Cognition 134 (2015) 21–38

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The construction of categorization judgments: Using subjective confidence and response latency to test a distributed model

Asher Koriat *, Hila Sorka

Department of Psychology, University of Haifa, Israel

ARTICLE INFO

Article history:
Received 29 April 2014
Revised 23 September 2014
Accepted 23 September 2014

Keywords:
Categorization
Subjective confidence
Response latency
Consensus
Self-consistency

ABSTRACT

The classification of objects to natural categories exhibits cross-person consensus and within-person consistency, but also some degree of between-person variability and within-person instability. What is more, the variability in categorization is also not entirely random but discloses systematic patterns. In this study, we applied the Self-Consistency Model (SCM, Koriat, 2012) to category membership decisions, examining the possibility that confidence judgments and decision latency track the stable and variable components of categorization responses. The model assumes that category membership decisions are constructed on the fly depending on a small set of clues that are sampled from a commonly shared population of pertinent clues. The decision and confidence are based on the balance of evidence in favor of a positive or a negative response. The results confirmed several predictions derived from SCM. For each participant, consensual responses to items were more confident than non-consensual responses, and for each item, participants who made the consensual response tended to be more confident than those who made the nonconsensual response. The difference in confidence between consensual and nonconsensual responses increased with the proportion of participants who made the majority response for the item. A similar pattern was observed for response speed. The pattern of results obtained for cross-person consensus was replicated by the results for response consistency when the responses were classified in terms of within-person agreement across repeated presentations. These results accord with the sampling assumption of SCM, that confidence and response speed should be higher when the decision is consistent with what follows from the entire population of clues than when it deviates from it. Results also suggested that the context for classification can bias the sample of clues underlying the decision, and that confidence judgments mirror the effects of context on categorization decisions. The model and results offer a principled account of the stable and variable contributions to categorization behavior within a decision-making framework.

1. Introduction

1.1. Theoretical views

Dividing the world into classes of things is a fundamental cognitive ability that allows people to treat distinct entities as equivalent in some way. The classical view of categorization holds that all instances of a concept share common fundamental features that are individually necessary and jointly sufficient for determining which instances are members of the concept. This view implies that categorization is rule-based. Extensive empirical research, however, has yielded several findings that challenge this view. Specifically, results indicate difficulties in specifying a set of defining attributes for natural concepts (Ashcraft,

* Corresponding author at: Department of Psychology, University of Haifa, Haifa 3498838, Israel. Tel.: +972 4 8249746; fax: +972 4 8249431.
E-mail address: akoriat@research.haifa.ac.il (A. Koriat).

http://dx.doi.org/10.1016/j.cognition.2014.09.009
0010-0277/

1978; Hampton, 2009; Rosch & Mervis, 1975), gradedness in category membership (Barr & Caplan, 1987; Hampton & Gardiner, 1983; McCloskey & Glucksberg, 1978; Oden, 1977; Rosch, 1973; Rosch & Mervis, 1975), cross-person and within-person inconsistency in categorization (Barr & Caplan, 1987; Estes, 2003; Hampton, 1979, 1998, 2009; McCloskey & Glucksberg, 1978), and contextual influences on categorization judgments (Anderson & Ortony, 1975; Barsalou, 1987, 1989; Hampton, 2011; Medin, Lynch, Coley, & Atran, 1997; Roth & Shoben, 1983; see Murphy, 2002). The theories that have been proposed to accommodate these findings, such as the prototype view (Rosch & Mervis, 1975) and the exemplar view (Brooks, 1978; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1988, 1991) assume that categories are defined in terms of family similarity rather than in terms of a set of criterial features. These views embody a probabilistic conception according to which category membership is graded rather than all or none. Hybrid models that include both rule-based and similarity-based categorization have also been proposed (see Smith & Sloman, 1994).

1.2. The present study

The aim of this study is to gain insight into the process underlying human categorization by examining confidence judgments in one's decision about the membership of an object in a certain category, and the time it takes to reach that decision. A model will be proposed, and predictions from the model will be tested. The model is clearly "fuzzy". However, it focuses on the process by which people make category decisions, and this process would seem to entail a variety of cognitive operations that do not follow any simple principle.

Our point of departure is the observation that the classification of objects to categories displays two seemingly inconsistent characteristics. On the one hand, there is a great deal of cross-person consensus and within-person consistency in the assignment of objects to natural categories (McCloskey & Glucksberg, 1978). This observation is, of course, basic to the idea that categorization is rule-based. On the other hand, categorization also exhibits some degree of within-person instability and cross-person variability. This observation, which suggests that the assignment of objects to categories is not clear-cut, has provided the motivation for the probabilistic views of categorization. The theoretical challenge is to offer a principled account for both the stable and variable contributions to categorization behavior.

In fact, the results documenting variability in categorization behavior also exhibit some order in this variability. For example, the typicality results reported by Rosch and Mervis (1975; see also Hampton & Gardiner, 1983) indicate a very reliable ranking of typicality across participants. In addition, participants disclose some awareness of the differences in degree of membership because they can rate the membership of exemplars on a continuous scale (Barr & Caplan, 1987). Typicality also predicts within-person consistency in categorization (Hampton, 1988, 1995), and correlates with cross-participants consensus in categorization decisions (Barr & Caplan, 1987; McCloskey & Glucksberg, 1978). Typicality was also found to predict response time in categorization: Categorization sentences of typical items were verified more quickly than sentences describing less typical items (McCloskey & Glucksberg, 1979; Rips, Shoben, & Smith, 1973; Rosch, 1973). Of particular relevance to the present investigation is the observation that within-person consistency and between-person consensus are correlated: Items that are in disagreement between participants also exhibit inconsistency in categorization across repeated presentations (McCloskey & Glucksberg, 1978). Thus, not only is there some stability in categorization, but the instability observed also follows a relatively stable pattern. Our proposal is that confidence judgments and decision latency can help track the stable and variable components of categorization responses in a manner that provides some information about the process underlying category membership decisions.

1.3. Category membership decisions: a process analysis

The model to be presented below is based on the assumption that category membership decisions are generally constructed on the fly depending on the clues and considerations that are accessible at the time of the judgment (see Barsalou, 1987). A similar assumption underlies the attitude-as-construction view, which assumes that attitudinal judgments are formed on the spot. Therefore, they can vary depending on the person's current goals and mood, and depending on the context in which the judgment is made (Bless, Mackie, & Schwarz, 1992; Schwarz, 2007, 2008; Schwarz & Strack, 1991; Tourangeau, 1992). A similar view has been proposed with regard to personal preferences (Lichtenstein & Slovic, 2006; Slovic, 1995): Preferences are generally constructed in the process of elicitation rather than retrieved ready-made from memory. This view was motivated by observations indicating that preferences can vary with the task, the context, and the goals of the respondents (see Bettman, Luce, & Payne, 1998; Shafir, Simonson, & Tversky, 1993; Warren, McGraw, & Van Boven, 2011).

We propose that category membership decisions are also constructed on the spot. Assume that a person is asked to decide whether a particular object belongs to a particular category, for example, whether olives belong to the fruit category. How does one choose between *yes* and *no*? Introspection and an informal think-aloud study suggest that a variety of clues and considerations come to mind in an associative manner. One might visualize a small olive and compare it to an apple, feeling uneasy to reach a *yes* decision. But then may think of other fruits such as a prune, even a green prune. One might try to recall whether olives are sold in the fruit section of a supermarket, or else one might think about the context in which olives are served or consumed (e.g., not in a fruit salad). Each such clue may tip the balance in favor of *yes* or *no* response. Some of the clues and considerations may involve similarity to a prototype, as postulated by Rosch and her associates, whereas others may concern the deep "essence" of a fruit (Medin, 1989). Others still may involve semantic or episodic associations that are irrelevant to the decision but can still bias the decision in one direction or the other.

To obtain further insight into the underlying process, we presented nine English-speaking participants with five object-category pairs. For each pair, they were asked to list the considerations that come to mind when having to decide whether the object belongs or does not belong to the category. They wrote down each consideration and indicated whether it supported a *yes* or a *no* judgment. To illustrate, for the question whether egg belongs to the animal category, participants listed such positive considerations as "eggs can be eaten just like animals"; "eggs need a mother's warmth in order to hatch, like a child"; "eggs come from chicken"; "eggs are the reproductive phase of some animals"; "eggs contain D.N.A. and therefore become an animal"; and "strict vegans will not eat it". Among the negative considerations, participants listed the following: "before eggs are hatched, they do not have organs"; "eggs cannot breathe"; "it is the beginning stages of life and can grow to become an animal"; "no – if the egg had not been fertilized"; "in Jewish kashrut - eggs are parve [not meat and not milk]"; "some plants have egg-like reproductive components, and they are not animals"; and "it does not have the ability to survive outside of the shell".

In general, unlike us researchers, who strive to provide a principled account of categorization, lay people who are asked to make a categorization decision are not bound by any general principle, and may navigate through diverse types of clues that come to mind. For example, when participants in the study of Rosch and Mervis (1975) were asked to list the attributes possessed by different items, they were instructed specifically: "But try not to *just* free associate – for example, if bicycles just happen to remind you of your father, *don't* write down *father*" (p. 578). However, in a natural situation, there is no reason why thinking about a particular object may not bring to mind some accidental details. These details may not be articulate or conscious, and may be logically irrelevant to the decision, but they may nevertheless influence that decision.

This portrayal of the categorization process implies a distributed model in which people sample clues from a rich network of clues that can be activated by the object-category pair at the time of making a categorization decision. The deliberation can be assumed to continue until the person feels that the "balance of evidence" (Vickers, 2001) favors one option rather than the other. If presented with the same question again after a few days, a new set of clues may come to mind that may lead to a different decision than the one reached on the first occasion.

How can this fuzzy distributed "model" produce testable predictions? We propose that the confidence with which a decision is reached, and the speed with which it is reached can help trace the sources of stability and variability in category membership decisions. In what follows, we describe the Self-Consistency Model (SCM) of subjective confidence, and its predictions for category membership decisions.

1.4. The Self-Consistency Model (SCM) of subjective confidence

SCM was originally developed to explain the *accuracy* of confidence judgments for two-alternative forced-choice (2AFC) general-information questions and perceptual judgments (Koriat, 2008, 2011), which yield typically a moderate-to-high confidence-accuracy (C/A) correlation. However, in several studies that included a sufficiently large number of consensually-wrong (CW) items, for which most participants choose the wrong answer, the C/A correlation was positive only for consensually-correct (CC) items for which most participants chose the correct answer. For CW items, in contrast, the correlation was consistently *negative*. This pattern was obtained across several studies that used a variety of tasks (Brewer & Sampaio, 2006, 2012; DeSoto & Roediger, 2014; Koriat, 1976, 2008, 2011, 2013). The results suggested that confidence judgments are correlated with the consensuality of the answer rather than with its accuracy, and shifted investigation to the *basis* of confidence judgments. Thus, SCM has been extended to tasks for which the answer does not have a truth-value such as social attitudes (Koriat & Adiv, 2011), social beliefs (Koriat & Adiv, 2012), and personal preferences (Koriat, 2013). In this study, we examine the possibility that the conceptual framework underlying SCM can also apply to category membership decisions.

In SCM, people's confidence judgments were modeled after the classical procedures of calculating statistical level of confidence when conclusions about a population are based on a sample of observations drawn from that population. It was proposed that when presented with a 2AFC item, it is by replicating the choice process several times that a person can appreciate the degree of doubt or certainty involved. Confidence is based on the consistency with which different replications agree in favoring a particular decision. Subjective confidence represents essentially an assessment of *reproducibility* – the likelihood that a new replication of the decision process will yield the same choice. Thus, reliability is used as a cue for validity. This is the logic underlying statistical inference when conclusions about a population are to be based on a sample of observations drawn from that population (Koriat, 2012).

Thus, SCM incorporates a sampling assumption that is common in many decision models (e.g., Juslin & Olsson, 1997; Stewart, 2009; Stewart, Chater, & Brown, 2006; Vickers & Pietsch, 2001; Vul, Goodman, Griffiths, & Tenenbaum, 2009). We propose that a similar sampling model applies to category membership decisions. Assume that each 2AFC object-category item is associated with a population of representations that can be potentially activated. In deciding whether an object is a member of a category, participants are assumed to sample a number of representations sequentially from memory, draw the implications of each representation for the decision, and reach a *yes*/*no* decision on the basis of the balance of evidence in favor of the two options (Vickers, 2001; see Baranski & Petrusic, 1998). The term *representation* refers to any clue or association that can tilt the pendulum in the direction of one decision or the other. The sample drawn in each occasion is assumed to be small because of the cognitive difficulty in integrating information across different clues to reach a final decision. Confidence in the decision reached is assumed to depend on self-consistency – the overall agreement across the sampled representations

in favoring the chosen option (see Alba & Marmorstein, 1987; Ross, Buehler, & Karr, 1998).

A critical assumption of SCM is that the population of clues associated with each object-category item is largely shared across all participants with the same background. Of course, participants may nevertheless reach different decisions. In the case of general-information questions, proponents of the ecological approach to cognition (Dhami, Hertwig, & Hoffrage, 2004; Gigerenzer, 2008) have stressed the idea that people's knowledge is not only shared, but is generally accurate by virtue of people's adaptation to the environment. A similar idea underlies the studies on the *wisdom of crowds*: Information that is aggregated across participants may be closer to the truth than the information provided by each participant (Galton, 1907; Mozer, Pashler, & Homaei, 2008). Thus, we assume that the ingredients that participants use in constructing their category membership decisions are drawn from a "collective wisdom". Because SCM has been described in detail elsewhere (Koriat, 2011, 2012), here we will present only a specific instantiation of the model. This instantiation is clearly over-simple, but is sufficient for bringing to the fore some of the predictions.

1.5. Predictions from a specific version of SCM

Assume that when presented with a category membership question, participants draw randomly 7 representations, each of which yields a binary subdecision favoring a *yes* or a *no* response. The ultimate, overt decision is based on the majority vote across the subdecisions. However, if 3 successively retrieved representations yield the same subdecision, the sampling is terminated, and that subdecision determines the choice (see Audley, 1960). Confidence is based on the degree of consistency among the subdecisions.

Assuming that participants sample their representations from the same item-specific population, the most important property of that population is p_{maj} – the probability of drawing a representation that supports the majority choice. To derive predictions about confidence and response latency, we ran a simulation experiment (see Koriat, 2012; Koriat & Adiv, 2011) which assumed 9 binary populations that differ in p_{maj}, with p_{maj} varying from .55 to .95, at .05 steps. For each population, 90,000 iterations were run, in each of which a sample of (3–7) representations was drawn. The ultimate choice that was based on the sample was classified as "majority" when it corresponded to the majority value in the population, and as "minority" when it corresponded to the minority value in the population.

Subjective confidence in the decision was assumed to depend on self-consistency, which is inversely related to the sample standard deviation. A self-consistency index was used, which was defined as $1 - \sqrt{\hat{p}\hat{q}}$ (range .5–1.0), when p and q designate the proportion of representations favoring the two choices, respectively. This index was calculated for each iteration over the actual number of representations sampled.

Based on the simulation results, Fig. 1A presents the self-consistency index for majority and minority choices

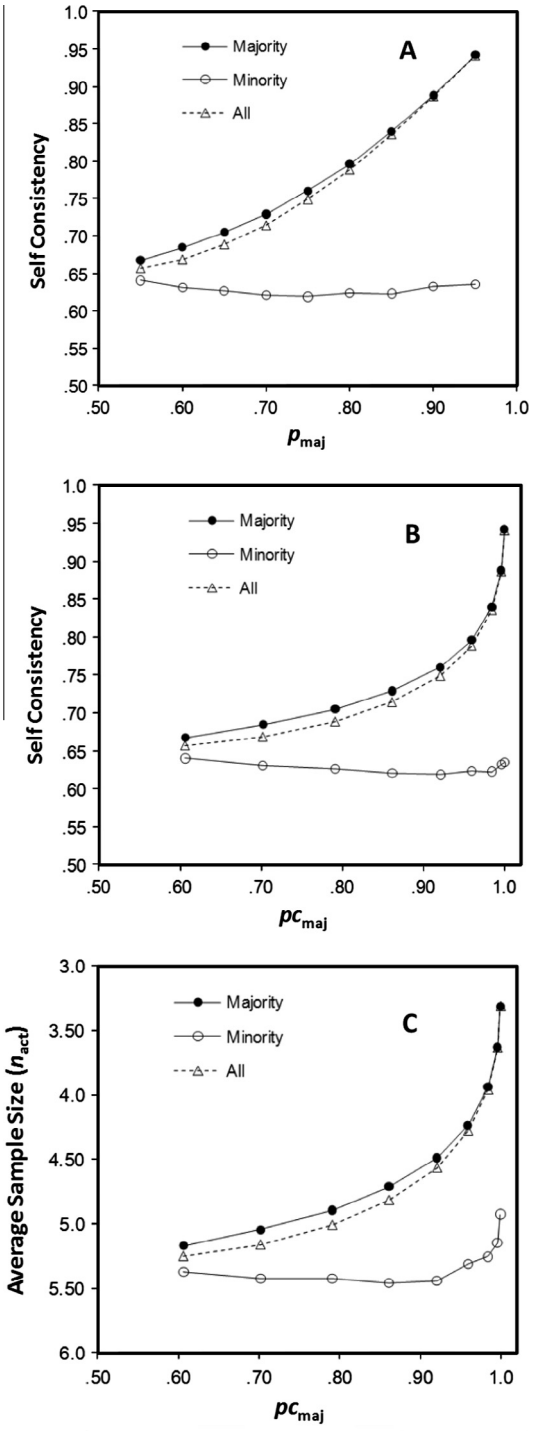

Fig. 1. Panel A: Self-consistency scores as a function of the probability of drawing a majority representation (p_{maj}). Panel B: Self-consistency scores as a function of the probability choosing the majority option (pc_{maj}). Panel C: Average sample size (n_{act}) as a function of the probability of choosing the majority option (pc_{maj}). The functions are based on the results of the simulation experiment (see text).

and for all choices combined as a function of p_{maj}. It can be seen that self-consistency increases monotonically with p_{maj}. More important, self-consistency is higher for

majority choices than for minority choices. The reason for this difference is that as long as p_{maj} exceeds .50, majority choices will be supported by a larger proportion of the sampled representations than minority choices. For example, for p_{maj} = .70, and sample size = 7, the likelihood that 6 or 7 representations will favor the majority choice is .329, whereas only in .004 of the samples will 6 or 7 representations favor the minority choice. Thus, the expectation is that confidence should be higher for majority choices than for minority choices.

Of course, p_{maj} for a particular item is not known. However, it can be estimated from pc_{maj} – the probability with which the majority alternative is chosen across individuals (consensus) or within-individuals across repeated presentations (consistency). The values of pc_{maj} can be derived from the simulation just described and can be substituted for the corresponding p_{maj} values to yield the results in Fig. 1B. These results specify the predicted pattern. For each object-category item, confidence should be higher for participants who choose the consensual response than for those who choose the nonconsensual response, and the difference in confidence between consensual and nonconsensual responses should increase with increasing item consensus – the proportion (or percentage) of participants who make the consensual response for that item. Similarly, if the same item is presented several times, and pc_{maj} is estimated from the proportion of times that each participant makes his or her more frequent choice, the frequent choice should be endorsed with higher confidence than the rare choice, with the difference increasing with item consistency – the proportion of times that the frequent choice is made across repetitions.

A very similar pattern is expected for response speed, assuming that response speed is an inverse monotonic function of n_{act} – the actual number of representations sampled. This number (between 3 and 7) can be derived from the simulation described above. It can be seen in Fig. 1C that response speed should be faster for consensual than for nonconsensual responses, with the difference between them increasing with majority size.

In sum, two sets of predictions follow from SCM. The first concerns inter-item differences (see "All" in Fig. 1B and C): Confidence and response speed should increase with item consensus, and for repeated presentations, they should both increase with increasing item consistency. These predictions are consistent with previous results (Estes, 2004; McCloskey & Glucksberg, 1979). In fact, much of the theorizing on categorization judgments has concerned inter-item differences in such measures as the probability of a positive category judgment, typicality ratings, and degree of stability over time. In SCM, item consensus and item consistency are assumed to reflect the polarity of the population of representations associated with each item, and this polarity is assumed to constrain the variability that can be observed in category membership decisions for each item.

The second set, in contrast, is unique to SCM. It involves systematic differences between different *choices*: When variability in the response choice is observed, confidence and response speed should differ depending on which of the two response alternatives is chosen. Confidence and response speed should be higher for consensual than for nonconsensual decisions, with the difference increasing with increased item consensus. Similarly, in a repeated presentation design, confidence and response speed should be higher for the more frequent response than for the less frequent response, with the difference increasing with increased item consistency. It is the focus on the specific response made on a specific occasion, and on systematic differences between different responses, that makes SCM a *process* model. It should be stressed that these predictions are based on the assumption that the same process underlies consensual/frequent decisions and nonconsensual/rare decisions: In each case, each participant chooses the response that is favored by the *majority* of representations in the sample of representations that he/she has retrieved.

These predictions were tested in Experiments 1 and 2. Experiment 1 used a paper-and-pencil task. Participants were presented with a list of 102 object-category pairs. They decided whether the object is or is not a member of the category, and indicated their confidence in their decision. Experiment 2 was a computerized experiment that permitted testing the predictions about confidence as well as response latency. In addition, the categorization task, which included 100 items, was administered 7 times over two sessions that took place on two separate days. Whereas the results from the first presentation permitted a test of the predictions regarding the effects of cross-person consensus on confidence and response latency, the results across the 7 presentations permitted a test of the predictions regarding within-person consistency. Experiment 3, in turn, explored contextual effects on category membership decisions. Finally, Experiment 4 focused on typicality ratings. It examined how the category membership decisions obtained in Experiments 1 and 2, and the confidence and latency associated with them, relate to typicality ratings.

2. Experiment 1

2.1. Method

2.1.1. Participants and stimulus materials

Twenty-one students, native English speakers, enrolled at the International school of the University of Haifa (12 males), volunteered to participate in the study. The experimental materials consisted of 102 candidate exemplars divided into nine categories. Exemplars and categories were chosen from Barr and Caplan (1987) and McCloskey and Glucksberg (1978) to represent a wide range of typicality and membership ratings.

2.1.2. Procedure

All instructions and materials were compiled in a ten-page booklet. The first page included the instructions, and each of the last nine pages included a list of candidate exemplars of one category, arranged alphabetically. The candidate exemplar and the category name were presented as a pair: The candidate exemplar appeared on the left-hand side, printed in lower-case letters and underlined, and the category name appeared on the right-hand side, printed in upper-case letters (e.g., apple – FRUIT).

The instructions were based on those of McCloskey and Glucksberg (1978). Participants were told: "Many nouns belong to semantic categories. For example, black is a member of the semantic category COLOR, fear is a member of the semantic category EMOTION, and so on". For each noun – CATEGORY pair, they were asked to decide whether or not the noun represents a member of the CATEGORY by circling *yes* or *No*. Participants were instructed not to skip any pair even if the decision is difficult, but to circle the "U" next to each pair when they are unfamiliar with the meaning of the word. Participants were further instructed to rate on a 0–100 scale how confident they were in their decision (0 – very unsure; 100 – very sure). They were encouraged to use the full range of the scale.

2.2. Results and discussion

One item (*Lamprey* – FISH) was eliminated from the analyses because 11 participants circled U for this item. There were 22 additional cases in which the candidate exemplar was marked as unfamiliar (7 of those for *Rutabaga*), and these cases were deleted from the analyses. For each of the 101 items, the response made more often was defined *ad hoc* as the consensual response for that item (see Table 1 in supplementary material). Item consensus averaged 80.76% across items (range 52.38–100%). For 27 items, all participants gave the same response (full-consensus items), and for the remaining 74 items, participants exhibited some disagreement (partial-consensus items). Across all participants and items, 61.86% of the responses were positive, 37.10% were negative, and 1.04% were marked as unfamiliar.

Let us examine the results for confidence beginning with a comparison between the full consensus and partial consensus items. According to SCM, p_{maj} – the probability of drawing a representation that supports the consensual choice – should be close to 1 for full-consensus items, and therefore the response to these items should be endorsed with very high confidence. Indeed, confidence for these items averaged 97.67 ($SD = 2.39$) and was significantly higher than that for the partial consensus items ($M = 87.34$, $SD = 5.24$), $t(94) = 13.54$, $p < .0001$, Cohen's $d = 2.24$.

We examine next the relationship between confidence and consensuality. All items were divided into 6 item consensus categories (51–59%, 60–69%, 70–79%, 80–89%, 90–99%, 100%). Fig. 2 presents mean confidence judgments for each of these categories for consensual and nonconsensual responses. The figure includes also the results for the full consensus items. Mean confidence judgments increased monotonically with item consensus: When mean confidence and mean item consensus were calculated for each item, the correlation between them over the 101 items was .72, $p < .0001$. This correlation is consistent with the idea that confidence increases with the polarity of the population of representations associated with an item. That polarity is assumed to constrain the variability that can be observed in category membership decisions.

Some clues to that variability are provided by the comparison between consensual and nonconsensual responses. Across the 74 items, mean confidence was significantly

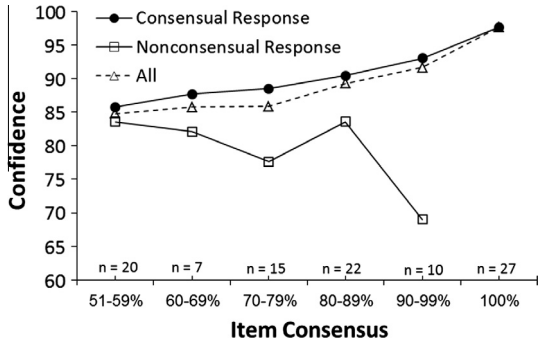

Fig. 2. Mean confidence judgments for consensual and nonconsensual responses and for all responses combined (All) as a function of item consensus – the percentage of participants who made the consensual response (Experiment 1).

higher for consensual responses ($M = 88.77$, $SD = 5.34$) than for nonconsensual responses ($M = 81.95$, $SD = 10.22$), $t(73) = 5.13$, $p < .0001$, $d = 0.84$. Furthermore, as expected, the difference increased with item consensus (see Fig. 1B). The statistical significance of this increase could not be tested on the results presented in the figure because each of the means was based on a different combination of participants. However, we calculated for each participant the functions depicted in Fig. 2 relating mean confidence in consensual and nonconsensual responses to grouped item consensus categories. The rank order correlation between the ordinal value of the item consensus category (1–5) and the difference in mean confidence between consensual and nonconsensual responses (using for each participant the observations for which this difference was computable) averaged .41 across participants, $p < .005$. This correlation was positive for 17 of the 20 participants (one had a tie), $p < .005$, by a binomial test.

In comparing confidence for consensual and nonconsensual responses, the means for each item were based on different participants. Therefore, the differences observed could reflect a between-individual effect: Participants who choose consensual responses tend to be more confident. Consistent with previous findings (see Kleitman & Stankov, 2001; Stankov & Crawford, 1997), there were marked and reliable individual differences in the tendency to make relatively high or relatively low confidence judgments. To control for these differences, the confidence judgments of each participant were standardized so that the mean and standard deviation of each participant were set as those of the raw scores across all participants. Average scores were then calculated for each item for consensual and nonconsensual responses. The results for the standardized scores were essentially the same as those for the raw scores, suggesting that the consensual–nonconsensual differences depicted in Fig. 2 are not due to chronic between-participant differences in confidence.

3. Experiment 2

Experiment 2 was a computerized experiment that allowed the testing of predictions about response latency

as well as those about confidence. In addition, the categorization task was repeated 7 times, allowing examination of the predictions regarding both cross-person consensus and within-person consistency.

3.1. Method

3.1.1. Participants

Thirty-three native English speaking undergraduates at the University of Haifa (21 females) participated in the experiment for payment.

3.1.2. Stimulus materials

The experimental materials consisted of 100 pairs of object-category pairs divided into 10 categories, five natural and five artifactual (see Estes, 2003), with 10 candidate exemplars for each category. Four candidate exemplars that were marked as unfamiliar in Experiment 1 were replaced; the WEAPON category was replaced with FLOWER, and the ANIMAL category was added. Two considerations guided the final selection of pairs, based on the results of Experiment 1 as well as those of Barr and Caplan (1987) and McCloskey and Glucksberg (1978). First, we attempted to include items for which there was some variability in category membership decisions across participants. Second, we tried to include a sufficient number of object-category pairs for which there was some degree of within-person fluctuation in category membership decisions across repeated presentations. Based on these considerations, 10 categories were selected, with a sufficient number of items that fulfill one or both of the criteria just mentioned. For each category, 10 candidate exemplars were selected to yield a wide range of membership ratings and typicality ratings.

Several tasks were used as fillers between different administrations of the categorization task. They were selected because they were judged to yield measures of individual differences that might be correlated with different parameters of the categorization task.[1] They included an automated version of the operation span task (Automated OSPAN; Unsworth, Heitz, Schrock, & Engle, 2005), the Category Width Scale (Pettigrew, 1958), the Need for Closure Scale (NFCS; Webster & Kruglanski, 1994) and the Rational-Experiential Inventory (REI; Pacini & Epstein, 1999).

3.1.3. Apparatus and procedure

The 7 administrations of the categorization task were divided between two sessions that took place on two separate days with a 1-week interval between them. The first session included four blocks in each of which the set of 100 pairs was presented, and the second included the remaining three blocks. The experiment was conducted individually on an IBM-compatible personal computer.

The instructions were the same as in Experiment 1 except that the option to mark an item as Unfamiliar was deleted. Participants initiated each trial by clicking a "start" box. A pair of two words then appeared; the left-hand word was a candidate exemplar, printed in lower-case letters, and the right-hand member was a category name, printed in upper-case letters. Participants clicked *yes* or *no* with the mouse and then clicked a "confirm" box, after which they could not change their response. Response latency was measured, defined as the interval between the presentation of the pair and the "confirm" box press. A confidence scale (0–100) was then added below, and participants marked their confidence by sliding a pointer on a slider using the mouse (a number in the range 0–100 corresponding to the location of the pointer on the slider was shown in a box). Participants were encouraged to use the full range of the confidence scale. The order of the pairs was determined randomly for each participant and block. In addition, each block was preceded by two warm-up items.

3.2. Results and discussion

The organization of the Results section will be as follows. We first test the predictions regarding cross-person consensus using only the results of Block 1. The results for confidence will be examined first, followed by those for response latency. We turn then to the predictions about within-person consistency, focusing again on confidence first and then on response latency. These predictions will be tested by examining the results across all 7 presentations. Finally, we examine the relationship between cross-person consensus and within-person consistency, and the validity of confidence judgments in predicting cross-person consensus and within-person consistency.

3.2.1. Confidence and latency as related to cross-person response consensus

3.2.1.1. Confidence as a function of cross-person consensus. The same analyses as those used in Experiment 1 were applied to the results of Block 1. For each of the 100 items, we first determined the consensual response, and calculated item consensus – the percentage of participants who made that response. Table 2 in supplementary material lists the consensual choice and item consensus for each item. In addition, the table presents mean confidence and response latency for the consensual and nonconsensual choices. Item consensus averaged 78.79% across items (range 51.52–100%). For 15 items all participants gave the same response (for 11 items the response was positive). Across all participants and items, 57.67% of the responses were positive and 42.33% were negative.

Fig. 3 presents mean confidence for consensual and nonconsensual responses for each of the 6 item consensus categories. The results are consistent with those of Experiment 1. First, confidence was significantly higher for the full consensus items ($M = 95.99$, $SD = 4.67$) than for the partial consensus items ($M = 82.89$, $SD = 9.57$), $t(98) = 5.18$, $p < .0001$, $d = 1.46$. Second, mean confidence judgments increased with item consensus. When mean confidence and mean item consensus were calculated for each item, the correlation between them across items was .62, $p < .0001$.

Third, across the 85 partial consensus items, confidence was higher for the consensual response ($M = 84.61$, $SD = 9.81$) than for the nonconsensual response ($M = 75.27$, $SD = 16.30$), $t(84) = 5.30$, $p < .0001$, $d = 0.70$.

[1] The results on individual differences will not be reported in this article.

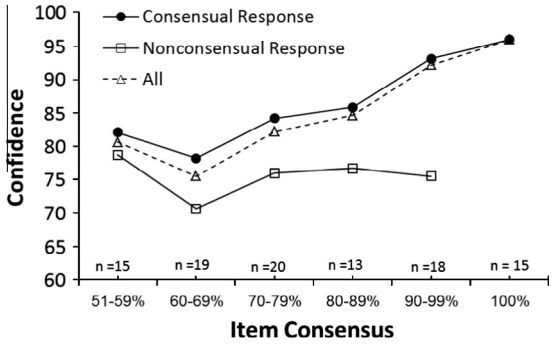

Fig. 3. Mean confidence judgments in Block 1 for consensual and nonconsensual responses and for all responses combined as a function of item consensus – the percentage of participants who made the consensual response (Experiment 2).

This difference was consistent across participants and items: Of the 33 participants, 31 participants evidenced this pattern, $p < .0001$, by a binomial test. In addition, for 65 items, the participants who made the consensual response exhibited higher confidence than those who made the nonconsensual response, in comparison with 20 items in which the pattern was reversed, $p < .0001$, by a binomial test.

Finally, the same analysis as in Experiment 1 confirmed that the difference in confidence between consensual and nonconsensual responses increased with item consensus: The rank order correlation between the ordinal value of the item consensus category (1–5) and the difference in mean confidence between consensual and nonconsensual responses averaged .35 across participants, $p < .0005$. This correlation was positive for 24 of the 33 participants, $p < .01$, by a binomial test.

When confidence judgments were first standardized to control for individual differences in confidence the results were very similar to those presented in Fig. 3.

On the whole, the results are in line with the idea that categorization decisions are based on the sampling of representations from a population of representations that is largely commonly-shared, and that when a participant draws a sample whose polarity deviates from that implied by the population, his/her confidence should be relatively low. Note that the consensual-nonconsensual difference in confidence was relatively consistent across items so that for each item, those individuals who made the consensual choice tended to express greater confidence than those who made the nonconsensual choice.

3.2.1.2. Response latency as a function of cross-person consensus. In the analyses of response latency, latencies below or above 2.5 SDs from each participant's mean in each block were eliminated (3.00% for Block 1 and 3.13% across all 7 blocks). Consistent with previous results (Kelley & Lindsay, 1993; Koriat, 2012; Koriat, Ma'ayan & Nussinson, 2006; Robinson, Johnson, & Herndon, 1997) there was an inverse relationship between confidence and response latency. Thus, focusing on the results of Block 1, when the items were divided for each participant at the median of response latency, confidence judgments were significantly higher for below-median latencies ($M = 92.09$, $SD = 5.81$) than for above-median latencies ($M = 77.27$, $SD = 11.46$), $t(32) = 11.70$, $p < .0001$, $d = 2.93$. The within-individual Pearson correlation between confidence and latency across all items averaged −.44, $p < .0001$, across participants.

As noted earlier, we assume that response latency is also diagnostic of self-consistency. Similar analyses to those of confidence were applied to response latency. The results (for Block 1) are presented in Fig. 4. The pattern mimics largely the one obtained for confidence. First, response latency was shorter for the full consensus items ($M = 3.12$ s, $SD = 0.71$ s) than for the partial consensus items ($M = 4.47$ s, $SD = 0.87$ s), $t(98) = 5.67$, $p < .0001$, $d = 1.61$. Mean latency decreased with item consensus: The correlation between latency and consensus was −.65 across the 100 items, $p < .0001$.

However, across the partial-consensus items, response latency was significantly shorter for consensual responses ($M = 4.34$ s, $SD = 0.94$ s), than for nonconsensual responses ($M = 5.22$ s, $SD = 1.53$ s) $t(84) = 4.53$, $p < .0001$, $d = 0.70$. Of the 33 participants, 31 exhibited this trend, $p < .0001$, by a binomial test. The effect was also consistent across items. For 56 items, participants who chose the consensual option responded faster than those who chose the nonconsensual option, $p < .005$, by a binomial test.

The correlation for each participant between the mean of the item consensus category and the difference in latency between the means of the consensual and nonconsensual responses averaged .39 across participants, $p < .0005$. This correlation was positive for 25 of the 32 participants (one had a tie), $p < .005$, by a binomial test. Thus, the consensual–nonconsensual difference in response speed increased with item consensus, as predicted (see Fig. 1C). When response latencies were first standardized to control for individual differences in response speed, the results were very similar to those presented in Fig. 4.

In sum, the results for response speed exhibit the same pattern as that observed for confidence. Response speed increased with degree of between-person consensus, but was significantly faster for consensual than for

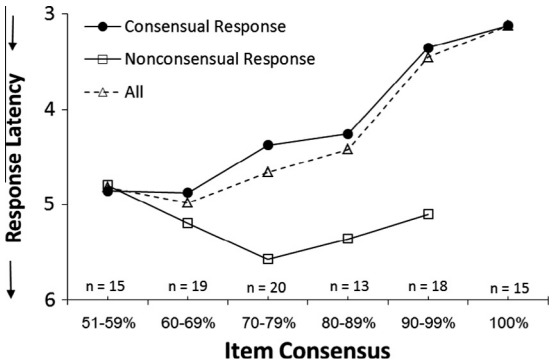

Fig. 4. Mean response latency in Block 1 for consensual and nonconsensual responses and for all responses combined as a function of item consensus – the percentage of participants who made the consensual response (Experiment 2).

nonconsensual responses, with the difference increasing with between-participant consensus.

3.2.2. Confidence and latency as related to within-person response consistency

In this section, we test the predictions of SCM for the within-person consistency. Participants tended to give the same response consistently throughout the 7 presentations. Thus, the likelihood of choosing the Block-1 response over the next six blocks averaged .89 across all participants. However, we expect systematic differences between items and responses as a function of within-person consistency.

3.2.2.1. Confidence as a function of response consistency. All items were classified for each participant into those that elicited the same response across all blocks (full consistency) and those exhibiting some degree of inconsistency (partial consistency). Confidence was significantly higher for the full-consistency items ($M = 89.98$, $SD = 6.64$) than for the partial-consistency items ($M = 72.01$, $SD = 15.89$), $t(32) = 10.08$, $p < .0001$, $d = 1.50$. All 33 participants exhibited this pattern, $p < .0001$, by a binomial test.

We next compare confidence for the participant's frequent and rare responses as a function of item consistency – the number of times that the frequent response was chosen. Fig. 5 presents the pertinent results and also includes the mean of the full-consistency items. As expected, confidence increased monotonically with item consistency (4–7): The correlation between mean confidence and item consistency averaged .34, $p < .0001$, across all items. However, across the partial-consistency items, confidence was significantly higher for the participant's frequent responses ($M = 73.92$, $SD = 16.05$) than for the rare responses ($M = 62.70$, $SD = 17.14$), $t(32) = 7.37$, $p < .0001$, $d = 0.69$. Of the 33 participants, 30 exhibited this pattern, $p < .0001$, by a binomial test. Thus, participants were less confident when their response deviated from *their own* modal response. The results also indicated that the frequent-rare difference in confidence increased with item consistency: Mean confidence difference was calculated for each participant between frequent and rare responses for each of the item-consistency categories. The correlation between this

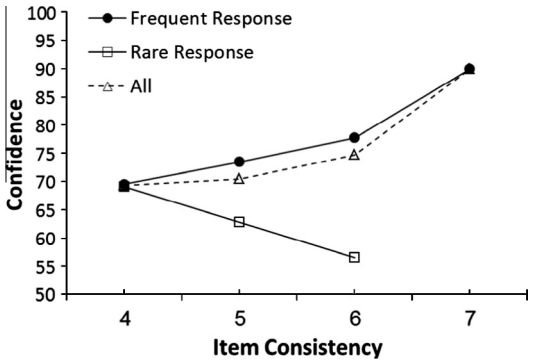

Fig. 5. Mean confidence judgments for the frequent and rare responses and for all responses combined as a function of item consistency – the number of times that the response was made across the 7 blocks (Experiment 2).

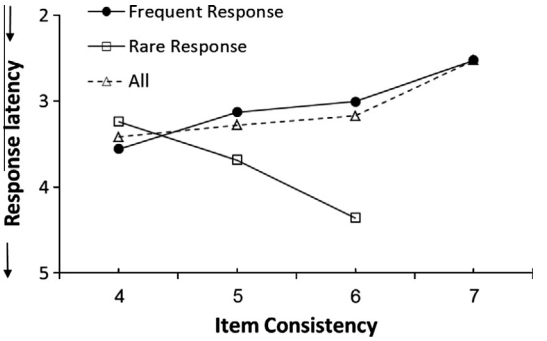

Fig. 6. Mean response latency for the frequent and rare responses and for all responses combined as a function of item consistency (Experiment 2).

difference and item consistency, for 31 participants with complete data, averaged .72 ($p < .0001$) across participants. It was positive for 30 participants, $p < .0001$, by a binomial test.

3.2.2.2. Response latency as a function of response consistency. Similar analyses were conducted on response latency, and the results, presented in Fig. 6, mimic roughly those obtained for confidence. Response latency was significantly shorter for the full-consistency items ($M = 2.52$ s, $SD = 0.40$ s) than for the partial-consistency items ($M = 3.27$ s, $SD = 0.67$ s), $t(32) = 12.52$, $p < .0001$, $d = 1.38$. All 33 participants exhibited this pattern, $p < .0001$, by a binomial test. In addition, for the partial-consistency items, response latency was shorter for the frequent response ($M = 3.22$ s, $SD = 0.68$ s) than for the rare response ($M = 3.74$ s, $SD = 0.99$ s), $t(32) = 3.64$, $p < .005$, $d = 0.62$. This pattern was observed for 24 of the 33 participants, $p < .05$, by a binomial test. The difference in response latency between consensual and nonconsensual responses increased with item consistency. The correlation between this difference and item consistency averaged $-.64$ ($p < .0001$) across participants. It was negative for 29 participants, $p < .0001$, by a binomial test.

3.2.2.3. The postdiction of confidence and latency from response repetition. The observed increase in confidence with item consistency might be due to carry-over effects across repeated presentations. Indeed, confidence increases with repeated solicitation of a judgment (Hasher, Goldstein, & Toppino, 1977). However, we showed that even for Block-1 responses, both confidence and latency discriminated between the more frequent and the less frequent responses. Specifically, Block-1 responses that were repeated three times or more were associated with higher confidence ($M = 86.33$) than those repeated two times or less ($M = 61.57$), $t(32) = 8.88$, $p < .0001$, $d = 1.6$. The respective means for response latency were 4.10 s and 6.09 s (using participants who had both means), $t(30) = 6.75$, $p < .0001$, $d = 1.26$. Thus, responses that were made more often yielded higher confidence and shorter latencies even in Block 1 than responses that were made less often.

3.2.3. Cross-person consensus, within-person consistency, confidence, and response latency

SCM assumes that the clues associated with a category membership pair are largely commonly shared. The implication is that properties of items, notably the likelihood of choosing the majority response and confidence in that response, should be reliable across participants and within participants. Inter-participant reliability for Block 1 was assessed using Cronbach's alpha coefficient (Crocker & Algina, 1986), which yielded a coefficient of .92 for response choice, and .94 for confidence judgments. These coefficients are remarkably high, supporting the assumption that participants base their choice and confidence on clues that are largely commonly shared.

In addition, responses that were consistently chosen by the same person were also more likely to be endorsed by others. Two scores were calculated for each item: (a) The proportion of times that the response made in Block 1 was repeated across the subsequent six blocks and (b) the proportion of *other* participants who made that response in Block 1. When each of these two scores was averaged for each pair across participants, the correlation between these averages (across the 100 pairs) was .81, $p < .0001$.

These results support the idea of a shared pool of representations underlying category membership decisions. The results also suggest that confidence taps into the shared pool of representations. A Pearson correlation was calculated for each participant in Block 1 between confidence judgments and the proportion of other participants who made the same choice. This correlation averaged .34 across participants, $p < .0001$. The correlation was positive for 32 participants. The respective correlation for response latency averaged −.32, $p < .0001$. The correlation was negative for all 33 participants.

In addition, confidence judgments in Block 1 predicted reproducibility – the likelihood of making the same judgment in the future. The Pearson correlation between confidence in Block 1 and the likelihood of making the same response over the subsequent 6 blocks averaged .37, $p < .0001$ across participants. The correlation was positive for all 33 participants. Note, however, that even perfect confidence (100%) in the first presentation was not associated with perfect reproducibility (see Hampton, Aina, Andersson, Mirza, & Parmar, 2012). The respective correlation for response latency averaged −.25, $p < .0001$. The correlation was positive for 32 participants.

Taken together, the results support the idea that confidence in a category membership judgment and the speed of making that judgment tap into the collective pool of clues from which participants sample the clues underlying their judgments in each occasion.

4. Experiment 3

Previous research indicated that category membership decisions are influenced to some extent by the context or perspective in which the decision is made (Barsalou, 1987; Hampton, 2011; Medin et al., 1997; Roth & Shoben, 1983). In Experiment 3 we examined the effects of context on confidence assuming that context can influence the clues that are sampled in making a decision. We expected the changes in confidence that occur because of changes in context to mirror the respective changes that occur in category membership decisions.

Participants made category membership decisions when primed by one of two different contexts for each exemplar-category pair. Context was expected to affect the distribution of the two responses to each item. In addition, confidence was expected to vary with context, being higher for category membership decisions that are compatible with the described context than for those that are incompatible with it.

4.1. Method

4.1.1. Participants

Thirty-six native English speaking undergraduates from the University of Haifa (20 females) participated in the experiment for payment.

4.1.2. Stimulus materials

Ten Object-category pairs were used. They were selected on the basis of the results of Experiment 2 to represent intermediate levels of item consensus. For each pair, two passages were used, each depicting a different context. One context (*neutral context*) was intended to prime the consensual, modal choice, as was found in Experiment 2, and the other was intended to induce a nonconsensual choice (*biasing context*). The items and the corresponding contexts appear in Table 3 in supplementary material. Five passages of each type were slated randomly to each of the blocks, with the assignment to each block counterbalanced across participants. In addition, the Category Width Scale (Pettigrew, 1958) was used, primarily to serve as a filler task between the two blocks.

4.1.3. Procedure

The task was administered in a paper-and-pencil format. Participants were presented with several short passages describing common situations followed by one question each. Participants were asked to read each passage and to imagine themselves being in the situation described in each passage. They were asked to answer the question according to the situation described, by circling *yes* or *no* next to each question. They also indicated their confidence by writing a number in the range 0–100 (0 – very unsure; 100 – very sure).

4.2. Results and discussion

One item was marked as unfamiliar by one participant ("is heather a flower?") and was eliminated from the analyses for that participant.

4.2.1. The effects of context on category membership decisions

For each item, the consensual response in Experiment 2 was defined as the *normative response*, and the nonconsensual response was defined as the *induced response*. Mean percentage of induced responses was significantly higher for the biasing context ($M = 64.17\%$, $SD = 16.45$) than for the neutral context ($M = 42.22\%$, $SD = 15.88$), $t(35) = 6.13$,

$p < .0001$, $d = 1.38$. Of the 36 participants, 30 exhibited this pattern, in comparison with 3 who exhibited the opposite pattern, $p < .0001$, by a binomial test (for 3 participants the two responses were equally frequent). Note that for the 10 items that appeared in Experiment 3, the percentage of induced choices averaged 32.12% in Experiment 2, in which these items appeared without any specific context.

A similar analysis was carried out with item as the unit of analysis. Table 4 in supplementary material lists the mean percentage of induced choices in the neutral and biasing contexts of Experiment 3, and in the no-context condition of Experiment 2. The mean was higher for the biasing context ($M = 69.02\%$, $SD = 21.15$) than for the neutral context ($M = 37.59\%$, $SD = 18.64$), $t(9) = 4.64$, $p < .005$, $d = 1.66$. All 10 items exhibited this pattern, $p < .005$, by a binomial test. The percentage of induced choices was slightly higher in the neutral context of Experiment 3 ($M = 37.59\%$) than in the no-context condition of Experiment 2 ($M = 32.12\%$).

It has been argued that the instability in categorization judgments stems from the lack of explicit context (Braisby, 1993; Braisby & Franks, 1997; Braisby, Franks, & Harris, 1997). We examined the influence of context on vagueness by comparing the results of Experiment 3 with those of Experiment 2 (no context). One item was eliminated from this analysis because it yielded a tie between the two responses in the neutral context. The proportion of responses that deviated from the majority response, calculated across items, was 24.22, 30.65 and 31.98, respectively, for the biasing, neutral and no-context conditions. A repeated measures ANOVA yielded a non-significant effect, $F(2,16) = 1.18$, $MSE = 131.25$, *ns*, $\eta_p^2 = 0.15$. Hampton, Dubois, and Yeh (2006) also failed to find support for the idea that explicit support should reduce vagueness in membership judgments.

4.2.2. The effects of context on response changes

The effects of context can also be gleaned from the changes in the response made in the two contexts. One participant did not change his responses across the two contexts, but 14 participants changed their responses for five items or more. Participants changed their response more often in the predicted direction ($M = 35.96\%$, $SD = 18.92$) than in the opposite direction ($M = 4.44\%$, $SD = 7.35$), $t(35) = 8.60$, $p < .0001$, $d = 2.23$. Thirty-one participants exhibited this pattern, and one exhibited the opposite pattern (for four participant there was a tie), $p < .0001$, by a binomial test.

Each of the 10 items exhibited this pattern, $p < .005$, by a binomial test. In fact, for three items there was a complete reversal so that the majority response differed for the two contexts: *Is dancing a sport*?; *Is fishing a sport*?; and *Is architecture a science*? In all three cases, the reversal was in the expected direction.

4.2.3. Confidence as a function of context

To examine the effects of context on confidence, mean confidence judgments were calculated for each participant for the normative and induced responses for each of the two context conditions. Fig. 7 presents the means across participants. ANOVA was used to test the effects of context

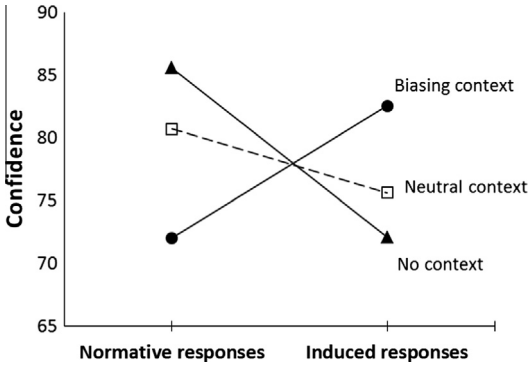

Fig. 7. Mean confidence judgments for the normative and induced responses for the biasing context, neutral context and no-context conditions.

on confidence. One participant made only induced responses in the biasing context, and was eliminated from this analysis. The main effects of context and of type of response (normative vs. induced), were not significant, both $F < 1$. However, the interaction was significant, $F(1,34) = 18.85$, $MSE = 166.0$, $p < .0001$, $\eta_p^2 = .36$. For the neutral context, confidence was higher for normative responses ($M = 80.23$, $SD = 13.03$) than for induced responses ($M = 72.03$, $SD = 21.10$), $t(34) = 2.36$, $p < .05$, $d = 0.57$. Of the 35 participants, 22 exhibited this pattern (one had a tie), $p < .10$, by a binomial test. Note that for these items a similar, but significant difference was observed for the no-context condition of Experiment 2 (see Fig. 7). In that experiment, three participants made the normative response for all ten items, but for the remaining participants, confidence was higher for the normative response ($M = 85.54$, $SD = 13.34$) than for the induced response ($M = 72.07$, $SD = 18.32$), $t(29) = 5.35$, $p < .0001$, $d = 0.85$.

In contrast, in the biasing context of Experiment 3, across 35 participants, mean confidence was in fact higher for the induced responses ($M = 82.72$, $SD = 14.03$) than for the normative responses ($M = 72.01$, $SD = 23.47$), $t(34) = 2.57$, $p < .05$, $d = 0.62$. Of the 35 participants, 26 exhibited this pattern, $p < .005$, by a binomial test. Thus, confidence judgments mirror the changes that occurred in the decision because of the changes in the context of the category membership task.

In conclusion, the results of Experiment 3 yielded a significant effect of context on category membership judgments, demonstrating the malleability of these judgments. It was proposed that context affects category membership judgments by biasing the sampling of clues retrieved, and that the exerted bias should be reflected in confidence judgments. Indeed, these judgments tended to be higher for normative responses in the no-context (Experiment 2) and in the neutral context (Experiment 3) conditions, but the reverse pattern was observed for the biasing context condition.

5. Experiment 4

The aim of Experiment 4 was to relate the results obtained in Experiments 1 and 2 to those of previous stud-

ies on membership typicality (Barr & Caplan, 1987; Hampton, 1995, 1998; Hampton & Gardiner, 1983; McCloskey & Glucksberg, 1978; Rosch, 1973; Rosch, 1975; Rosch & Mervis, 1975). Whereas the typicality task requires ratings on an ordinal scale, the category-membership task requires a binary decision. However, the probability of a positive categorization was found to increase monotonically with the mean typicality rating of an item in the category (Hampton, 1998). In experiment 4, we collected typicality ratings for the pairs used in Experiments 1 and 2.[2] Our aim was to examine how the category membership decisions obtained in these experiments, and the confidence and latency associated with them, relate to typicality ratings.

5.1. Method

5.1.1. Participants, stimulus materials, and procedure

A paper-and-pencil format was used. Participants were 11 native English students. The instructions, which appeared on the first page a booklet, were based on those of Hampton and Gardiner (1983) and asked participants to rate each exemplar according to how typical or atypical it is to the category it was presented with. The remaining 11 pages included 165 pairs, which comprised all the pairs used in Experiments 1 and 2. Each page included a list of candidate exemplars of one category, arranged alphabetically. For each category, there were between 10 and 18 candidate exemplars. The candidate exemplar and the category name were presented as a pair as in Experiment 1 (e.g., apple – FRUIT). Participants rated typicality on a 9-point scale ("1 – atypical; 9 – typical").

5.2. Results and discussion

5.2.1. The relationship between category membership decisions and typicality ratings

Mean percentage of positive categorization was calculated for each item in Experiment 1, and for each item in Block 1 of Experiment 2. The Pearson correlation between the proportion of positive categorization judgments in Experiment 1 and mean typicality ratings was .90, $p < .0001$, across the 101 items. The respective correlation for Block 1 of Experiment 2 was also .90, $p < .0001$, across the 100 items. These results replicate those of Hampton (1998), indicating that by and large, the two tasks tap the same type of gradedness in the membership of exemplars in categories (see Barr & Caplan, 1987; Hampton, 1995; Rosch, 1975).

Typicality ratings were also correlated with within-person consistency in categorization in Experiment 2. For each participant, the responses were classified as consistently positive (the answer "yes" was chosen across all 7 presentations) or as not consistently positive. The percentage of participants who made consistent positive responses for each item correlated .91, $p < .0001$, with mean typicality ratings across the 100 items.

5.2.2. The relationship between confidence and typicality ratings

In Experiments 1 and 2, participants were less confident when they made a nonconsensual choice. We examined whether the confidence ratings obtained in these experiments discriminate also between items that received consensual and nonconsensual typicality ratings.

First, we classified exemplars as typical (scored 5 or above) or non-typical (scored below 5). For each participant in Experiments 1 and 2, we calculated mean confidence for typical and non-typical items in positive and negative categorization responses. One participant in Experiment 1, who made only positive responses for typical items, was eliminated from the analysis. For the remaining participants, mean confidence for positive and negative categorization judgments are plotted in Fig. 8A for typical and non-typical items. A two-way ANOVA on these means yielded non-significant effects for type of response (positive vs. negative), $F(1,19) = 2.36$, $MSE = 81.35$, *ns*, $\eta_p^2 = .11$, and for typicality (typical vs. non-typical), $F(1,19) = 2.08$, $MSE = 2.28$, *ns*, $\eta_p^2 = .10$. The interaction, however, was highly significant, $F(1,19) = 26.69$, $MSE = 55.01$, $p < .0001$, $\eta_p^2 = .58$. For positive categorization judgments, mean confidence was higher for typical items than for non-typical items, $t(19) = 5.30$, $p < .0001$, $d = 1.72$, whereas for negative categorization judgments, mean confidence was lower for typical items than for non-typical items, $t(19) = 3.53$, $p < .005$, $d = 1.14$. The interactive pattern is consistent with the idea that deviant judgments ("no" for typical items and "yes" for non-typical items) are associated with lower confidence than normative judgments.

Fig. 8B presents the same results for Block 1 of Experiment 2. The same ANOVA as before yielded $F(1,32) = 18.05$, $MSE = 51.68$, $p < .0005$, $\eta_p^2 = .36$, for type of response, and $F(1,32) = 9.56$, $MSE = 48.99$, $p < .005$, $\eta_p^2 = .23$, for typicality. In addition, the interaction was significant, $F(1,32) = 34.14$, $MSE = 82.79$, $p < .0001$, $\eta_p^2 = .52$. For positive categorization judgments, confidence was higher for typical items than for non-typical items, $t(32) = 7.14$, $p < .0001$, $d = 1.78$, whereas for negative categorization judgments, confidence was lower for typical items than for non-typical items, $t(32) = 2.54$, $p < .05$, $d = 0.63$.

5.2.3. The relationship between response latency and typicality ratings

A similar pattern to that observed for confidence judgments was found for response latency. A similar ANOVA as before yielded $F(1,32) = 24.95$, $MSE = 1.41$, $p < .0001$, $\eta_p^2 = .44$, for type of response, and $F(1,32) = 0.33$, $MSE = 1.35$, *ns*, $\eta_p^2 = .01$, for typicality. The interaction, however, was significant, $F(1,32) = 31.06$, $MSE = 1.56$, $p < .0001$, $\eta_p^2 = .49$. For positive categorization judgments, response latency was shorter for typical items ($M = 3.46$, $SD = 0.71$) than for non-typical items ($M = 4.78$, $SD = 1.31$), $t(32) = 8.36$, $p < .0001$, $d = 2.07$, whereas for negative categorization judgments, response latency was

[2] Because the items used in Experiments 1 and 2 were taken from different studies that differed in the typicality scales used (1–7 or 1–10), our participants were asked to rate all the items on a 9-point typicality scale.

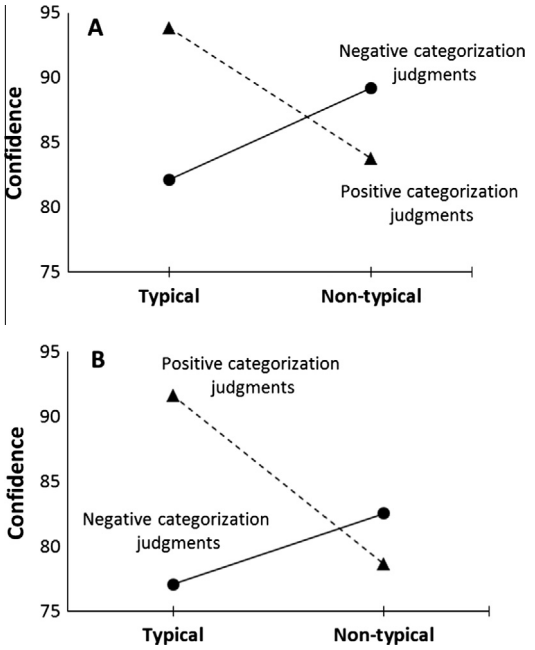

Fig. 8. Mean confidence for positive and negative categorization judgments for typical and non-typical items. Panel A presents the results for Experiment 1, whereas Panel B presents the results for Block 1 of Experiment 2.

longer for typical items ($M = 5.70$, $SD = 2.46$) than for non-typical items ($M = 4.61$, $SD = 0.96$), $t(32) = 2.82$, $p < .01$, $d = 0.70$. This pattern of results is consistent with the findings indicating faster verification times for typical items than for non-typical items (Rips et al., 1973; Rosch, 1973).

The results overall are consistent with the idea that confidence is higher and response latency is shorter for items that are consensually classified with respect to typicality than for the nonconsensual items. It should be stressed, however, that because of the very strong relationships between categorization judgments and typicality ratings, it is difficult to tell whether the results on the relationship between confidence and typicality ratings provide additional information beyond that obtained in Experiments 1 and 2.

6. General discussion

The present study focused on the concrete task faced by a participant who has to decide whether a certain object belongs or does not belong to a particular category. This task was analyzed within a behavioral decision-making perspective. In line with several models of binary decisions (e.g., Alba & Marmorstein, 1987; Koriat, Lichtenstein, & Fischhoff, 1980; Shafir et al., 1993; Slovic, 1975; Stewart, 2009), it was proposed that in attempting to reach a decision, participants summon a variety of discrete clues and considerations sequentially in favor of the two response options, and base their decision on the balance of evidence between the two options. In our theoretical proposal, we evaded the question of the content of the clues underlying categorization, and focused on structural properties of the process. We assumed that participants retrieve associatively a variety of clues and considerations that do not obey any single principle. Think-aloud protocols, and a small paper-and-pencil study that we carried out, suggest that category membership decisions are not based on a single parameter (e.g., similarity to a prototype) or even on a global principle (e.g., family resemblance). Indeed, other authors also considered the possibility of a multiplicity of features underlying categorization (e.g., Hampton, 1998, 2012; Medin, 1989; Rosch, 1973; Smith, Patalano, & Jonides, 1998).

6.1. The distributed model of human categorization decisions

The model proposed assumes that categorization judgments are constructed on the fly depending on the clues and considerations that are sampled at the time of making a category membership decision. The population of clues from which the activated clues are sampled was assumed to consist of a rich, distributed, associative network of representations of many different sorts. Although representations may differ in their accessibility and in their weight in affecting the decision (see Koriat, 2012), we assumed for simplicity that all representations are equally weighted, and each representation has a specific valence favoring a positive or a negative response.

To account for the systematic inter-item differences observed in categorization behavior, we assumed that the population of representations associated with an item is commonly shared by people with the same experience. In turn, to account for between-person and between-occasion variability, we postulated a sampling process in which discrete representations are accumulated sequentially over time, and the sampling is terminated when several representations in a row support the same response (see Audley, 1960). The maximum number of representations sampled in each occasion was assumed to be small because of the cognitive difficulty in integrating evidence across representations (Koriat, 2012).

The assumptions underlying SCM for confidence and response latency have much in common with those of other sampling models of choice and confidence (Juslin & Olsson, 1997; Stewart et al., 2006; Vul et al., 2009). Unlike these models, however, SCM brings to the fore the possibility that even a random sampling of clues can result in systematic differences between different decisions in both subjective confidence and response speed. These differences were assumed to shed light on the online construction of category membership judgments. A very simple implementation of the model was sufficient to bring to the fore the main predictions. Indeed, the results, to be summarized in the next section, provided consistent support for these predictions.

6.2. Summary of the Findings

We will summarize the main findings of the study before discussing them.

6.2.1. Confidence and response consensus

The results of Experiments 1 (Fig. 2) and those of Block 1 of Experiment 2 (Fig. 3) supported the predictions of SCM: Confidence generally increased with increasing cross-person consensus, but consensual responses were endorsed with higher confidence than nonconsensual responses. This was true in comparing the two types of responses either for each individual or for each item. As predicted, the difference in confidence between consensual and nonconsensual responses increased with item consensus – the proportion of participants who made the majority response for the item.

6.2.2. Response latency and consensus

Precisely the same pattern of results was observed for response speed in Block 1 of Experiment 2 (Fig. 4).

6.2.3. Confidence and response latency as a function of response consistency

The pattern of results obtained for cross-person consensus was replicated in the results for response consistency: Mean confidence and mean response speed increased with increasing within-person consistency (Figs. 5 and 6). In comparing the more frequent and the less frequent responses of each participant, the more frequent responses were associated with higher confidence and shorter response latencies than the less frequent responses. The difference in confidence and response speed between frequent and rare responses increased with item consistency – the proportion of times that the frequent response was made across presentations.

6.2.4. Cross-person consensus and within-person consistency

The results supported the assumption that participants base their category membership decisions on representations that are drawn from a commonly shared item-specific population of representations, and that confidence and latency are diagnostic of the properties of that population.

6.2.5. The effects of context

The results of Experiment 3 were consistent with the idea that the context for classification can bias the sample of representations underlying the decision. Context affected category membership decisions to the extent that the consensual response changed from one context to another. Confidence judgments were found to track the changes that occurred with context, mirroring the overall effects of context on category membership decisions (Fig. 7).

6.2.6. Typicality ratings

The percentage of positive categorizations in Experiments 1 and 2 increased with typicality ratings, suggesting that category membership decisions reveal the same type of gradedness as that tapped by typicality ratings. In addition, for typical items, confidence in positive categorization responses was higher than confidence in negative categorization responses, whereas the opposite was found for nontypical items (Fig. 8). A similar pattern was observed for response speed.

6.3. Stability and variability in categorization

Let us now examine the implications of these results for the source of stability and variability in category membership decisions. The results overall indicated that on the one hand, there was a great deal of cross-person consensus and within-person consistency in categorization. On the other hand, there was some between-individual variation for most items, and also some within-person variability in the response to the same item across presentations. These results are consistent with those of previous studies (see Barr & Caplan, 1987; Estes, 2003; McCloskey & Glucksberg, 1978). What is notable, however, is that the confidence in one's decision, and the speed of forming that decision were sensitive not only to inter-item differences in categorization judgments, as has been demonstrated by others (e.g., Estes, 2004; McCloskey & Glucksberg, 1979); they were also sensitive to inter-response differences, providing a clue to the on-line construction of categorization judgments.

In discussing the construction of attitudinal judgments, Koriat and Adiv (2011) proposed that the distinction between the stable and variable components of these judgments can be conceptualized in terms of the distinction between *availability* and *accessibility* (see Tulving & Pearlstone, 1966). Likewise, we propose that in making category membership decisions for a given item, the stable components derive from the constraints imposed by the population of representations *available* in memory. The critical property of that population is the distribution of representations that speak for a positive or a negative response. The polarity of that population (p_{maj}) is assumed to account for the systematic inter-item differences in categorization but also to constrain the extent of cross-person and within-person variation. Thus, for some very typical exemplars, most or all representations that come to mind would favor a positive response. Importantly, mean confidence and mean response speed for an item were assumed to tap the polarity of the population of representations, and indeed, both were found to increase with item consensus and with item consistency as predicted.

The response made in each occasion, however, was assumed to depend on the specific sample of representations that are *accessible* at the time of the decision. The systematic differences observed in confidence and response speed between consensual and nonconsensual responses were assumed to reflect differences in the specific set of representations that are sampled by different individuals for the same item. Likewise, the differences between frequent and rare responses were assumed to reflect differences between the samples drawn in different occasions. Thus, assuming that the consensual/frequent response for a given item is the response that follows from p_{maj}, that response was found to yield higher confidence and shorter response latency than the nonconsensual/rare response, as predicted.

What is notable is that for both confidence and response speed the difference between consensual and nonconsensual responses increased with item consensus. Similarly, the difference between frequent and rare responses also increased with item consistency. This

interactive pattern is consistent with the idea that as p_{maj} increases, not only does the proportion of minority decisions decreases, but the within-sample proportion of representations supporting these decisions also decreases, resulting in lower confidence and longer response times. Thus, variability is not entirely random but obeys a certain pattern that is consistent with the sampling assumption.

6.4. The explanation of inter-person variability and within-person instability

Our explanation of the variability in category membership decisions may be compared to that offered by other researchers. The results of McCloskey and Glucksberg (1978) and Hampton (1998) indicate that instability is related to intermediate degrees of typicality. However, why do people differ in their categorization responses, and why does the same person make different responses on different occasions? Some researchers proposed that variability in categorization stems from the lack of a clear context for classification (Braisby, 1993; Braisby & Franks, 1997; see also Hampton et al., 2006). Barsalou (1987) suggested the possibility of consistent individual differences in concept representation that may contribute to inter-participant variability in categorization. In addition, recent experience can affect category representation, resulting in within-person fluctuations in categorization across occasions. Another suggestion was that instability in categorization stems from random variation in the process of computing similarities and variation in the placement of the criterion differences in the threshold of similarity criteria for deciding that a candidate exemplar belongs to a particular category (Hampton, 1995; Hampton et al., 2012; Verheyen, Hampton, & Storms, 2010). According to the Threshold Theory (Hampton, 1995, 2007), the threshold criterion can vary from one person to another and from one occasion to another (Hampton, 1995; Verheyen et al., 2010).

The sampling assumption underlying the distributed model provides a more principled account of instability in categorization. Variability is assumed to be inherent in the sampling process, and would be expected even when all conditions are equal. When the distribution of representations associated with an item is not strongly polarized, the sampling of a small set of representations from a large pool is bound to yield some variability across occasions. Indeed, many theories of judgment and decision incorporate a sampling assumption to account for variability. Many models of confidence in psychophysical tasks assume random fluctuations that are due to internal noise (e.g., Audley, 1960; Merkle & Van Zandt, 2006; Vickers, 1970; for a review, see Baranski & Petrusic, 1998). Unlike these models, however, SCM assumes that confidence judgments and response latency can provide some clue to the sampling underlying binary decisions even if this sampling is completely random (see Koriat, 2012). As the simulation experiment described in the introduction indicates, a random sampling of clues is bound to yield lower confidence and longer response times for responses that deviate from the response that is implied by p_{maj}.

It is important to stress that in the sampling model proposed, the same process is assumed to underlie consensual and nonconsensual responses. Both types of responses (as well as frequent and rare responses in a repeated presentation design) are assumed to be based on the "majority vote" in the specific sample underlying the category membership decision.

6.5. The effects of context

As noted earlier, several researchers attributed the vagueness of the categorization task to the lack of a clear discourse context provided to participants (e.g., Braisby & Franks, 1997). In terms of the model proposed here, context can bias the sampling of representations that is assumed to underlie category membership decisions. Indeed, in Experiment 3, the biasing contexts tended to shift the choice towards the induced responses. In addition, consistent with SCM, confidence judgments were found to vary with context in a manner that mirrored the changes that occurred in categorization. Thus, in each context condition, confidence was higher for the consensual response in that condition than for the nonconsensual response.

Hampton et al. (2006), however, found little evidence that clarifying the context reduces disagreement and inconsistency. A comparison of the results of Experiment 3, in which context was always provided, with the respective results obtained in Experiment 2, in which context was not specified, also yielded little evidence that the provision of a specific context reduces vagueness, as indexed by cross-person variability.

However, should the specification of context always reduce disagreement? According to our model, it should do so only under specific conditions. Consider a control, no-context condition in which an exemplar-category pair is associated with $p_{maj} = .70$ favoring a positive response. This pair is expected to yield a .87 (pc_{maj}) agreement (e.g., the pair "Coconut – FRUIT" in Experiment 2). A biasing context that induces a shift to a negative response should increase agreement only if it produces a sample that corresponds to that of $p_{maj} > .70$. However, it would be expected to increase disagreement if it produces samples that are more like those that are characteristic of $p_{maj} = .60$. In fact, it should be difficult to create a scenario that reverses the dominant category membership decision and still achieve a degree of agreement that corresponds to that of $p_{maj} > .70$. Thus, only when the biasing context acts to reinforce the "default" context, would we expect an increase in agreement, but that increase would be difficult to detect.

Given the difficulties in confirming the effects of context on categorization, it is particularly important that these effects could be detected by confidence judgments. Studies of the effects of context can benefit from the collection of confidence judgments and response latency in clarifying the effects of context on the degree of between-person agreement and within-person consistency in category membership decisions.

6.6. Basic assumptions and a comparison with other approaches

In concluding this article, we review some of the distinctive characteristics of our approach, and examine how other views may accommodate the findings obtained in this study. We then point out some of the problems and difficulties in our own proposal.

The general approach that we took in the present investigation differs from that of previous studies in several respects. First, we focused on the concrete task faced by a participant when having to decide whether a certain object belongs or does not belong to a certain category. Verbal reports suggest that participants retrieve a mixture of clues and considerations that are not governed by any simple rule. Many of the clues that come to mind consist of associations, hunches, and images that are not readily expressed in the form of declarative statements.

Second, we assumed that categorization decisions are generally constructed on the spot rather than retrieved ready made from memory. This assumption is common among students of social attitudes (Schwarz & Strack, 1991; Tourangeau, 1992) and personal preferences (Slovic, 1995). Like in these domains, this assumption can help account for contextual effects and for the instability in categorization decisions.

Third, we analyzed the categorization process within a decision-making framework. As Medin and Smith (1984) noted, "given that the distinction between categorization and decision making is rather fuzzy, there has been surprisingly little interplay between formal models in these two areas" (p. 126). Following some theories of choice and decision (Alba & Marmorstein, 1987; Allwood & Montgomery, 1987; Koriat et al., 1980; Shafir et al., 1993), we proposed that in making a category membership decision, participants retrieve a variety of clues and considerations, weigh the pros and cons for each answer, and then settle on one option.

Finally, we assumed that category membership decisions rely on a process in which discrete pieces of information are accumulated and consulted sequentially. The final decision is based on the online aggregation of "subdecisions" (see Koriat, 2012). Medin and Smith (1984) argued that "although most categorization models assume that people are essentially making similarity judgments, often it is unclear whether these judgments constitute a holistic impression of overall similarity or a more analytic accumulation of matches and mismatches of components" (p. 127). Our assumption that discrete representations are sampled is critical for the SCM predictions regarding confidence and latency. In addition, the latency results also suggest that the retrieval of clues is a self-terminating process. We leave open the question whether typicality judgments also rely on an analytic process, or are based on an overall, holistic impression.

Can our results be accommodated by other models of categorization? Clearly, our model and findings are most consistent with the probabilistic view that concept representations are based on properties that are only characteristic or typical of category examples rather than on defining properties. SCM is particularly compatible with models in which information about similarity relations is assumed to accumulate over time until a certain criterion is reached (e.g., Verheyen et al., 2010). Two examples will be mentioned. McCloskey and Glucksberg (1979) proposed that when a category membership sentence is presented for verification, properties of the subject and predicate concepts are retrieved and compared. Each comparison yields either evidence that the sentence is true or evidence that it is false. Both types of evidence accumulate, and the process terminates when sufficient evidence has been collected to exceed a "true" or "false" decision criterion.

Our results can also be accommodated by exemplar views in which exemplars are assumed to be retrieved and compared in turn to the test item during categorization. In Nosofsky and Palmeri's (1997) exemplar-based random walk model, exemplars are assumed to race for retrieval during speeded categorization, with rates determined by their similarity to the test item. The retrieved exemplars provide incremental information that enters into a random walk counter, and once the counter reaches a pre-established criterion, the appropriate categorization response is made.

These models, like SCM, can account for differences between object-category pairs in the probability of positive responses and in response latency. It is unclear, however, how they can be accommodated to explain the differences between different responses in confidence and latency. The advantage of SCM is that it accounts for differences between items and between responses within the same conceptual framework.

There is no question that the model that we proposed for category membership decisions is very rudimentary and does not capture the complexity of the processes involved. However, it accounts rather parsimoniously for the major qualitative patterns of results obtained in this study. Importantly, these qualitative patterns have been confirmed for a wide range of tasks (see Koriat, 2012), suggesting that these tasks, as well as the categorization task, can be analyzed within a decisional framework that incorporates the sampling of clues from a shared pool of clues.

However, one theoretical problem should be pointed out in the application of SCM to category membership decisions. SCM has been developed to account for decisions in which the sampled representations may favor either one of two alternative responses. Thus, it was applied to 2AFC general-information questions (Koriat, 2008) and perceptual comparisons (Koriat, 2011). It has also been applied to social attitudes items in which the response options were *yes* (favor) and *no* (oppose). In the case of category membership judgments, in contrast, it is unclear how information can be accumulated to favor a negative response. The question of the basis for negative responses has also been discussed in the context of similarity-based approaches to categorization (e.g., Hampton, 1993; McCloskey & Glucksberg, 1979; Rosch, 1973). However, in the context of SCM, it is unclear whether items for which the consensual response is negative (e.g., chicken – FRUIT) can be said to be associated with a commonly shared population of representations favoring a *no* response. This might be the case, as suggested by the observation that non-typical items yielded the same trend of higher

confidence for consensual than for nonconsensual responses. Clearly, however, further theoretical and empirical work is needed to address this and other problems.

Acknowledgements

The work reported in this paper was conducted as part of the second author's Master thesis. The research was supported by the Max Wertheimer Minerva Center for Cognitive Processes and Human Performance at the University of Haifa. We are grateful to Shiri Adiv for her assistance in this work and to Miriam Gil for her help in the analyses. Tamar Jermans helped in the collection of the data. We also thank Etti Levran (Merkine) and Ornit Tzuri for their help in copyediting.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2014.09.009.

References

Alba, J. W., & Marmorstein, H. (1987). The effects of frequency knowledge on consumer decision making. *Journal of Consumer Research, 14*, 14–25.

Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior and Human Decision Processes, 39*, 365–383.

Anderson, R. C., & Ortony, A. (1975). On putting apples into bottles: A problem of polysemy. *Cognitive Psychology, 7*, 167–180.

Ashcraft, M. H. (1978). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition, 6*, 227–232.

Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological Review, 67*, 1–15.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 929–945.

Barr, R. A., & Caplan, L. J. (1987). Category representations and their implications for category structure. *Memory & Cognition, 15*, 397–418.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 101–140). New York: Cambridge University Press.

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge: Cambridge University Press.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research, 25*, 187–217.

Bless, H., Mackie, D. M., & Schwarz, N. (1992). Mood effects on encoding and judgmental processes in persuasion. *Journal of Personality and Social Psychology, 63*, 585–595.

Braisby, N. R. (1993). Stable concepts and context-sensitive classification. *Irish Journal of Psychology, 14*, 426–441.

Braisby, N. R., & Franks, B. (1997). What does word use tell us about conceptual content. *Psychology of Language & Communication, 1*, 5–16.

Braisby, N. R., Franks, B., & Harris, J. (1997). Classification and concepts: Fuzzy or perspectival? In D. Dubois (Ed.), *Catégorisation et cognition: De la perception au discours* (pp. 163–189). Paris: Kimé.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metacognitive approach. *Memory, 14*, 540–552.

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language, 67*, 59–77.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. *Cognition and Categorization, 3*, 170–211.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

DeSoto, K. A., & Roediger, H. L. III, (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science, 25*, 781–788.

Dhami, M., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. *Memory & Cognition, 31*, 199–214.

Estes, Z. (2004). Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural. *Psychonomic Bulletin and Review, 11*, 1041–1047.

Galton, F. (1907). Vox populi. *Nature, 75*, 450–451.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*, 20–29.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*, 441–461.

Hampton, J. A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 12–32.

Hampton, J. A. (1993). Prototype models of concept representation. In I. van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67–95). London: Academic Press.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*, 686–708.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137–165.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*, 355–384.

Hampton, J. A. (2009). Stability in concepts and evaluating the truth of generic statements. In F. J. Pelletier (Ed.). *Kinds, things, and stuff: Concepts of generics and mass terms. New directions in cognitive science* (Vol. 12, pp. 80–99). Oxford: Oxford University Press.

Hampton, J. A. (2011). Concepts and natural language. In R. Belohlavek & G. J. Klir (Eds.), *Concepts and fuzzy logic* (pp. 233–258). Cambridge: MIT Press.

Hampton, J. A. (2012). Thinking intuitively: The rich (and at times illogical) world of concepts. *Current Directions in Psychological Science, 21*, 398–402.

Hampton, J. A., Aina, B., Andersson, J. M., Mirza, H., & Parmar, S. (2012). The Rumsfeld effect: The unknown unknown. *Journal of Experimental Psychology: Learning Memory & Cognition, 38*, 340–355.

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

Hampton, J. A., Dubois, D., & Yeh, W. (2006). The effects of pragmatic context on classification in natural categories. *Memory & Cognition, 34*, 1431–1443.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16*, 107–112.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104*, 344–366.

Kelley, C. M., & Lindsay, S. D. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*, 321–341.

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition, 4*, 244–248.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*, 945–959.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General, 140*, 117–139.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*, 80–113.

Koriat, A. (2013). Confidence in personal preferences. *Journal of Behavioral Decision Making, 26*, 247–259.

Koriat, A., & Adiv, S. (2011). The construction of attitudinal judgments: Evidence from attitude certainty and response latency. *Social Cognition, 29*, 577–611.

Koriat, A., & Adiv, S. (2012). Confidence in one's social beliefs: Implications for belief justification. *Consciousness and Cognition, 21*, 1599–1616.
Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107–118.
Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69.
Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. New York, NY: Cambridge University Press.
McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*, 462–472.
McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11*, 1–37.
Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469–1481.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology, 35*, 113–138.
Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*, 49–96.
Merkle, E. C., & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology: General, 135*, 391–408.
Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science, 32*, 1133–1147.
Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700–708.
Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 3–27.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300.
Oden, G. C. (1977). Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory & Cognition, 5*, 198–204.
Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*, 972–987.
Pettigrew, T. F. (1958). The measurement and correlates of category width as a cognitive variable. *Journal of Personality, 26*, 532–544.
Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*, 1–20.
Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*, 416–425.
Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.
Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.
Ross, M., Buehler, R., & Karr, J. (1998). Assessing the accuracy of conflicting, autobiographical memories. *Memory and Cognition, 26*, 1233–1244.
Roth, E. M., & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15*, 346–378.
Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25*, 638–656.
Schwarz, N. (2008). Attitude measurement. In W. Crano & R. Prislin (Eds.), *Attitudes and persuasion* (pp. 41–60). Philadelphia: Psychology Press.
Schwarz, N., & Strack, F. (1991). Context effects in attitude surveys: Applying cognitive theory to social research. *European Review of Social Psychology, 2*, 31–50.
Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*, 11–36.
Slovic, P. (1975). Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 280–287.
Slovic, P. (1995). The construction of preference. *American Psychologist, 50*, 364–371.
Smith, E. E., & Sloman, S. A. (1994). Similarity versus rule-based categorization. *Memory & Cognition, 22*, 377–386.
Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65*, 167–196.
Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*, 93–109.
Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62*, 1041–1062.
Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1–26.
Tourangeau, R. (1992). Context effects on responses to attitude questions: Attitudes as memory structures. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 35–47). New York: Springer.
Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5*, 381–391.
Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505.
Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica, 135*, 216–225.
Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*, 37–58.
Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the seventeenth annual meeting of the international society for psychophysics* (pp. 148–153). Lengerich, Germany: Pabst Science.
Vickers, D., & Pietsch, A. (2001). Decision making and memory: A critique of Juslin and Olsson's (1997) sampling model of sensory discrimination. *Psychological Review, 108*, 789–804.
Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the 31st annual conference of the cognitive science society* (Vol. 1, pp. 66–72).
Warren, C., McGraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *Interdisciplinary Reviews: Cognitive Science, 2*, 193–205.
Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*, 1049–1062.