). (1994). *Handbuch deutschsprachiger Wortnormen*. Göttingen: Hogrefe.

Hu, X. (1999). Multinomial processing tree models: An implementation. *Behavior Research Methods, 31,* 689–695.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59,* 21–47.

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, 31,* 220–234.

Johnson, K. E., Mervis, C. B., & Boster, J. S. (1992). Developmental-changes within the structure of the mammal domain. *Developmental Psychology, 28,* 74–83.

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika, 68,* 373–389.

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology Learning, Memory, and Cognition, 33,* 680–703.

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology, 75,* 1155–1178.

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24,* 492–516.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton Mifflin.

Linzer, D. A., & Lewis, J. (2010). *poLCA: Polytomous variable latent class analysis*. Retrieved March 9, 2011 from http://userwww.service.emory.edu/~dlinzer/poLCA/

Macmillan, N. A., & Creelman, C. (2008). *Detection theory: A user's guide*. New York: Psychology Press.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109,* 235–250.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M. Y., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40,* 230–262.

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., et al. (2006). Folkbiology of freshwater fish. *Cognition, 99,* 237–273.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42,* 42–54.

Nadarevic, L., & Erdfelder, E. (2011). Cognitive processes in implicit attitude tasks: An experimental validation of the trip model. *European Journal of Social Psychology, 41,* 254–268.

Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods, 40,* 457–466.

R Development Core Team. (2011). *The R-project for statistical computing*. Available at http://www.r-project.org/

Read, T. R., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive-processes. *Psychological Review, 95,* 318–339.

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist, 31,* 163–177.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus—a theory of culture and informant accuracy. *American Anthropologist, 88,* 313–338.

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for macintosh computers. *Behavior Research Methods, 31,* 696–700.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 29,* 641–649.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory—applications to dementia and amnesia. *Journal of Experimental Psychology. General, 117,* 34–50.

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class processing tree models. *Behavior Research Methods, 39,* 267–273.

Steffens, M. C., Buchner, A., Martensen, H., & Erdfelder, E. (2000). Further evidence on the similarity of memory processes in the process dissociation procedure and in source monitoring. *Memory & Cognition, 28,* 1152–1164.