Journal of Behavioral Decision Making
J. Behav. Dec. Making, **21**: 253–281, (2008)
Published online 2 January 2008 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/bdm.587

Confidence and Decision Type Under Matched Stimulus Conditions: Overconfidence in Perceptual but Not Conceptual Decisions

SARA KVIDERA* and WILMA KOUTSTAAL
Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA

ABSTRACT

Within the domain of metacognition, there is disagreement whether different processes underlie evaluations of confidence in perceptual versus conceptual decisions. The relationship between confidence and accuracy for perceptual and conceptual decisions was compared using newly created stimuli that could be used to elicit either decision type. Based on theories of Brunswikian and Thurstonian uncertainties, significant underconfidence for perceptual decisions and overconfidence for conceptual decisions were predicted. Three within-subjects experiments did not support this hypothesis. Participants showed significant *overconfidence* for perceptual decisions and *no* overconfidence for conceptual decisions. In addition, significant hard-easy effects were consistently found for both decision types. Incorporating our findings with past results reveals that both over- and underconfidence are attainable on perceptual tasks. This conclusion, in addition to the common presence of hard-easy effects and significant across-task correlations in over/underconfidence, suggests that confidence judgments for the two decision types may depend on largely shared processes. Possible contributions to confidence and over/underconfidence are explored, focusing on response time factors and participants' knowledge bases. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS over/underconfidence; calibration; decision making; discrimination; confidence judgments; post-test performance estimate (PTPE)

INTRODUCTION

The ability to consider and evaluate our own thinking, or to demonstrate what has been termed ''metacognition'' (Flavell, 1979), is a crucial human capacity (for reviews, see Fernandez-Duque, Baird, & Posner, 2000; Nelson, 1996; Nelson & Narens, 1994). Whether such meta-cognitive evaluations provide accurate or useful information and, in particular, how an individual's level of confidence relates to actual performance are central questions in this domain. Additional questions, such as ''How well do people

* Correspondence to: Sara Kvidera, Department of Psychology, 75 East River Road, Minneapolis, MN 55455, USA.
E-mail: kvid0007@umn.edu

understand the possible reasons for their uncertainty?'' and ''How well can they utilize these insights to assess their level of confidence in a decision?'' arise from findings that confidence is not always a reliable indicator of accuracy (see Keren, 1988, 1991, 1997; McClelland & Bolger, 1994, for reviews).

Answers to these questions would allow for a better basic understanding of human cognition, but they also have important practical implications. Studies of *confidence calibration* (how well confidence matches accuracy) are significant as an individual's confidence often guides that person's decisions and the decisions of others (e.g., Simmons & Nelson, 2006). Therefore, errors in confidence calibration often can have negative, and sometimes disastrous, consequences. For instance, unwarranted high levels of confidence in eyewitness testimony have been shown to influence juries to convict innocent people, particularly if highly confident assertions are assumed to be accurate (Deffenbacher, 1980; see Price & Stone, 2004, for evidence that individuals may use a ''confidence heuristic,'' assuming that more confident advisors also are more knowledgeable).

Pioneering work on the relationship between confidence and decision accuracy done by Adams and Adams (1961) indicated that people tended to be overconfident when making decisions. *Overconfidence* indicates that, on average, the amount of confidence expressed in relation to a group of decisions (as indicated as a percentage) is higher than the accuracy of those decisions (as a percentage). A number of subsequent studies have obtained similar results (e.g., Stankov, 1998; see Keren, 1991 for a review). However, evidence suggests that this overconfidence phenomenon occurs most often when participants are making conceptual decisions, such as answering general knowledge questions (Lichtenstein, Fischoff, & Phillips, 1982) regarding one's knowledge of history, geography, or literature (Fischhoff, Slovic, & Lichtenstein, 1977), or when answering problem-solving questions (Crawford & Stankov, 1996).

In contrast to the typical pattern of results found for conceptual tasks, early experiments in psychophysics (e.g., Peirce & Jastrow, 1884) found that people could often detect perceptual differences (e.g., which weight is heavier) when they were *unaware* they could do so, thus demonstrating underconfidence for these perceptual decisions. The opposite of overconfidence, *underconfidence* arises when, on average, the confidence one expresses in a group of decisions is lower than the accuracy of those decisions. A number of recent studies have replicated the finding of underconfidence for perceptual decisions. For instance, Bjorkman, Juslin, and Winman's (1993) participants showed an average of 13% underconfidence on length and weight discrimination tasks, and Stankov (1998) found underconfidence of 16% on a line-length judgment task (see also Stankov & Crawford, 1996). Baranski and Petrusic (1994) also found that underconfidence occurred on their line-length discrimination tasks. However, they discovered overconfidence can also occur for this perceptual task, especially on the hardest items (a phenomenon called the *hard-easy effect*, typically found for conceptual decisions). Additionally, Bar-tal, Sarid, and Kishon-Rabin (2001) recently observed both a hard-easy effect and general *over*confidence in a perceptual task using an auditory stimulus detection paradigm.

The frequent, although not uniformly, observed finding of differing relations between accuracy and confidence for conceptual versus perceptual decisions, with overconfidence most commonly observed for conceptual decisions but underconfidence for perceptual decisions, has led to the hypothesis that in order to establish confidence for the two different types of decisions (conceptual versus perceptual), people rely on two separate processes (Bjorkman et al., 1993; Runeson, Juslin, & Olsson, 2000). In particular, it has been proposed that overconfidence for conceptual decisions results from Brunswikian uncertainty, a discordance between the cues participants have learned from the world and their effectiveness in being utilized to answer the questions that are presented in the experimental situation, that is, errors of ''cue validity'' (Juslin & Olsson, 1997). In contrast, errors in confidence for perceptual tasks are thought to be based solely on Thurstonian uncertainty, or the inaccuracies of the sensory/perceptual system (Winman & Juslin, 1993).

From these two sources of error, overconfidence can be predicted on conceptual tasks in which participants depend on cues (e.g., heuristics) that are unreliable guides to decisions. For instance, in their development of the Brunswikian viewpoint, Gigerenzer, Hoffrage, and Kleinbölting (1991) have argued that

DOI: 10.1002/bdm

overconfidence on general knowledge tests arises from systematic biases in the selection of the items that are tested, such that rules of thumb that normally would apply often do not apply to the selected items.

Evidence for the role of misleading items in overconfidence is mixed. Although random sampling of items sometimes reduces overconfidence on general knowledge tasks (Gigerenzer et al., 1991), it does not always eliminate it (e.g., Griffin & Tversky, 1992; see Brenner, Koehler, Liberman, & Tversky, 1996 for a review). Nonetheless, item analyses have sometimes pointed to systematically incorrect answers to particular items suggesting reliance on inappropriate heuristics (for an extensive review of heuristics, see Kahneman, Slovic, & Tverksy, 1982). Indeed, it has been suggested that ''misleading'' items (i.e., items that are consistently answered incorrectly or, on average, at below chance levels) will be found more frequently on conceptual tasks, whereas such items are less likely to be found for perceptual tasks, because errors arise from random, rather than systematic, errors within the perceptual system. Consistent with this proposal, Stankov (1998) presented evidence that overconfidence on a vocabulary knowledge task was largely attributable to a small number of items containing a ''familiar attractor'' incorrect alternative. This pattern was not observed for a perceptual line-length discrimination task where the vast majority of the items elicited underconfidence.

In their subjective distance theory, Bjorkman et al. (1993) outline how underconfidence can be predicted for perceptual tasks. In their theory, perceptual decision-makers rely on the subjective difference between the stimuli to rank their confidence. The stimuli which have the least amount of subjective difference in a two-choice task are rated at the 50% level of confidence, or ''guessing.'' Thurstonian uncertainty dictates that these subjective differences will be normally distributed. Therefore, at the lowest level of confidence (50%), there will always be more correct answers than incorrect answers. These researchers also argue that this discrepancy will continue to be seen at higher levels of confidence.

However, not all researchers support the need for these two separate models of confidence. For instance, Ferrell (1995) argues that the more comprehensive signal detection model can account for variance in over/underconfidence in both perceptual and conceptual tasks. He also notes that the hard-easy effect has been observed in perceptual tasks and is predicted by the signal detection model.[1]

Therefore, the debate continues: do confidence judgments made with regard to conceptual decisions depend on different cognitive processes than do confidence judgments for perceptual decisions? Research directly addressing this issue has been conducted with mixed results. For example, Dawes (1980) found overconfidence for his ''intellectual'' tasks (knowledge questions), but only intermittent underconfidence for perceptual judgments (tone-length discrimination and area judgments) depending on the participant pool. Keren (1988) found overconfidence on questions probing geographical and populace knowledge and underconfidence on a visual-perceptual Landolt ring task. However, Baranski and Petrusic (1995) found similar levels of over/underconfidence and calibration for their conceptual (questions) and perceptual (pictures) tasks.

Notably, however, there are important limitations to all of the previous research addressing this issue. Many comparisons between conceptual and perceptual decisions have been done across studies (e.g., Dawes, 1980), thus allowing for the intrusion of potential confounds (e.g., between-subject differences in overall level of confidence, or differing experimenter instructions). Even the experimental designs that have been used to examine the two decision types within a study have employed extremely different stimuli (e.g., Keren, 1988) and/or procedures for the two types of tasks. For instance, researchers often use general knowledge questions for the conceptual decisions and line-length judgment stimuli for the perceptual decisions. This disparity makes it unclear whether the differences in over/underconfidence for the perceptual versus conceptual tasks are due to differences in the way people determine confidence for the two decision types, or rather, to differences in the stimuli and various other aspects of the decision context.

---

[1]Erev, Wallsten, and Budescu (1994) also argue that the observation of over/underconfidence does not depend on the types of judgments made (in their case, revision-of-opinion tasks vs. general knowledge tasks). Instead, they suggest that such disparities result from differences in the way the data are analyzed.

DOI: 10.1002/bdm

The experiments we report here are critically different; each experiment maximizes experimental control by the use of a within-subjects design[2] and the *presentation of the same stimuli for both the decision types*. Furthermore, all other aspects of the procedure for the two decision types (e.g., number of items, presentation parameters) likewise are held constant.[3] Our study also obtains two measures of confidence: (1) confidence ratings for each item answered (''item-by-item'' confidence ratings) and (2) confidence ratings for a set of decisions of a given type, elicited after a group of those decisions had been made, asking the participant to give a global or aggregate ''post-test performance estimate'' (PTPE) of his/her decision accuracy. Finally, as noted above, levels of over/underconfidence can be affected by the difficulty of the task (hard-easy effect). Thus, task difficulty (i.e., proportion correct) is equated when making comparisons of over/underconfidence for perceptual and conceptual decisions in this study. Overall, this new paradigm presents a unique method for ruling out stimulus and procedural differences as explanations for differences seen in the confidence–accuracy relationship for conceptual and perceptual decisions.

If the dual processes account of confidence judgments is correct, we predict that overconfidence will be seen on the item-by-item confidence ratings for the conceptual decisions and underconfidence will be seen for the perceptual decisions. However, if the differences in over/underconfidence as a function of decision type that have been observed in previous research are artifacts of the experimental designs, then we predict that the level of over/underconfidence will be similar for both perceptual and conceptual decisions.

Beyond considering over/underconfidence measures, this study evaluates the extent to which differences in levels of confidence ''discriminate'' between correct versus incorrect decisions. Measures of discrimination are important because, as discussed above, confidence can influence behavior (both that of the actor, and that of observers, who make decisions on the basis of expressed confidence). One can imagine that high levels of confidence may lead people to act quickly whereas low levels may lead to inaction and/or continued investigation and information gathering regarding the available options. Thus, confidence that better discriminates between correct and incorrect decisions would enhance one's ability to choose the best option.

Finally, as noted above, a global measure of confidence, the PTPE also was obtained in these experiments. In accordance with previous research (May, 1988, cited in Brenner et al., 1996; Liberman, 2004), we predict that confidence (and also overconfidence) will be significantly lower on the PTPE than for the average of the individual ''item-by-item'' confidence ratings. One account of the lower levels of confidence often found on the PTPE is that participants fail to take into consideration the likelihood of correctly answering by guessing (Liberman, 2004). However, lower levels of confidence on the PTPE than on item-by-item ratings also have been reported even when participants were directly reminded, at the time of making their PTPE, of the proportion they could expect to obtain correctly by chance alone (e.g., Fu, Koutstaal, Fu, Poon, & Cleare, 2005).

Gigerenzer et al. (1991) propose that the lower levels of confidence often found on the PTPE are due to participants relying on different ''reference classes'' in making PTPE versus item-by-item confidence judgments. Specifically, when making item-by item confidence ratings in the absence of knowledge which would produce certainty, decision-makers consult a ''reference class'' that includes the stimuli that are present in the experimental and task context (e.g., all cities in Germany when answering general knowledge questions about German cities). In contrast, when making PTPE confidence ratings, decision-makers consult a reference class that includes all previous experiences with such testing (or similar) situations. Thus, for

[2]Within-subjects designs eliminate between-group variability effects, but are not unique to our research; others have used such designs (e.g., Juslin, Winman, & Olsson, 2003; Stankov, 1998).

[3]Using the same stimuli for both tasks would also allow for a valid comparison of the two types of confidence assessments in any future research using neuroimaging techniques such as functional Magnetic Resonance Imaging and Event-Related Potentials (see, e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001; Cutmore & Muckert, 1998; Luu, Collins, & Tucker, 2000; Scheffers & Coles, 2000).

DOI: 10.1002/bdm

item-by-item confidence, decision-makers consider their knowledge in a certain domain, whereas for PTPE confidence, decision-makers consider their general ability to perform on similar decision-making tasks.

In the current research, participants were required to give their PTPEs on the same scale as the item-by-item confidence ratings (i.e., only responses from 50 to 100% were allowed). This provides a more tightly controlled comparison of the two forms of confidence assessment across perceptual and conceptual tasks; the consequences of a failure to take into account the effectiveness of guessing will be mitigated in this design. Therefore, if lower overconfidence nonetheless is still observed for the PTPE than for the item-by-item confidence measure, this outcome would give qualified support to Gigerenzer et al.'s (1991) reference class hypothesis.

EXPERIMENT 1

Method

Materials

The stimuli were designed so that they could be used to elicit *either* conceptual or perceptual decisions (see Figure 1, panels A–C). The stimuli consisted of pairs of words which were manipulated in area, length, or height in such a way that the participant could be asked to judge either the physical size of the *referent* of the words (conceptual decision) or the physical size of the words *as displayed on the computer screen* (perceptual decision).[4]

Three categories of stimuli for each dimension were created: the ''smaller'' dimension included birds, cars, and countries; the ''longer'' dimension included fish, mammals, and rivers; and the ''taller'' dimension included buildings, dogs, and trees. Items were selected for inclusion because they were the most common or the most popular items in the category within the United States, with the constraints that (1) items with numbers in their names (e.g., Ford F-150), (2) items with size information in their names (e.g., miniature

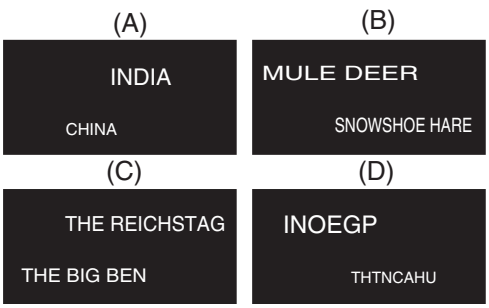

Figure 1. Example stimuli for the smaller (A), longer (B), and taller (C) decision tasks. Panel D shows an example of a Remixed stimulus from the smaller dimension (the original pairing was PIGEON-NUTHATCH)

[4]In using the same stimuli for both decision types, our perceptual stimuli necessarily vary on a greater number of aspects than did the stimuli in past perceptual tasks. For instance, for line-length discriminations only one aspect (length) varies between the two options, whereas our perceptual length stimulus options vary by length, number of letters, and (in the case of real-word stimuli) the objects named and the participant's knowledge-based associations with those objects. Therefore, on the one hand, there is the potential for additional sources of error involved in these decisions, beyond Thurstonian noise. The potential presence of such influences is addressed to some extent with the tests for ''Stroop-like effects'' discussed further below. On the other hand, our new paradigm also enables a test of the generality of the perceptual versus conceptual task differences in confidence assessment, beyond the frequently used general knowledge and line-length judgment tasks.

DOI: 10.1002/bdm

schnauzer), and (3) items that contained the same word that appeared in a more popular/common item (e.g., United Arab Emirates; United States of America) were all excluded. Once lists of 40 items per category were compiled, the words were paired semi-randomly, with the constraint that the real world size of the objects could not overlap (thereby ensuring that there would be a single correct answer for the conceptual decision). Thus, 20 word pairs were created per category. This pairing procedure was performed twice to create two unique sets of word pairs to permit greater generalizability across stimuli.

The word pair stimuli were produced using Macromedia FreeHand MX software (Macromedia, Inc., 1988–2003), which allowed for the precise spatial manipulation of the stimuli in each dimension (area, width, and height). All stimuli were typed in Arial font in capital letters (white font on a black background). The text blocks of each pair were first equated on the dimension of interest. For instance, for word pairs in the longer dimension, this was done by first finding the average length of both text blocks; then, each text block was manipulated so that its length was equal to that average length. The same process was performed for the smaller stimuli using the average area of the text blocks. Equating was unnecessary for the taller dimension, because all the text was originally written in the same font size.

Once the text blocks were equated on their dimension of interest, the stimuli were randomly assigned to one of five levels of difference in that dimension to create varying levels of difficulty for the perceptual tasks. Text blocks were altered so that they differed by 1, 2, 3, 6, or 12 mm in the longer dimension, by 10, 30, 60, 100, or 120 mm^2 in the smaller dimension, and by 0.05, 0.10, 0.12, 0.20, or 0.60 mm in the taller dimension. Each text block was altered by half of the intended difference. Thus, for a longer stimulus pair assigned to differ by 6 mm, one text block would be stretched 3 mm, whereas the other would be reduced by 3 mm. In addition, text blocks were offset horizontally by one of four levels of magnitude (12, 16, 20, or 24 mm) using the same procedure of modifying the horizontal location of each text block by half the difference. These offsets were intended to further increase the variability of difficulty for the perceptual task as they prevented the stimuli from being directly visually aligned with one another.

Two versions of each pairing were created: one in which the perceptual answer agreed with the conceptual answer (i.e., the correct answer to both decisions was the same) and one in which the perceptual answer disagreed with the conceptual answer. Counterbalancing of these two versions so that, for each participant, equal numbers of agree and disagree items were included, prevented participants from relying on the correct answer for one decision type to correctly answer the other decision type. The stimuli were piloted and iteratively modified so as to obtain about 60% accuracy for each decision type and dimension.

Note that for the perceptual task, the words themselves referred to objects that have a real world size, but the object's actual size was not relevant to the perceptual decision participants were asked to make. Nonetheless, it is possible that either the match, or mismatch, between the ''real world size'' of the referents and the correct answer for the visual display size of the words might influence performance on the perceptual decision trials, either facilitating or impeding decision making in a ''Stroop-like'' manner.

To address this issue, stimuli for a control condition were created. These stimuli used the same random word pairings described above. However, the letters and spaces within each text block were scrambled in such a way that it created a random letter string (See Figure 1, panel D). The letter string stimuli were otherwise identical to the stimuli using actual words. One half of the participants made perceptual decisions in response to the word stimuli, whereas the other half of the participants made perceptual decisions to the remixed (non-word) stimuli. Contrasting the decision accuracy and over/underconfidence measures for the intact words versus remixed word conditions allowed assessment of the possible contribution of salient, but irrelevant, semantic/conceptual information in perceptual decision making and confidence evaluation.

Participants

An initial group of 36 students from the University of Minnesota participated in the study in order to receive extra credit in one of their introductory courses. All participants were native speakers of English and reported

DOI: 10.1002/bdm

normal or corrected-to-normal vision. Five participants were replaced because they achieved less than 30% accuracy in at least one of the three dimensions (longer, smaller, or taller). We assumed that accuracy this low on a paired forced-choice task indicated a poor understanding of the instructions from the participant (e.g., choosing the larger rather than the smaller option, or responding to the items for the longer dimension based on the number of letters in each word rather than the length of the text blocks in *mm*). One additional participant was replaced because he/she failed to answer more than one third of the questions in a dimension. The final dataset thus included 36 (26 female) students; participants' ages ranged from 18 to 27 ($M = 20.66$) and they reported an average of 14.89 years of formal education.

Procedure

All participants were tested individually, in single experimental sessions of approximately 50 minutes. The experiment was conducted in two identical halves, one half containing perceptual decisions, the other conceptual decisions, the order of which was counterbalanced across participants. The stimuli were presented in a blocked form by decision type (perceptual vs. conceptual), dimension (smaller, taller, longer), and category (e.g., bird or dog). The orders of the dimensions and categories were also counterbalanced across participants. Three practice trials were given at the beginning of each dimensional block in order to allow participants to acclimate to each procedure.

Participants were instructed to choose the item from a pair of words that they believed to be correct, given the instructions for that block (e.g., select the word that visually *appears* to be the longest on the screen, or select the word for which the *referent* of that word is taller). They indicated their response using one of two keys on the keyboard (a small reminder card indicated response mappings). Using E-Prime software, the stimuli were displayed for 3.5 seconds each; participants were required to respond during this time. The time limit was imposed in order to increase the ease of potential transfer of the paradigm to the neuroimaging context (fMRI or ERPs). After each decision, participants performed a self-paced rating of their confidence in that decision on a six-point scale that ranged from 50 (i.e., complete guess) to 100% (i.e., complete certainty in the correctness of their response). As it was a two-alternative forced-choice task, confidence ratings of less than 50% were not allowed.

Once the participant had completed half of the experiment (i.e., all of the conceptual or perceptual decisions), they were asked to estimate the overall percentage of questions they believed they had answered correctly (PTPE), again using the same six-point scale of 50–100%. Next, they were asked to fill out a short demographic questionnaire. This brief activity was included in order to create temporal and cognitive separation between the two judgment types. Upon completion of the questionnaire, the participants continued with the second half of the experiment (i.e., conceptual or perceptual judgments, whichever they had not answered in the first half), again making confidence ratings after each decision and an overall confidence rating (PTPE) at the conclusion. After participants completed all the trials, they filled out a post-experimental questionnaire and were debriefed.

Results

Results are presented separately by analysis type, focusing on: (1) preliminary analyses of differences in decision accuracy between the different types of stimuli; (2) analyses of over/underconfidence and calibration to determine how closely participants' confidence matched their accuracy; (3) discrimination analyses to determine how well participants were able to use their levels of confidence to distinguish between items they answered correctly and those they answered incorrectly; (4) an inspection of the correlates of confidence; (5) item analyses exploring differences in accuracy and over/underconfidence for individual items; and (6) across task analyses, including examination of within-individual correlations on the key

DOI: 10.1002/bdm

dependent variables across the two decision types. Analyses were limited to questions for which the participants had chosen an answer within the time limit.

Preliminary analyses

Initial analyses were conducted to examine whether there was any evidence for a Stroop-like effect (i.e., the correct answer conceptually to an item biasing answering in the perceptual task or vice versa). In order to test for this effect, participants who had been given real word pairs were compared to participants who had been given scrambled words on the perceptual task. The two groups showed very similar levels of accuracy, confidence, and overconfidence (confidence minus accuracy) on these perceptual items, all $t(34) < .70$, all $ps > .50$. However, those shown remixed words had quicker reaction times (RT) ($M = 1823.51$) than did those shown real words ($M = 2048.76$), $t(34) = 2.09$, $p = .04$. Therefore, while the presence of a readable word slightly slowed the reactions of the participants in the perceptual task, it did not significantly affect their ability to correctly answer the items or their confidence ratings.

An absence of Stroop-like interference effects was further demonstrated by comparing those items for which the correct conceptual answer *agreed* with the perceptual answer to those items which *disagreed* with the perceptual answer. Analyses on accuracy and confidence showed no significant differences for perceptual decisions, both $t(17) = <1.5$, $p > .20$, or conceptual decisions, both $t(35) < .60$, $p > .60$. Thus, it can be likewise concluded that the perceptual differences in the word pairs did not significantly affect the participants' ability to correctly answer the conceptual items or their usage of the confidence ratings.

Next, analyses were conducted on accuracy so as to be able to compare confidence ratings under conditions of equal accuracy. A 2 (decision type) × 3 (dimension) repeated measures ANOVA indicated that there was a significant main effect of dimension, $F(2, 70) = 12.68$, $p < .001$. Contrasts indicated that accuracy on the taller dimension ($M = 63.74$) was significantly lower than accuracy on the longer ($M = 71.78$) and smaller ($M = 70.08$) dimensions. However, no main effect of decision type was found. Accuracy on the perceptual decisions ($M = 69.42$) was not significantly different from accuracy on the conceptual decisions ($M = 67.65$), $F(1, 35) = 1.90$, $p = .18$. In addition, no significant interaction effect was found, $F(2, 60) = 2.95$, $p > .05$. Therefore, given the similar levels of accuracy for perceptual and conceptual decisions, any significant differences in the confidence on the two decision types are unlikely to be due to variance in accuracy.

Analyses of under/overconfidence and calibration

We first focused on the measure of over/underconfidence (i.e., confidence minus accuracy). A 2 (decision type) × 3 (dimension) repeated measures ANOVA produced no significant main effect of dimension, $F(2, 70) = 2.38$, $p = .10$, and the interaction effect of decision type and dimension was also not significant, $F(2, 58) = 1.69$, $p = .20$. However, a significant main effect of decision type was detected, $F(1, 35) = 17.66$, $p < .001$. Numerical overconfidence was evidenced for both types of decisions but, contrary to predictions, significantly greater *over*confidence was found for the perceptual decisions ($M = 7.95$) as compared to the slight overconfidence for the conceptual decisions ($M = 1.93$). Overconfidence for perceptual decisions was found to be significantly greater than zero, $t(35) = 4.44$, $p < .001$. However, the overconfidence for conceptual decisions was not found to be significantly greater than zero, $t(35) = 1.26$, $p = .22$; therefore, participants were reasonably well calibrated for conceptual decisions. Furthermore, a similar result of *comparatively* greater overconfidence for perceptual decisions ($M = 1.41$) than for conceptual decisions ($M = -6.01$) was found when analyzing participants' PTPE data, $t(35) = 3.89$, $p < .001$.

Figure 2 graphically depicts differences in confidence and accuracy by dimension and decision type. From the graph, it can be seen (as established above) that overall accuracy is well matched across the two decision types. In addition, within each dimension, there is an indication of greater overconfidence on the perceptual

DOI: 10.1002/bdm

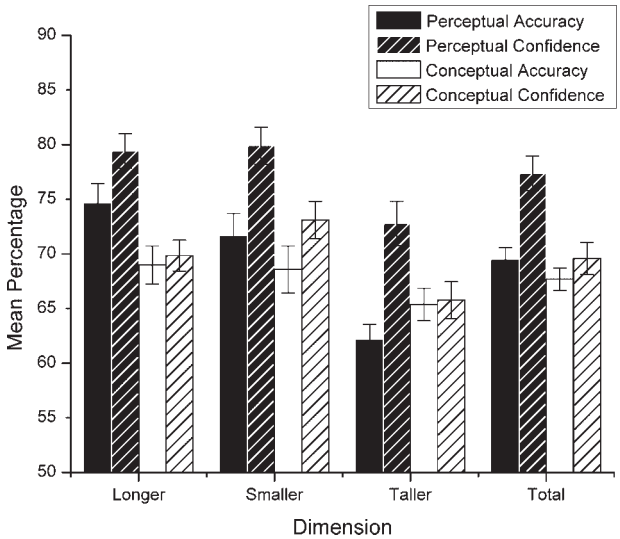

Figure 2. Confidence and accuracy by decision type (perceptual, conceptual) and dimension (longer, smaller, taller) for Experiment 1. Error bars indicate one standard error above and below the mean

task. This occurs regardless of relative accuracy for the two decision types. Specifically, on the longer dimension, perceptual accuracy is higher than conceptual accuracy; on the smaller dimension, the same (albeit less pronounced) perceptual advantage is seen. However, on the taller dimension, conceptual accuracy actually exceeds perceptual accuracy. Nevertheless, in all instances, there is greater overconfidence for the perceptual than for the conceptual decisions.

The finding that greater overconfidence on perceptual than on conceptual items persisted even in the presence of a numerical conceptual accuracy advantage (taller) points to the robustness of this unexpected effect. Nonetheless, there was still concern that differences in accuracy (possibly at the dimensional or individual level) may have been driving the observed main effect.

To test for this possibility, further analyses were performed on groups of the perceptual and conceptual items in which each participant's performance was exactly matched. For each participant, categories within a given dimension (longer, taller, smaller) were chosen in which the participant had achieved the exact same level of accuracy for perceptual and conceptual decisions. For instance, if a participant had attained 70% accuracy in the perceptual bird category and 70% accuracy in the conceptual car category, bird and car were selected for his/her smaller categories. To qualify, the categories must have contained responses to at least nine items (out of a possible 10). Adherence to this criteria and the lack of exact accuracy matches for some participants on some dimensions led to lower *N*s for these analyses. However, the same pattern of significant results was obtained. Because of the procedure, accuracy for both decision types was perfectly matched (see Table 1 for means). Nonetheless, significantly greater confidence was still found for perceptual decisions than for conceptual decisions in all three dimensions, all ts > 2.5, ps $< .05$.

Calibration curves were used to examine how well participants were calibrated at different levels of confidence. Figure 3 illustrates that the pattern of calibration for conceptual and perceptual decisions is similar, and calibration for both types of decision is quite good at the lower levels of accuracy. However, overconfidence begins to be displayed at higher confidence levels, with greater overconfidence on the perceptual decisions versus conceptual decisions consistently seen at 70% confidence and above.

Specific analyses were also performed to test for the presence of hard-easy effects on both perceptual and conceptual decisions. Difficulty for the perceptual decisions was defined by the five levels of differences

DOI: 10.1002/bdm

Table 1. Mean accuracy, confidence, and over/underconfidence for the paired-matching comparisons for Experiments 1, 2, and 3

	Dimension		
	Smaller/larger	Longer	Taller
Experiment 1			
N	19	23	22
Accuracy	75.45	74.44	65.00
Perceptual			
Confidence	79.89	80.44	75.33
Over/underconfidence	*4.44*	*6.00**	*10.33***
Conceptual			
Confidence	73.78	72.09	65.63
Over/underconfidence	*−1.67*	*−2.35*	*0.63*
Experiment 2			
N	25	28	30
Accuracy	73.20	77.86	71.00
Perceptual			
Confidence	82.32	78.07	74.13
Over/underconfidence	*9.12***	*0.21*	*3.13*
Conceptual			
Confidence	75.84	72.14	67.33
Over/underconfidence	*2.64*	*−5.72***	*−3.67*
Experiment 3			
N	23	23	26
Accuracy	72.90	73.04	69.02
Perceptual			
Confidence	84.82	78.09	71.27
Over/underconfidence	*11.92***	*5.05*	*2.25*
Conceptual			
Confidence	73.91	70.35	64.71
Over/underconfidence	*1.01*	*−2.69*	*−4.31*

*Denotes over/underconfidence significantly different from zero at the .05 level.
**At the .01 level.

between the text block sizes described in the Materials section. As conceptual items were designed by a random-pairing technique, they were not initially assigned to difficulty levels. In order to establish varying levels of difficulty similar to the perceptual tasks, conceptual items on each list were first ranked by the difference in size between the two real-world objects in each pair. Then, the four items with the greatest difference in size were assigned to the easiest difficulty level (1), the next 4 to the difficulty level 2, and so on to produce difficulty rankings for conceptual items on the same 1–5 scale as used for the perceptual decisions. ANOVAs on accuracy as a function of difficulty level, and also the linear pattern in the means (see Table 2), support the validity of these difficulty level assignments for both perceptual, $F(4, 140) = 59.23$, $p < .001$, and conceptual decisions, $F(4, 140) = 28.68$, $p < .001$.

A 2×5 ANOVA on the level of over/underconfidence using decision type and difficulty as factors yielded a significant main effect of difficulty, $F(4, 140) = 45.41$, $p < .001$. In addition, the interaction effect was not found to be significant, $F < 1$. Therefore, the strong effect of difficulty was not significantly modulated by decision type. Means obtained in these analyses (see Table 2) illustrate the typical pattern found in hard-easy effects for *both* decision types, with underconfidence for the easiest decisions and overconfidence for the more difficult decisions.

DOI: 10.1002/bdm

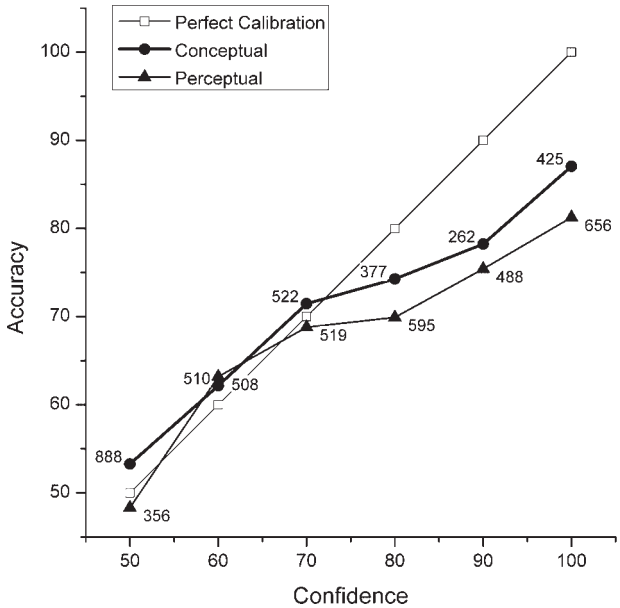

Figure 3. Perceptual (thin) and conceptual (thick) calibration curves for Experiment 1. Each point is labeled with the number of responses represented

Discrimination analyses

The measures reported above examine how well participants' confidence relates to their accuracy. In contrast, discrimination (sometimes called resolution) measures examine how well participants are able to use their confidence ratings to distinguish between items that they answer correctly and items that they answer incorrectly.[5] One such measure is slope. Slope is equivalent to confidence on correct items minus confidence on incorrect items. Thus, larger slopes indicate that the participants had greater confidence in their correct responses than their incorrect responses, consequently indicating better discrimination. Participants demonstrated modestly but significantly lower slope scores for perceptual decisions ($M = 7.20$) than for conceptual decisions ($M = 8.91$), $t(35) = 2.15$, $p = .04$. The results of additional discrimination measures can be found in Table 3.

Correlates of confidence

It is assumed that in order for confidence to be useful, it is correlated with accuracy. The gamma statistic and Spearman's rho (both appropriate for use when correlating ordinal variables) were calculated on the raw data to test for this expected correlation. For perceptual decisions, confidence and accuracy were significantly correlated, $\gamma = .29$, $\rho = .20$, both ps $< .001$. Consistent with the slope results, a slightly stronger significant positive relationship between confidence and accuracy was seen for conceptual decisions, $\gamma = .38$, $\rho = .25$, both ps $< .001$.

Next, as RT data have been negatively correlated with confidence in previous research (classically: Henmon, 1911; Johnson, 1939; Kellogg, 1931; Volkmann, 1934; more recently: Baranski & Petrusic, 1994, 2001; Petrusic & Baranski, 2003; Robinson, Johnson, & Herndon, 1997), response times for the primary

[5]Measures of discrimination also depend to some extent on the difficulty distribution of the stimuli, with higher discrimination scores possible with larger ranges of difficulty.

DOI: 10.1002/bdm

Table 2. Mean confidence, accuracy, and over/underconfidence in relation to task difficulty (hard-easy effects) for Experiments 1, 2, and 3

	Level of ease—1 = easiest, 5 = most difficult				
	1	2	3	4	5
Experiment 1					
Perceptual					
Confidence	82.58	79.69	76.26	74.93	73.14
Accuracy	85.34	79.57	67.30	58.92	54.97
Over/underconfidence	*− 2.76*	*0.12*	*8.96*	*16.01*	*18.17*
Conceptual					
Confidence	74.94	68.92	67.94	69.75	66.40
Accuracy	81.51	73.68	64.37	63.84	55.80
Over/underconfidence	*− 6.57*	*− 4.76*	*3.57*	*5.91*	*10.60*
Experiment 2					
Perceptual					
Confidence	84.98	80.96	75.65	75.28	72.90
Accuracy	90.89	87.50	70.99	62.35	55.09
Over/underconfidence	*− 5.91*	*− 6.54*	*4.66*	*12.93*	*17.81*
Conceptual					
Confidence	77.32	71.43	68.40	70.31	68.07
Accuracy	81.49	71.91	63.70	65.75	55.98
Over/underconfidence	*− 4.17*	*− 0.48*	*4.70*	*4.56*	*12.09*
Experiment 3					
Perceptual					
Confidence	82.28	78.95	75.45	73.99	73.11
Accuracy	87.41	83.27	69.88	61.87	57.13
Over/underconfidence	*− 5.13*	*− 4.32*	*5.57*	*12.12*	*15.98*
Conceptual					
Confidence	74.95	69.15	67.03	67.87	64.86
Accuracy	80.01	71.10	60.75	66.32	54.71
Over/underconfidence	*− 5.06*	*− 1.95*	*6.28*	*1.55*	*10.15*

decision were correlated with confidence. While linear correlations between RT and confidence for each participant were not significant for either decision type (P: $r = .17$, $p = .33$; C: $r = .07$, $p = .67$), significant curvilinear relationships between the variables were detected in the raw data (P: $\varepsilon = .34$, $p < .001$; C: $\varepsilon = .23$, $p < .001$). Participants chose higher confidence ratings when they were able to make their decisions more quickly, but made many of their ''complete guesses'' quickly as well. In addition, a 2×3 ANOVA on RT produced similar results to the ANOVA done on overconfidence. There was again a significant main effect of decision type, $F(1, 35) = 104.42$, $p < .001$; RTs were significantly faster when participants made perceptual decisions ($M = 1936.14$ milliseconds) than when they made conceptual decisions ($M = 2410.55$ milliseconds).

The differences in RT between the two decision types, together with the fixed time limit of 3.5 seconds for all trials, implied that participants may have missed more items on the conceptual decisions than the perceptual decisions. A 2×3 ANOVA on the number of responses bore this out; a significant main effect of

DOI: 10.1002/bdm

Table 3. Mean slope, Confidence Judgment Accuracy Quotient (CAQ), and Brier score for Experiments 1, 2, and 3

	Decision type		
	Perceptual	Conceptual	Across-task correlation
Experiment 1			
Slope	7.20*	8.91*	.57**
CAQ	.540	.603	.54**
Brier score	.216	.209	.59**
Experiment 2			
Slope	8.44	9.00	.18
CAQ	.636	.595	.14
Brier score	.191**	.211**	.49**
Experiment 3			
Slope	8.02	8.32	.03
CAQ	.583	.544	−.03
Brier score	.200*	.218*	.12

Note: The final column shows the correlation of these measures across the two decision types.
*Denotes significant mean differences in discrimination for the perceptual versus conceptual decisions, or correlations at the .05 level.
**At the .01 level. See Keren (1991) for descriptions of the CAQ and Brier score measures.

decision type was found, $F(1, 35) = 28.27$, $p < .001$. On average, participants responded to a modestly but significantly greater number of the perceptual items ($M = 28.91$) than to the conceptual items ($M = 27.63$), both out of a possible 30 items. This issue will be addressed further in Experiments 2 and 3.

Furthermore, correlations between participants' response rates and confidence measures produced interesting results. Response rates were more strongly related to PTPE confidence assessments (P: $r = .48$, $p = .003$; C: $r = .24$, $p = .16$) than to item-by-item confidence (P: $r = .15$, $p = .39$; C: $r = .10$, $p = .58$) for both decision types. In addition, replicating past research, confidence indicated by the PTPE (P: $M = 70.83$; C: $M = 61.67$) was significantly lower than confidence indicated by the average of the item-by-item ratings (P: $M = 77.37$; C: $M = 69.58$) for both decision types, both $t(35) > 5$, $p < .001$.[6]

Item analyses

Item analyses were limited to items that, combining across all participants, had been responded to at least eight times. This provided a pool of 180 items for each decision type. In order to explore the possible usage of heuristics, we focused on items that showed particularly low levels of accuracy (i.e., those below the chance performance level of 50%). The key outcomes of these analyses, shown separately by decision type and experiment, can be found in Table 4. On all measures, the results were consistent with a greater usage of heuristics on conceptual decisions than perceptual decisions. There were more items in the conceptual task on which participants, on average, achieved less than 50% and less than 40% accuracy, and the distribution of item accuracies was more negatively skewed for the conceptual task. These outcomes suggest that, particularly for the conceptual task, participants were likely depending on a heuristic to answer these questions, which produced a consistently inaccurate response. Furthermore, on the conceptual items with accuracy <40%, the average confidence rating was 64.30%, indicating that participants felt they had some information on which to base their decision. For instance, for the country stimulus MOROCCO–GERMANY, only 15.38% (2/13) participants correctly chose Germany as the smaller country. However, they averaged

[6]This finding was also replicated in both Experiments 2 and 3, all $t > 2.9$, $p < .01$.

DOI: 10.1002/bdm

Table 4. Results for the item analyses for Experiments 1, 2, and 3, and for all three experiments combined

		Experiment			
Decision type	Measure	1	2	3	Combined
Perceptual	$N < 50\%$ accuracy	20	16	17	17
	$N < 40\%$ accuracy	10	8	4	3
	Lowest accuracy (%)	16.67	16.67	31.25	36.95
	Distribution skew	−.57	−.47	−.26	−.15
	Mean difficulty	4.09	4.25	4.50	4.33
Conceptual	$N < 50\%$ accuracy	43	40	43	45
	$N < 40\%$ accuracy	31	30	28	28
	Lowest accuracy (%)	7.14	0.00	8.33	10.87
	Distribution skew	−.70	−.62	−.59	−.60
	Mean difficulty	3.87	3.93	3.75	3.71

Note: Shown are the number of items with low-accuracy rates, the accuracy percentage for the lowest scoring item, the degree of skew of the accuracy distributions, and the average difficulty of the low-accuracy items.
Note: The Combined column presents item analyses when collapsing all data over the three experiments (not the average of the experiments taken individually).

74.62% confidence. For the perceptual task, it may be suspected that participants were relying on some heuristic to answer these items as well. However, given the small percentage of perceptual items which fell into the <40% accuracy category, it seems likely that these items simply represent the bottom of the normal distribution of accuracy.

Across task analyses

Overall, participants demonstrated across task consistency in their confidence assessments. Participants' confidence ratings for perceptual and conceptual decisions were significantly correlated ($r = .68$, $p < .001$). A significant correlation between participants' level of over/underconfidence on the conceptual task and that on the perceptual task ($r = .65$, $p < .001$) was also found. Individual consistency across decision type also was seen for the discrimination measure, slope ($r = .57$, $p < .001$).

EXPERIMENT 2

Previous studies (e.g. Dawes, 1980; Baranski & Petrusic, 1994; Stankov, 1998; Winman & Juslin, 1993) did not often employ a time limit on participants' decisions. Therefore, this difference in time pressure may have led to our divergent findings. For instance, given that the decision response times were, on average, significantly longer for the conceptual than for the perceptual task, it might be argued that the lack of significant overconfidence for the conceptual decision task arose because participants experienced greater subjective time pressure on this task than on the perceptual task. Such a between-task difference in perceived time pressure may have had one of two effects.

First, if participants sometimes felt that they did not as fully consider their answer as they might have done, then the perceived time pressure may have led to a global reduction in confidence for the conceptual task. Second, and perhaps more speculatively, it may have led participants to focus their attention successively or serially on the decision task versus the confidence rating to a greater degree than was true for the perceptual task. Stated differently, given that there was, on average, less ''leeway'' between the time required to make the conceptual decisions and the response limit for the conceptual task than for the perceptual task,

DOI: 10.1002/bdm

participants may have been less likely to make their conceptual decisions in parallel with, or in conjunction with, their confidence ratings in those decisions.

Baranski and Petrusic (1998) have suggested that, for decisions made under speed stress, as seems true for our conceptual decisions, ''confidence is determined *postdecisionally* and involves a memory-based, computational algorithm. This strategy frees the primary decision of processing time and permits the accurate diagnosis of decision errors'' (p 929, italics added). By contrast, for decisions made under accuracy stress (as seems more true for our perceptual decisions), ''the determination of confidence is initiated, or can even be completed, *during the primary decision process*. This strategy permits confidence to be used in the adaptive regulation of the decisional parameters during the decision process but *yields poorer diagnosticity of errors* when they occur'' (p 929, italics added). Therefore, Experiment 2 was designed to determine if the same pattern of results would be obtained when participants were not under time pressure.

Method

All procedures and stimuli were held constant with the exception of two changes: (1) participants were allowed an unlimited amount of time to make each decision and (2) the ''smaller'' dimension was changed to ''larger'' so that for all three decision dimensions—longer, taller, and larger—participants were instructed to choose the option with the greater magnitude.

Participants

A new complement of 36 students was recruited for this study from the University of Minnesota undergraduate Research Experience Program pool, which provides participants with extra credit in one of their introductory level psychology courses. The sample included 26 female students; participants' ages ranged from 18 to 30 ($M = 20.39$) and they had an average of 13.72 years of formal education.

Results

Preliminary analyses

Initial checks for Stroop-like effects again produced no significant effects. There was no effect of condition (agree vs. disagree) on measures of accuracy, confidence, or over/underconfidence (all Fs < 1). Effects of word versus non-word also were found to be non-significant, largest $t(34) = 1.50$, $p = .14$ for RT.

Significant effects of dimension and decision type on accuracy, however, were found. An ANOVA on accuracy by decision type and dimension produced a significant main effect of decision type, $F(1, 35) = 22.52$, $p < .001$; participants were more accurate on their perceptual decisions ($M = 73.36$) than on their conceptual decisions ($M = 67.56$). As before, there was also a significant main effect of dimension, $F(2, 70) = 25.87$, $p < .001$, with lower accuracy for the taller dimension ($M = 64.95$) than for the larger ($M = 72.18$) or longer ($M = 74.26$) dimensions. Furthermore, the interaction effect was significant, $F(2, 70) = 3.20$, $p = .047$. This was driven by the greater disparity in accuracy between the taller dimension and the other two dimensions in the perceptual task (see Figure 4).

Analyses of under/overconfidence and calibration

Using the overall results (without matching on decision accuracy), the ANOVA on over/underconfidence including decision type and dimension as within-subject factors showed no main effect of decision type, $F < 1$; participants' overconfidence did not differ significantly between their perceptual ($M = 3.49$) and conceptual ($M = 4.53$) decisions. There was a significant main effect of dimension, $F(2, 70) = 7.22$, $p = .001$, with much less overconfidence in the longer dimension ($M = .72$) than in the larger ($M = 6.39$) and taller

DOI: 10.1002/bdm

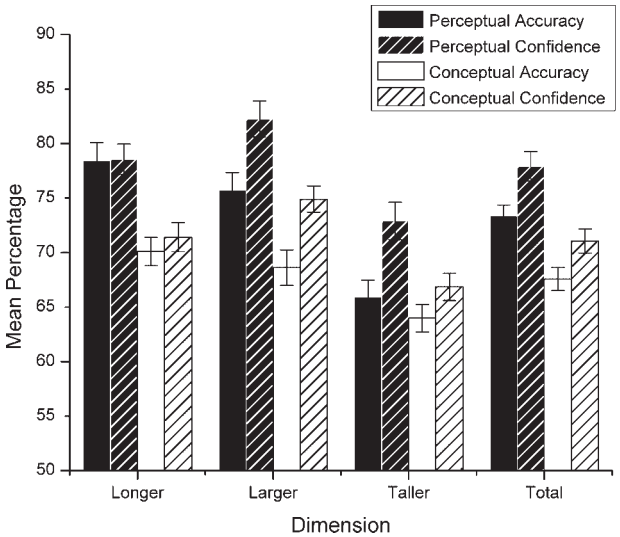

Figure 4. Confidence and accuracy by decision type (perceptual, conceptual) and dimension (longer, larger, taller) for Experiment 2. Error bars indicate one standard error above and below the mean

($M = 4.92$) dimensions. The interaction effect was not significant, $F(2, 70) = 1.64$, $p = .20$. Data from the PTPE showed similar results, with average PTPE confidence only 0.06% lower than accuracy for perceptual decisions, and 3.12% lower than accuracy for conceptual decisions, $t(35) = 1.44$, $p = .16$.

However, the significant accuracy differences between the two decision types prompted us to implement our paired-matching technique on these data, selecting for additional analyses those perceptual and conceptual decision blocks for each participant that were precisely matched in level of accuracy. These analyses produced results that replicated the findings reported for Experiment 1. *When perfectly matched for accuracy*, perceptual confidence significantly exceeded conceptual confidence in each dimension, all ts > 2.5, ps $\leq .01$ (see Table 1 for means).

Figure 5 illustrates that the pattern of calibration for conceptual and perceptual decisions is somewhat similar. At low levels of accuracy, calibration is good for perceptual decisions, but slight underconfidence is seen for conceptual decisions. However, overconfidence begins to be displayed at higher confidence levels for both decision types.

The presence of hard-easy effects on both perceptual and conceptual decisions was again tested for. As in Experiment 1, the 2×5 ANOVA on the level of over/underconfidence using decision type and difficulty as factors produced a significant main effect of difficulty, $F(4, 140) = 57.84$, $p < .001$. In contrast, the interaction effect was found to be significant in this experiment, $F(4, 140) = 4.27$, $p = .003$; however, this was likely driven by the lack of increase of overconfidence at difficulty level 4 (which corresponded to a lack of change in accuracy) for conceptual decisions compared to the steady increases in overconfidence for perceptual decisions. Notably, the means (see Table 2) again illustrated patterns consistent with hard-easy effects for *both* decision types.

Discrimination analyses

Slope scores were, on average, slightly but not significantly lower for perceptual decisions ($M = 8.44$) than for conceptual decisions ($M = 9.00$), $t(35) < 1$.

DOI: 10.1002/bdm

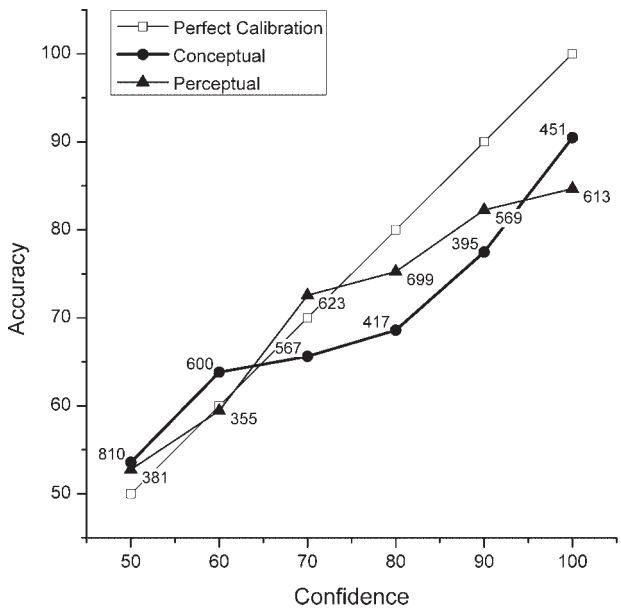

Figure 5. Perceptual (thin) and conceptual (thick) calibration curves for Experiment 2. Each point is labeled with the number of responses represented

Correlates of confidence

As before, for perceptual decisions, confidence and accuracy were modestly but significantly correlated, $\gamma = .34$, $\rho = .23$, both ps $< .001$, as were the same correlations for conceptual decisions, $\gamma = .35$, $\rho = .24$, both ps $< .001$.

Again, significant curvilinear relationships between RT and confidence were detected in the raw data (P: $\varepsilon = .23$, $p < .001$; C: $\varepsilon = .16$, $p < .001$). Generally, participants chose higher confidence ratings when they were able to make their decisions more quickly, but made many of their ''complete guesses'' quickly as well. The 2×3 ANOVA effects were similar to those in Experiment 1; however, they did not mirror the results of the ANOVA done on overconfidence. Dimension did not show a significant effect on RT, $F < 1$, whereas there was again a significant main effect of decision type, $F(1, 35) = 24.93$, $p < .001$. RTs were significantly faster for perceptual decisions ($M = 3738.77$ milliseconds) than for conceptual decisions ($M = 4500.18$ milliseconds).

Item analyses

The patterns observed for individual items that yielded low levels of accuracy (see Table 4) were similar to the patterns found in Experiment 1; the conceptual task produced more inaccurate items, a more negatively skewed distribution, and a more extreme lowest accuracy than did the perceptual task.

Across task analyses

Participants' confidence ratings and levels of over/underconfidence for perceptual and conceptual decisions were again significantly correlated ($r = .40$, $p = .017$, $r = .49$, $p = .002$, respectively). However, individual consistency across decision type was not seen for discrimination, that is, slope ($r = .18$, $p = .30$).

DOI: 10.1002/bdm

EXPERIMENT 3

The outcomes from Experiment 2 argue against a time-limit account of the findings we have obtained; in both experiments, after exactly matching for accuracy level, there was significantly greater confidence for *perceptual* than for *conceptual* decisions. This result was observed even though, in Experiment 2, participants were allowed to self-pace their decisions. Our final experiment was conducted in order to further replicate our findings and to further examine the influence of time pressure on confidence evaluations for conceptual versus perceptual decisions. For this purpose, we used the empirical response time data obtained in Experiment 2, which involved no time stress, to establish response time limits that would create approximately equivalent time pressure for each type of decision.

Method

Using the RT data from Experiment 2 (where participants made decisions under self-paced conditions), time limits were set separately for the three types of decisions (conceptual, perceptual words, and perceptual non-words) in order to attempt to equalize the subjective ''time pressure'' participants felt while making their decisions. Specifically, we aimed to arrive at decision response time limits that would place participants under some degree of time stress, but that (a) would not differentially influence either the conceptual or perceptual decisions, and (b) would still allow decisions to be made for the majority of items. To meet these aims, we opted for a within-task time limit that corresponded to the RT at the 75th percentile under self-paced conditions. In the calculation of the RT limits for Experiment 3, Experiment 2 RT outliers were first excluded (excessively slow responses, including 175 and 202 items for the conceptual and perceptual tasks, respectively). Then, the 75th percentile values for the three dimensions for each decision type were obtained. The values for the three dimensions (larger, longer, and taller, respectively) were then averaged for each decision type and rounded up to the nearest 100 milliseconds. This procedure yielded the following response time limits:

$$\text{Conceptual} = (5245.5 + 4949.5 + 5153)/3 = 5116;\ \text{rounded} = 5200\ \text{milliseconds}.$$

$$\text{Perceptual words} = (4055.75 + 4921 + 4853)/3 = 4610;\ \text{rounded} = 4700\ \text{milliseconds}.$$

$$\text{Perceptual non-words} = (3556 + 3867 + 3529.5)/3 = 3651;\ \text{rounded} = 3700\ \text{milliseconds}.$$

Except for the imposition of these new (empirically determined) time limits, the methods and procedure were identical to those in Experiment 2.

Participants

A further 36 students from the University of Minnesota Research Experience Program who had not participated in either of the previous experiments were recruited. Participants (27 female) ranged in age from 18 to 23 ($M = 19.11$); they had an average of 13.32 years of formal education.

Results

Preliminary analyses

Given that this experiment was designed to re-introduce time pressure without differentially affecting the number of decisions that were made (response rates), we first performed a 2 (decision type) $\times$ 3 (dimension) repeated measures ANOVA on response rates. No significant effects were found (all Fs < 1); thus, response rates were equivalent for each dimension and decision type (all cell means approximately 29.4 out of 30 items).

DOI: 10.1002/bdm

Next, checks for Stroop-like effects were performed. No significant effects of condition (agree vs. disagree) were found, all Fs < 1.0, ps $> .15$. Most effects of word versus non-word were also found to be non-significant, largest $t(34) = 1.56$, $p = .13$ for RT. Unexpectedly, a significant effect of word/non-word was found for the PTPE, $t(34) = 2.55$, $p = .015$. Those making decisions about real words ($M = 74.44$) averaged higher PTPE ratings than those who were shown remixed words ($M = 66.67$).[7]

Significant task effects on accuracy, however, were again detected. The ANOVA on accuracy by decision type and dimension produced a significant main effect of decision type, $F(1, 35) = 19.85$, $p < .001$. Participants were more accurate on their perceptual decisions ($M = 72.03$) than on their conceptual decisions ($M = 66.28$). The significant main effect of dimension, $F(2, 70) = 10.32$, $p < .001$, again demonstrated that items in the taller dimension ($M = 65.12$) were responded to less accurately than those in the larger ($M = 71.49$) or longer ($M = 70.84$) dimensions. The marginally significant interaction effect, $F(2, 70) = 3.08$, $p = .052$, was again caused by the greater disparity in accuracy between the taller dimension and the other two dimensions of the perceptual task (see Figure 6).

Analyses of under/overconfidence and calibration

As in Experiment 2, a 2 (decision type) $\times$ 3 (dimension) ANOVA on the overall levels of over/underconfidence (before matching on decision accuracy) showed *no* significant main effect of decision type, $F(1, 35) = 1.94$, $p = .17$. Overall, participants were only slightly more overconfident in their perceptual decisions ($M = 4.74$) than in their conceptual decisions ($M = 2.48$). The main effect of dimension was significant, $F(2, 70) = 3.36$, $p = .04$, with more overconfidence in the larger ($M = 6.15$) dimension than in the longer ($M = 2.89$) and taller ($M = 1.80$) dimensions. The interaction effect was not significant, $F < 1$.

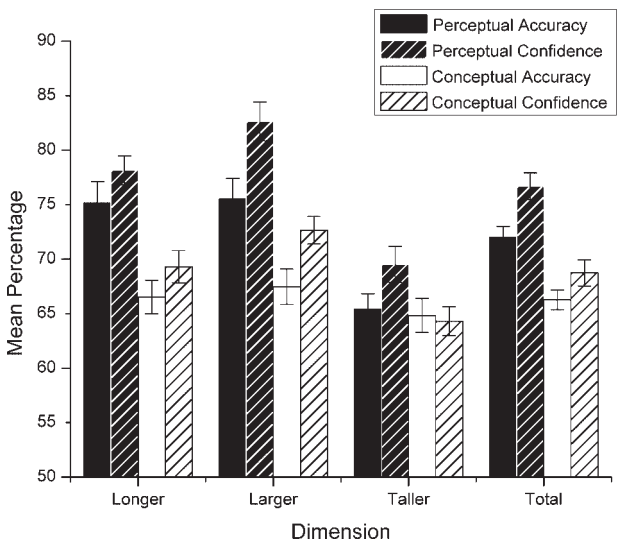

Figure 6. Confidence and accuracy by decision type (perceptual, conceptual) and dimension (longer, larger, taller) for Experiment 3. Error bars indicate one standard error above and below the mean

[7]This unexpected PTPE effect did not seem to be attributable to outliers, but instead seems likely to be a Type 1 error as this pattern was not supported in Experiment 1: real = 70.56, remix = 71.11, $t(34) < 1$, or in Experiment 2: real = 74.44, remix = 72.22, $t(34) < 1$.

DOI: 10.1002/bdm

Similarly, data from the PTPE showed only slightly less retrospective underestimation relative to actual performance for perceptual decisions (1.44% lower than accuracy) than for conceptual decisions (3.77% lower than accuracy), $t(35) = 1.11$, $p = .27$.

Nevertheless, the observation of significant accuracy differences between the two decision types led us to run our paired-matching technique on these data as well. Results again echoed and replicated our previous findings—*when perfectly matched for accuracy*, perceptual confidence surpassed conceptual confidence in each dimension, all ts > 2.5, all ps $< .02$ (see Table 1 for means).

Figure 7 illustrates similar patterns of calibration for conceptual and perceptual decisions. At low levels of accuracy, calibration is good for both decision types, with slightly greater levels of overconfidence becoming evident at higher accuracy levels.

Hard-easy effects were again examined. As in Experiments 1 and 2, a significant main effect of difficulty was found, $F(4, 140) = 39.87$, $p < .001$. The interaction effect was also significant in this experiment, $F(4, 140) = 5.28$, $p = .001$, driven by the dip in overconfidence at level 4 for conceptual decisions (which corresponded to the jump in accuracy) as compared to the steady increase in overconfidence for perceptual decisions. Yet, notably, the means (see Table 2) again demonstrated patterns consistent with hard-easy effects for *both* decision types.

Discrimination analyses

Slope scores were again found to be slightly, but not significantly, lower for perceptual decisions ($M = 8.02$) than for conceptual decisions ($M = 8.32$), $t < 1$.

Correlates of confidence

In line with findings from Experiments 1 and 2, confidence and accuracy were modestly but significantly correlated for both perceptual decisions ($\gamma = .31$, $\rho = .21$, both ps $< .001$) and conceptual decisions ($\gamma = .33$,

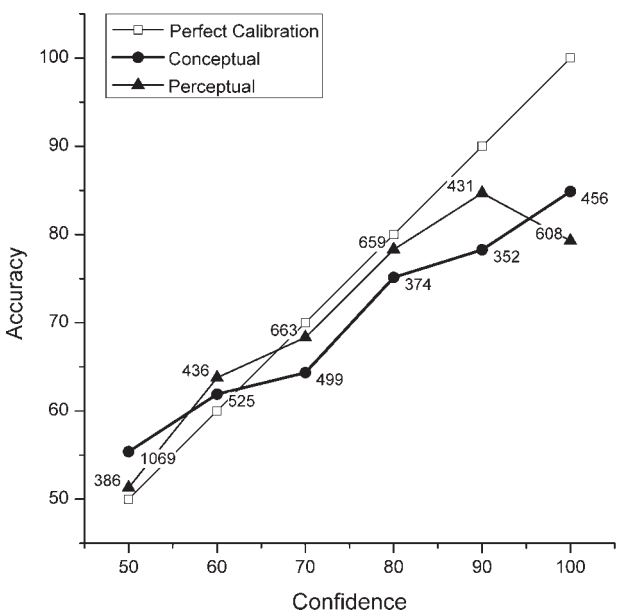

Figure 7. Perceptual (thin) and conceptual (thick) calibration curves for Experiment 3. Each point is labeled with the number of responses represented

DOI: 10.1002/bdm

$\rho = .22$, both $ps < .001$). In addition, once again, significant curvilinear relationships between confidence ratings and RT were detected in the raw data (P: $\varepsilon = .27$, $p < .001$; C: $\varepsilon = .21$, $p < .001$), reflecting faster responding for items when participants were highly confident and also when they were very non-confident.

Item analyses

The patterns of low accuracy items (see Table 4) mirrored those found in Experiments 1 and 2, again showing more low accuracy items for conceptual than for perceptual decisions. In addition, there was considerably greater *across-experiment* consistency in the items that were most often responded to inaccurately for the conceptual items. Of the 28 low accuracy conceptual items in this experiment, 22 were items on which participants in both or either Experiments 1 and 2 also had accuracies of less than 40%, and an additional 4 of the low-accuracy items were items which in either or both of Experiments 1 or 2 had accuracies of less than or equal to 50%. These data further support the notion that participants were relying on heuristics mainly on the conceptual task, which produced inaccurate responses to particular items consistently across participants and even across experiments. In contrast, only one of the four low-accuracy perceptual items was an item on which participants in *either* Experiment 1 or Experiment 2 also had an accuracy of less than 40%. Lastly, data combining across experiments further illustrate a likely greater use of heuristics for the conceptual decisions with a strengthening of the patterns demonstrated in each separate experiment (see Table 4).

Across task analyses

Participants' confidence ratings and levels of over/underconfidence for perceptual and conceptual decisions were again significantly correlated ($r = .72$, $p < .001$, $r = .44$, $p = .007$, respectively). However, individual consistency across decision type was not seen for slope ($r = .03$, *n.s.*).

GENERAL DISCUSSION

Together, these three experiments have yielded several key findings. Focusing on the findings that were most consistent across the three experiments, we found that: (1) contrary to predictions, whenever accuracy was perfectly matched, overconfidence on the item-by-item confidence measure for perceptual decisions always robustly exceeded confidence for conceptual decisions (average effect size $d = 1.36$ for the three experiments). Furthermore, (2) measures of participants' confidence and over/underconfidence were solidly, and significantly, correlated across the two decision types each time (across experiment average Pearson correlation using the Fisher *Z-r* transformation (Rosenthal & Rosnow, 1991): $r = .62$ and $r = .53$, respectively).

In addition, (3) in agreement with predictions, examination of performance on individual items showed that a consistently greater number of inaccurate items were found among the conceptual decisions than in perceptual decisions. Furthermore, (4) confidence ratings were consistently found to significantly correlate with accuracy (average across the three experiments: $\gamma = .33$, $\rho = .23$) and also (5) with RT, with the latter relation reflecting faster responding for items on which individuals were either highly confident, or not at all confident (average ε across the three experiments $= .24$). Also, (6) in agreement with predictions, global retrospective estimates of performance using PTPE showed significantly lower levels of overconfidence than did the average of the item-by-item ratings. Lastly, (7) participants' levels of discrimination showed less consistency, tending to be either equal across decision types or greater for the conceptual decisions and only significantly correlating across tasks in Experiment 1. We will discuss each of these main findings in turn.

DOI: 10.1002/bdm

Greater confidence and overconfidence in perceptual than in conceptual decisions

In each experiment, we were able to equalize accuracy levels for the perceptual and conceptual decisions, thus allowing us to make stronger claims about differences in the relationship between confidence and accuracy for the two decision types. Further, our results indicated that our participants were not strongly influenced by any ''Stroop-like'' interference, arising either from the inconsistent semantic content of the words for the perceptual judgments, or from the match or mismatch between the conceptual answers and the perceptual answers for particular items. However, despite these matched decision accuracy levels (achieved overall in Experiment 1, and through our pair-wise blocked matching procedure in all three experiments), the hypothesized result of overconfidence for the conceptual decisions and underconfidence for the perceptual decisions was not obtained. Indeed, the opposite result was obtained, with *slight to significant overconfidence for the perceptual decisions* and either *closely matched confidence and accuracy or underconfidence for the conceptual decisions*.

Our results clearly contradict Bjorkman et al.'s (1993) subjective distance theory of confidence assessment for perceptual tasks. We found in a number of cases that participants demonstrated significant *over*confidence for perceptual decisions, whereas this model holds that underconfidence will always be present for perceptual tasks. In addition, there was consistent evidence for hard-easy effects[8] on the perceptual tasks (as well as the conceptual ones), which was not predicted by the subjective distance theory. Our findings also challenge the sensory sampling model proposed by Juslin and Olsson (1997), which stresses an increased likelihood of underconfidence for perceptual tasks.

It may be suggested that the overconfidence observed for our perceptual decisions resulted because our perceptual stimuli were not as ''purely'' perceptual as those typically used (i.e., a sensory discrimination task dominated by Thurstonian error; see footnote 4). However, our results also replicate and extend earlier findings by researchers Baranski and Petrusic (1994, 1995, 1999; Petrusic & Baranski, 1997; cf., Bar-tal et al., 2001) in which clear overconfidence was reported in more ''pure'' perceptual tasks. Taken together, this provides strong evidence that while underconfidence may be a common finding for the typical perceptual decision tasks used (see Juslin, Olsson, & Winman, 1998) underconfidence is not the inevitable outcome for all perceptual decisions.

The results of these experiments also are inconsistent with previous findings of high levels of overconfidence for conceptual decisions (e.g., Adams & Adams, 1961; Crawford & Stankov, 1996; Fischhoff et al., 1977; Lichtenstein et al., 1982; Stankov, 1998); none of our analyses yielded significant overconfidence for the conceptual decisions. The generally close correspondence between overall levels of confidence and overall accuracy shown for conceptual decisions in our study may be due in part to the design of our stimuli. As Gigerenzer et al. (1991) illustrated, overconfidence can be eliminated or greatly attenuated when using randomly paired stimuli (as we did) rather than experimenter-selected items (which may be unintentionally biased toward the inclusion of misleading or surprising items). However, as can be seen from our item analyses, it is evident that our conceptual tasks nonetheless included a number of ''misleading'' items (see More Items were Consistently Wrong for Conceptual than for Perceptual Decisions). In addition, Brenner et al. (1996) have shown that high levels of overconfidence can still be seen when using randomly selected items. Taken together, these points suggest that it is unlikely that the lack of overconfidence seen for the conceptual decisions in our experiments was due solely to our method of stimuli selection. It is more likely that the underconfidence or near match between accuracy and confidence for conceptual decisions observed in our studies may be related to the ''ease'' of our task. The overall accuracy for conceptual decisions across experiments was 67.13%, not what one would typically think of as an ''easy'' task. However, the work of Juslin et al. (1998) summarizing the relationship between accuracy and confidence for research on 21 sensory

[8]See Juslin, Winman, and Olsson (2000) for a useful examination of the various factors contributing to the hard-easy effect.

DOI: 10.1002/bdm

discrimination tasks and 44 general knowledge tasks indicated that underconfidence begins to be seen for conceptual tasks at accuracy levels of about 70%.

These same data then suggest that overconfidence on conceptual decisions may not be the norm it has been touted to be. It appears that overconfidence for such decisions may only be present for the most difficult tasks (those with average accuracies at 65% or lower). Indeed, other researchers have also either observed a close correspondence between overall accuracy and confidence, or underconfidence, on conceptual tasks. For instance, Stankov (1998) obtained good agreement between accuracy and confidence on both a vocabulary task and Raven's matrices test. Furthermore, the task requiring recollection of co-workers' eye colors put forth by Dawes (1980) as ''perceptual'' was actually quite similar to our conceptual task. Both required participants to make judgments about physical characteristics using information from their knowledge bases. Interestingly, Dawes also found underconfidence in his task. Finally, Juslin et al. (2000) concluded that there is not a cognitive overconfidence bias when the effects of item selection procedures are controlled for (see More Items were Consistently Wrong for Conceptual than for Perceptual Decisions).

However, it is possible that our findings of low levels of confidence (i.e., occasional significant underconfidence) for the conceptual decisions may also be due in part to participants having judged that they generally lacked knowledge regarding certain categories (e.g., fish lengths). Kruger and Dunning (1999) state that in order for overconfidence to occur, the decision-maker must reach a certain ''minimum threshold of knowledge'' which suggests they have an ability to produce correct answers. Although the stimuli for the current experiment were selected using the most common exemplars of a category, participants still sometimes verbally reported not knowing what the items were.

Confidence in perceptual and conceptual decisions was significantly correlated

Within individuals, confidence ratings for conceptual decisions were consistently significantly positively correlated with confidence ratings for perceptual decisions (across experiment average $r = .62$). The same was true for participants' level of over/underconfidence (across experiment average $r = .53$). These results are consistent with additional work in our lab (Koutstaal, Kvidera, & Matthews, 2007) demonstrating significant across-task correlations of confidence and over/underconfidence using facial recognition and line-length judgments for perceptual decisions and general knowledge and fluid intelligence tasks for conceptual decisions. Such findings also agree well with other research findings showing within-person across-task correlations in confidence assessment using tasks such as Raven's matrices, vocabulary, digit span, and line-length judgment (e.g., Stankov & Crawford, 1996; Stankov, 1998).

Particularly when taken together with the capacity for participants to show a reversal of the ''typical'' pattern of over/underconfidence for perceptual and conceptual decisions discussed above, the presence of such substantial *within-person but across-decision task* correlations may be taken as evidence against the proposal that the grounds and/or processes for confidence assessment in perceptual tasks differ from those for conceptual tasks. If entirely different processes or sources of information contributed to the evaluation of one's confidence in perceptual versus conceptual decisions, then a positive correlation across tasks would not be expected: one would expect no or little correlation. Instead, we found that, on average, approximately 36% of the variance in the absolute level of confidence ratings in one type of decision task could be accounted for by confidence ratings on the other type of decision; similarly, nearly 28% of the variance in over/underconfidence on one decision type could be predicted based on the other decision task.

Considered in conjunction with our other findings pointing to parallel hard-easy effects on the two decision types, these data provide strong evidence that at least some processes or sub-processes of confidence assessments for decisions that are perceptually based versus conceptually based may be shared. However, it is also important to note here that our perceptual decisions were themselves somewhat ''higher level'' (i.e., perceptual judgments rather than purely sensory judgments); it is possible that the across-task within-person correlations would be of smaller magnitude for decisions that were more purely ''sensory'' versus purely

DOI: 10.1002/bdm

''conceptual.'' Nonetheless, overall these outcomes argue for important within-person contributions to the meta-cognitive evaluation of confidence in decision making that are independent of the particular decision domain. As such, these data support the view of a single process model of confidence assessments, thereby also calling into question the need to distinguish between Brunswikian and Thurstonian variability or uncertainty. Other investigators have reached the same conclusion on the basis of computational as well as empirical considerations (e.g., see Vickers and Pietsch (2001) and their extensive critique of Juslin and Olsson's (1997) sampling model of sensory discrimination).

More items were consistently wrong for conceptual than for perceptual decisions

In each experiment, we consistently found more conceptual than perceptual items that fell below chance on accuracy. These data are in agreement with the results of Stankov (1998), who found a number of ''familiar attractors'' among his vocabulary items, but no such misleading items in the line-length judgment task. The presence of some items that were often answered incorrectly suggests that, consistent with the probabilistic mental models account of decision-making for conceptual tasks, heuristic processes (e.g., cue familiarity) may substantially influence the outcomes. This coincides with the conclusions of Juslin et al. (2000) whose meta-analysis of representative versus selected item samples illustrated a lack of overconfidence for random samples but significant overconfidence for selected samples, even when matching the sampling types on accuracy. However, the lack of significant overconfidence for our conceptual tasks suggests that reliance on heuristics does not inevitably generate overconfidence as may seem implied by Gigerenzer et al. (1991).

In addition, heuristically based processes do not appear to operate in the same way or to the same extent for perceptual tasks (including the new types of perceptual decisions introduced here). The 17 perceptual items consistently below chance accuracy across experiments might suggest that heuristics may play a small role in perceptual decision making; however, it is more likely that these items simply make up the lower end of the accuracy distribution. The findings from our item analyses further emphasize that although random selection of items can reduce the influence of misleading items, random-pairing processes do not entirely exclude the presence of systematically ''misleading'' items, but rather more nearly equate the rate of their occurrence within the task to that in the natural environment.

The results of the item analyses do seem to suggest that the processes that individuals use to make conceptual versus perceptual *decisions* differ in the extent of their reliance on heuristics. However, the co-existence of a close match between accuracy and confidence and a strong reliance on heuristics for conceptual decisions, together with the overconfidence and weak reliance on heuristics for perceptual decisions across our three experiments, demonstrates that *confidence assessments* of the two decision types may not be as differentially affected by differences in heuristic versus non-heuristic information processing as has been proposed. Thus, our findings argue that levels of over/underconfidence can be dissociated from the degree of heuristic-guided decision processing.

Confidence ratings significantly correlated with accuracy in all experiments

While much of the research on this topic focuses on errors in calibration, it is also important (and highly encouraging!) to see that confidence is significantly and positively related to accuracy. However, the strength of this relationship is not overly high (average $\gamma = .33$, $\rho = .23$). In addition, Chua, Rand-Giovannetti, Schacter, Albert, and Sperling (2004) dissociated two brain regions involved in accuracy and confidence; whereas activation of the left inferior prefrontal cortex was associated with subjective reports of high confidence, regardless of the accuracy of the decision, activation of medial temporal lobe regions was only associated with high confidence for correctly made decisions. These results illustrate that confidence and accuracy are related, but separable factors. It also re-emphasizes the need to be cautious in using confidence

DOI: 10.1002/bdm

levels to determine courses of action for oneself, or as a means of attempting to evaluate the likely accuracy of the decisions or judgments of another person.

Confidence ratings significantly curvilinearly correlated with response time

This result replicates numerous findings of a relationship between RT and confidence (e.g., Baranski & Petrusic, 1994; Henmon, 1911; Petrusic & Baranski, 2003; Robinson et al., 1997). The typical pattern obtained reflects a negative correlation with faster responding corresponding to higher levels of confidence. Our findings illustrate a modification of this pattern for paired-choice tasks, especially under time pressure: participants tend to have more confidence when they can respond quickly, but also tend to make many of their ''complete guesses'' quickly as well.

Global retrospective estimates of performance (PTPE) versus item-by-item confidence

The results of the PTPE analyses were in line with predictions; in each experiment, lower levels of confidence were indicated on the PTPE measure than on the item-by-item confidence measure. Given that lower than chance PTPE ratings were not allowed, this may be taken as evidence in support of Gigerenzer et al.'s (1991) hypothesis that item-by-item confidence judgments and global post-test performance estimates are made using different reference classes. It is possible that a lack of knowledge about probabilistic mathematics still contributed to participants' lower PTPE confidence ratings (e.g., a participant who felt unsure for most of the items may have rated their PTPE as 50%, failing to adjust the base rate upward for those items for which he/she felt greater confidence). However, the greater correlations between response rates and PTPE than item-by-item confidences in Experiment 1 also support the reference class hypothesis.[9] That is, since this hypothesis posits that PTPE confidence is based on a reference class of participants' perceived ability to perform this type of task, one can infer that their PTPE confidence would correlate with factors that may signal to the participants that they are struggling with the task. The ability to come to a decision within the allotted time represents one such factor. In contrast, participants' item-by-item confidence was not significantly related to their response rates, demonstrating that participants' local confidence assessments were not influenced by this indicator of the ''task ability'' reference class.

Discrimination shows less consistency, but may be lower for perceptual decisions

Our measurement of discrimination demonstrated less consistent patterns. However, when there was not a significant difference in accuracy between decision types (Experiment 1), there was a tendency for participants to demonstrate better discrimination on conceptual decisions. These findings are comparable to those obtained from Stankov and Crawford (1996) as well as Stankov (1998), who found that resolution and slope were better on conceptual tasks such as vocabulary than on a perceptual line-length judgment task. Furthermore, our intermittent observation of significant across-task correlations for slope mirrors results obtained from Stankov and Crawford (1996) showing weaker relatedness between discrimination measures across tasks. These researchers suggested that the weakness of these effects is likely driven by the lower reliabilities of these measures (as tested through split-half reliability and parallel forms). The weak correlations across decision type in addition to the inconsistent patterns of differences in discrimination between conceptual and perceptual decisions seem to support this view. Further studies with a more reliable

[9]These correlations were not run for Experiment 2 (as there was no variation in response rates) or for Experiment 3 (as there was a ceiling effect on response rates).

DOI: 10.1002/bdm

measure of discrimination (if one can be devised) should address differences in discrimination under matched accuracy to determine if the patterns found here persist.

Limitations and conclusion

Certain limitations of the current study should be addressed in future research. For instance, care should be taken in the design of future studies to create perceptual and conceptual items that elicit equal decision RTs and decision accuracies. Also, as noted above, the stimuli used for our conceptual decisions included some items with which some of the participants were not familiar. In addition, as these studies used newly developed stimuli, the conclusions of these experiments are somewhat limited to these specific stimuli (although stimuli did include multiple pairings and both real word and remixed letter strings for perceptual items). Design of stimuli that can be used for both decision types is labor intensive, but we are currently attempting to develop a similar experimental paradigm using face stimuli to further examine the generalizability of these findings. Furthermore, there is a lack of clarity as to what constitutes a ''perceptual'' versus ''conceptual'' decision; previous research treated these two decision types as separate categories. In contrast to this dichotomy, it seems more likely that perceptual and conceptual decisions represent ends of a continuum (see Roediger, Weldon, & Challis, 1989, for a description of the perceptual ''data-driven''–conceptual ''conceptually driven'' continuum). Indeed, in developing the distinction between Brunswikian and Thurstonian uncertainty, Juslin and Olsson (1997) acknowledge this continuum. Further research examining the confidence accuracy relationship along this continuum is needed.

Nonetheless, these experiments demonstrate that differences in the relationship between confidence and accuracy between perceptual and conceptual decisions can be seen when ruling out any differential contributions from the varying stimuli and experimental design for the decision types, and also equating level of accuracy in the decisions. The current research shows that, contrary to predictions based exclusively on Brunswikian versus Thurstonian perspectives, overconfidence in perceptual decisions may sometimes exceed that demonstrated for conceptual decisions. An integration of our findings with past results demonstrates the potential to obtain either overconfidence or underconfidence on both decision types; our results also demonstrated significant positive correlations in over/underconfidence across the two decision types, and similar hard-easy effects regardless of decision type. Thus, while there are obvious differences in the information utilized to make perceptual versus conceptual decisions, our results are consistent with the idea that confidence assessments regarding the accuracy of those decisions rely on common mechanisms (or at least largely shared common mechanisms).

In summary, in this research we developed a new paradigm with which to test the ways in which confidence assessments differ for perceptual versus conceptual decisions, thereby precisely controlling for many (although not all) task and stimulus parameters across the two decision types. Based on the outcomes from three experiments, we conclude that the ''cognitive overconfidence bias'' is not the pervasive phenomenon it is often portrayed as, and that the ''underconfidence phenomenon'' for perceptual decisions is not universal either. Furthermore, the cognitive and meta-cognitive representational processes that are called upon in making confidence assessments about *what we see* versus *what we know*, are to a substantial degree shared or overlapping, rather than distinct from one another.

ACKNOWLEDGEMENTS

The authors would like to thank a number of graduate students for their assistance in the pilot work for this experiment. We would also like to thank three undergraduate research assistants, Brett Fank, Courtney A. Poster, and Aleta Reese, for their help in recruiting and testing participants.

DOI: 10.1002/bdm

REFERENCES

Adams, J. K., & Adams, P. A. (1961). Realism in confidence judgments. *Psychological Review*, *68*, 33–45.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*, 412–428.

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology*, *49*, 397–407.

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.

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics*, *61*, 1369–1383.

Baranski, J. V., & Petrusic, W. M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, *55*, 195–206.

Bar-tal, Y., Sarid, A., & Kishon-Rabin, L. (2001). A test of the overconfidence phenomenon using audio signals. *Journal of General Psychology*, *128*, 76–80.

Bjorkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics*, *54*, 75–81.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*, 212–219.

Chua, E. F., Rand-Giovannetti, E., Schacter, D. L., Albert, M. S., & Sperling, R. A. (2004). Dissociating confidence and accuracy: Functional magnetic resonance imaging shows origins of the subjective memory experience. *Journal of Cognitive Neuroscience*, *16*, 1131–1142.

Crawford, J. D., & Stankov, L. (1996). Age differences in the realism of confidence judgments: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences*, *8*, 83–103.

Cutmore, T. R. H., & Muckert, T. D. (1998). Event-related potentials can reveal differences between two decision-making groups. *Biological Psychology*, *47*, 159–179.

Dawes, R. M. (1980). Confidence in intellectual judgments vs. confidence in perceptual judgments. In E. D. Lantermann, & H. Ferfer (Eds.), *Similarity and choice* (pp. 327–345). Bern: Hans Huber Publishers.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, *4*, 243–260.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition*, *9*, 288–307.

Ferrell, W. R. (1995). A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination. *Perception & Psychophysics*, *57*, 246–254.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552–564.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906–911.

Fu, T., Koutstaal, W., Fu, C. H. Y., Poon, L., & Cleare, A. J. (2005). Depression, confidence, and decision: Evidence against depressive realism. *Journal of Psychopathology and Behavioral Assessment*, *27*, 243–252.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*, 186–201.

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology*, *241*, 1–52.

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.

Juslin, P., Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior and Human Decision Processes*, *73*, 3–26.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384–396.

DOI: 10.1002/bdm

Juslin, P., Winman, A., & Olsson, H. (2003). Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes*, *92*, 34–51.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
Kellogg, W. N. (1931). The time of judgment in psychometric measures. *American Journal of Psychology*, *43*, 65–86.
Keren, G. (1988). On the ability of monitoring non-verdical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologia*, *67*, 95–119.
Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*, 217–273.
Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making*, *10*, 269–278.
Koutstaal, W., Kvidera, S., & Matthews, S. C. (2007). Effects of age and type of decision on the relation between decision accuracy and confidence: Older but not wiser? Manuscript under revision.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.
Liberman, V. (2004). Local and global judgments of confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 729–732.
Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of subjective probabilities: The state of the art up to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.
Luu, P., Collins, P., & Tucker, D. M. (2000). Mood, personality, and self-monitoring: Negative affect and emotionality in relation to frontal lobe mechanisms of error monitoring. *Journal of Experimental Psychology: General*, *129*, 43–60.
McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities, theories and models: 1980–1994. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). New York: John Wiley & Sons, Ltd.
Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*, 102–116.
Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 2–25). Cambridge, MA: MIT Press.
Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Proceedings of the National Academy of Sciences*, *3*, 75–83.
Petrusic, W. M., & Baranski, J. V. (1997). Context effects in the calibration and resolution of confidence. *American Journal of Psychology*, *110*, 543–572.
Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review*, *10*, 177–183.
Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*, 39–57.
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.
Roediger, H. L. III., Weldon, M. S., & Challis, B. A. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger III, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Erlbaum.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.
Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological Review*, *107*, 525–555.
Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 141–151.
Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General*, *135*, 409–428.
Stankov, L. (1998). Calibration curves, scatterplots, and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences*, *10*, 29–50.
Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*, 971–986.
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.
Volkmann, J. (1934). The relation of time of judgment to certainty of judgment. *Psychological Bulletin*, *31*, 672–673.
Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology*, *34*, 135–148.

DOI: 10.1002/bdm

Authors' biographies:

Sara Kvidera is a graduate student in the Psychology Department at the University of Minnesota. Her research interests include human metacognitive abilities such as confidence assessments, and the monitoring and modification of thinking and learning strategies (for instance, those that rely on more automatic vs. more controlled processing).

Wilma Koutstaal is an Associate Professor of Psychology at the University of Minnesota and Visiting Scientist at the University of Reading, Reading, UK. Her research interests focus on memory, thinking, and judgment.

Authors' addresses:

Sara Kvidera, Department of Psychology, 75 East River Road, Minneapolis, MN 55455, USA.

Wilma Koutstaal, Department of Psychology, 75 East River Road, Minneapolis, MN 55455, USA.

DOI: 10.1002/bdm