The disutility of the hard-easy effect in choice confidence

Edgar C. Merkle
Department of Psychology
Wichita State University

A common finding in confidence research is the hard-easy effect, where judges exhibit greater overconfidence for more difficult sets of questions. Many explanations have been advanced for the hard-easy effect, including systematic cognitive mechanisms, experimenter bias, random error, and statistical artifact. In this paper, I mathematically derive necessary and sufficient conditions for observing a hard-easy effect, and I relate these conditions to previous explanations for the effect. I conclude that all types of judges exhibit the hard-easy effect in almost all realistic situations. Thus, the effect's presence cannot be used to distinguish between judges or to draw support for specific models of confidence elicitation.

Keywords: Confidence, hard-easy effect, overconfidence, confidence modeling

Confidence and its calibration are oft-studied topics in the decision sciences. The topics are relevant to many applied areas, including finance (Thomson, Önkal-Atay, Pollock, & Macaulay, 2003), meteorology (Murphy & Winkler, 1984), and eyewitness testimony (Wells, 1981). Psychological research on confidence also has implications for the elicitation of prior distributions in Bayesian models (e.g., O'Hagan et al., 2006). This general applicability of confidence elicitation contributes to its popularity as a research area.

In the above applications, confidence is usually expressed as a probability: given a single event, zero expresses certainty that the event will not occur, 1 expresses certainty that the event will occur, and intermediate probabilities signify intermediate levels of certainty. This is known to decision researchers as a No Choice-100 (NC100) task (terminology from Ronis & Yates, 1987). In an alternative task, the Choice-50 (C50) task, judges choose between two alternatives and then report confidence in their choice. Confidence is bounded at .5 and 1 because, if the judge's confidence is below .5, she should have chosen the other alternative.

Regardless of the task, researchers often examine a judge's calibration by comparing average confidence ($\overline{f}$) over a set of events to proportion correct ($\overline{d}$) over the same set. This results in the overconfidence statistic, OC:

$$\text{OC} = \overline{f} - \overline{d}. \quad (1)$$

Judges are said to be well-calibrated if OC$= 0$; that is, if their average confidence matches proportion correct. It is very common to find that OC> 0; that is, that judges are *overconfident*.

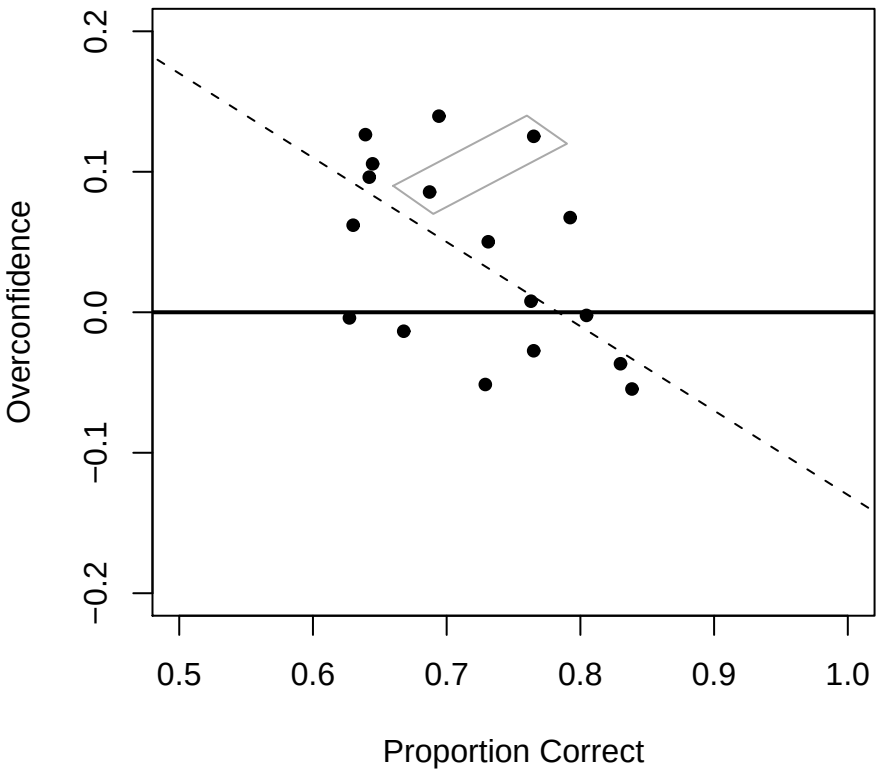
A second ubiquitous finding in confidence research deals with the magnitude of OC at different difficulty levels. This finding, termed the *hard-easy effect*, was described in detail by Lichtenstein, Fischhoff, and Phillips (1982; see also Lichtenstein & Fischhoff, 1977). The effect states that people tend to exhibit more overconfidence for hard sets of questions, vs. for easy sets of questions. Across experiments or question sets, a hard-easy effect for the C50 task is displayed in Figure 1. Proportion correct is on the x-axis, overconfidence is on the y-axis, and each point represents a hypothetical experiment or question set. The points show the general hard-easy trend: as $\overline{d}$ increases, OC decreases.

Many explanations have been advanced for the hard-easy effect, including insufficient placement of confidence criteria in a signal detection framework (Ferrell & McGoey, 1980; Suantak, Bolger, & Ferrell, 1996), random error (Erev, Wallsten, & Budescu, 1994), the insensitivity of confidence to task difficulty (Price, 1998; von Winterfeldt & Edwards, 1986), and cognitive bias (Griffin & Tversky, 1992). While these explanations are internal to the judge, other researchers have proposed that the experimental design itself contributes to the hard-easy effect. For example, Gigerenzer et al. (1991) show that the biased selection of test questions can yield a hard-easy effect: if an experimenter chooses more trick questions than are usually found in some domain, for example, then we might expect a judge's confidence to be artificially high and accuracy to be artificially low.

Juslin, Winman, and Olsson (2000) quantify the hard-easy effect as a regression weight between $\overline{d}$ and OC. In this paper, I expand upon this insight and study the covariance between proportion correct and overconfidence. If this covariance is negative, then we observe a hard-easy effect: as proportion correct increases, overconfidence tends to decrease. In deriving conditions under which the covariance is negative, I am able to derive necessary and sufficient conditions for observing a hard-easy effect in empirical data. These conditions provide a unifying framework for comparing and con-

The author thanks William Batchelder, Henrik Olsson, and Thomas Wallsten for their insightful comments on this article. Any mistakes that remain are the sole responsibility of the author. Correspondence may be sent to Edgar C. Merkle, Department of Psychology, 1845 Fairmount Box 34, Wichita, KS 67260. Email: edgar.merkle@wichita.edu.

Figure 1. Graph of the hard-easy effect. Each point on the graph represents a hypothetical experimental outcome, and the negative trend of the points represents the hard-easy effect. The two points in the grey box demonstrate a reversal of the hard-easy effect.



trasting the explanations for the hard-easy effect described above, as well as for determining the extent to which these explanations are empirically separable.

The analyses in this paper formalize and extend a number of other researchers' verbal arguments and simple demonstrations. Wallsten (1996) argues that confidence researchers have placed too much emphasis on calibration and too little emphasis on cognitive processes. He shows the extent to which different analyses can impact the experimental results (see also Dawes & Mulford, 1996; Erev et al., 1994), and he demonstrates how the hard-easy effect can be a symptom of the test (versus of the judge). In this paper, I am not concerned with specific explanations for the hard-easy effect; I instead argue that the effect cannot help us distinguish between the potential explanations. To be specific, I mathematically derive conditions that are necessary and sufficient for observing a hard-easy effect. In relating these conditions to a general model of judges, I show that the hard-easy effect will occur in almost all experiments. Thus, presence of the effect tells us nothing about the confidence elicitation process.

In the following pages, I first derive necessary and sufficient conditions for observing the hard-easy effect. Next, I use an error model of confidence to examine situations in which the sufficient condition is satisfied. The situations include both C50 and NC100 experiments, described at the beginning of the article. Finally, I discuss the general implications of my analyses for confidence research and modeling.

Necessary & Sufficient Conditions for a Hard-easy Effect

As described in the introduction, the hard-easy effect can be viewed as a description of the covariance between proportion correct and overconfidence ($cov(\overline{d}, OC)$): experimental findings show that there tends to be a negative relationship between these two measures. This relationship is shown in Figure 1, which is a scatter plot of $\overline{d}$ versus OC. Each point represents the $\overline{d}$ and OC calculated for a single test, and the negative trend of the points represents a hard-easy effect. The dotted line is the regression line for these data.

Following Figure 1, my analyses focus on hard-easy effects at the test level (i.e., where each point in the graph represents a single test). While I focus on a single judge completing a series of tests, the number of judges who take

the tests is unimportant: we could plot a single judge's data across a series of tests as well as a group's data across a series of tests. I ignore "within-test" analyses, where accuracy is calculated within different confidence bins and calibration curves are employed (see, e.g., Dougherty, 2001; Ferrell & McGoey, 1980; Lichtenstein et al., 1982; Merkle & Van Zandt, 2006). This is because the latter analyses do not immediately take into account the proportion of responses within each confidence bin, which can lead to misleading results (e.g., Wallsten, 1996).

To examine the ubiquity of the hard-easy effect, we can expand $cov(\overline{d}, \mathrm{OC})$. In the equations below, $\overline{d}$ is proportion correct and $\overline{f}$ is average confidence at the test level. Following much empirical decision-making research, I assume that the f_i are scaled from 0–1 (or, for C50 tasks, from 0.5–1).

$$\begin{aligned} cov(\overline{d}, \mathrm{OC}) &= cov(\overline{d}, \overline{f} - \overline{d}) \\ &= cov(\overline{d}, \overline{f}) - var(\overline{d}) \\ &= sd(\overline{d}) sd(\overline{f}) corr(\overline{d}, \overline{f}) - var(\overline{d}) \\ &= sd(\overline{d})[sd(\overline{f}) corr(\overline{d}, \overline{f}) - sd(\overline{d})]. \quad (2) \end{aligned}$$

Based on Equation (2), it is possible to derive necessary and sufficient conditions for the hard-easy effect.

Proposition 1. *Assume that the hard-easy effect is defined across tests as a negative covariance between $\overline{d}$ and OC. Then:*
(A) we will observe a hard-easy effect if and only if: $sd(\overline{f}) corr(\overline{d}, \overline{f}) < sd(\overline{d})$ (necessary condition).
(B) we will observe a hard-easy effect if: $var(\overline{f}) < var(\overline{d})$ (sufficient condition).

Proof. (A): Examining Equation (2), all standard deviations are greater than or equal to 0. Thus, the sign of Equation (2) depends entirely on the difference in brackets.
(B): By definition, $corr(\overline{d}, \overline{f}) \leq 1$. Thus, $sd(\overline{f}) corr(\overline{d}, \overline{f}) \leq sd(\overline{f})$. The sufficient condition (expressed in Proposition 1 as variances instead of as standard deviations) automatically satisfies the necessary condition. □

An immediate implication from Proposition 1A (the necessary condition) is that judges who report confidence judgments that are completely unrelated to the stimulus will always exhibit a hard-easy effect. In such a case, $corr(\overline{d}, \overline{f}) = 0$ (and $sd(\overline{d}) > 0$ for realistic tests). Thus, through nothing interesting on his or her part, the occasional undergraduate who completes a 45-minute confidence experiment in 3 minutes will exhibit a hard-easy effect. An interesting aspect of Proposition 1B (the sufficient condition) is that it contains a definition of "insensitivity to task difficulty:" confidence is not affected by task difficulty as much as it should be, and, in turn, mean confidence varies less than does proportion correct. I will return to this issue in the General Discussion.

Judges (and mathematical models of confidence) whose confidence and choice satisfy the Proposition 1B condition will exhibit a hard-easy effect. Thus, it is of interest to determine the ease with which this sufficient condition is satisfied. If every judge satisfies the condition on every test, then it is useless to seek a scientific explanation for it. That is, if a hard-easy effect is always present, a model's ability to exhibit the effect is meaningless.

Upon initial examination, it appears that Proposition 1B is not very stringent. In reviewing sets of confidence experiments, some researchers have noted that the variance of accuracy tends to be larger than that of confidence (Dawes & Mulford, 1996; Juslin et al., 2000). Consider further a Choice-50 task, where judges choose one of two alternatives and then give confidence in their chosen alternative.[1] In this task, confidence ranges from .5 to 1 because the judge should choose the alternative that she believes is more likely to be correct (that is, if confidence in the chosen alternative is below .5, then the judge should have chosen the other alternative). As a result, $\overline{d}$ is based on a series of 0's and 1's, while $\overline{f}$ is based on numbers that range from .5 to 1. Within any one experiment, this leads us to surmise that the variance of accuracy would tend to be greater than the variance of confidence. The greater within-experiment variance of accuracy might naturally lead to greater between-experiment variance. If this is the case, then the condition in Proposition 1B is satisfied. We thus observe a hard-easy effect, regardless of the psychological mechanisms underlying confidence elicitation.

While the above arguments are intuitive, they are not formal enough to draw any definitive conclusions. In the next section, I mathematically develop the arguments to show that the condition in Proposition 1B is satisfied in almost all realistic situations.

Satisfying the Sufficient Condition

To study when the condition in Proposition 1B is satisfied, I use an error model of confidence[2] (Juslin, Olsson, & Björkman, 1997) to resemble realistic judges in two-alternative confidence experiments. For each test item, the model assumes that judges have a well-calibrated, internal confidence judgment in the correctness of each alternative. Choice is based on these internal confidence judgments, so that the judge chooses the alternative that is more likely to be correct. Random error then enters into the translation from internal confidence to reported confidence. For a judge responding to item k ($k = 1, \ldots, K$) on test j ($j = 1, \ldots, J$), this is expressed as:

$$f_{jk} = t_{jk} + e_{jk}, \quad (3)$$

where f_{jk} is reported confidence (bounded between 0 and 1), t_{jk} is internal confidence (bounded between 0 and 1), and

[1] This task is commonly employed by decision researchers; see, e.g., Arkes et al. (1987), Dawes (1980), Gigerenzer et al. (1991), and Lichtenstein and Fischhoff (1977).

[2] Juslin et al. (1997) refer to this model as the "response error model." I generically call it an "error model" here, because the phrase "response error" is ill-defined and could lead readers to believe that the error term only captures one type of error. For differing definitions of "response error," see Merkle et al. (in press) and Olsson et al. (in press).

$e_{jk} \sim N(0, \sigma^2)$ (unbounded[3]). t_{jk} also reflects test difficulty, because we assume that these internal confidence judgments are perfectly calibrated.

The above model is very general; Merkle et al. (in press) show that the error term in this model can account for a number of systematic biases. Thus, I can use the model to generally examine the frequency with which the hard-easy effect occurs. I make no arguments that the error model best describes the confidence elicitation process; I simply use the error term to encompass many potential explanations for the hard-easy effect.

Assume that a single judge completes J tests which have K items each. Further assume that item difficulty (described by the t's) arises from the same distribution across tests. If $K = \infty$ (infinite number of items per test), then each test would have the same difficulty. However, given the relatively small values of K in many experiments, we can obtain tests of varying difficulty while sampling all t's from the same distribution. This is a simplifying assumption that could be relaxed if necessary.

Given a specific distribution of internal confidence judgments (t's) for the tests and constant error variance, the error model's predictions for $var(\overline{f})$ and $var(\overline{d})$ are given by (see Appendix A for more details; E is the expectation operator):

$$var(\overline{d}) = \frac{1}{JK}\left[E(t(1-t)) + var(t)\right] \quad (4)$$

$$var(\overline{f}) = \frac{1}{JK}\left[var(t) + \sigma^2\right]. \quad (5)$$

For a specific experimental paradigm, these equations represent the error model's predictions of how $\overline{d}$ and $\overline{f}$ vary across tests. If Equation (4) is larger than Equation (5), then we will observe a hard-easy effect. Substituting these equations into Proposition 1B and rearranging, we will observe a hard-easy effect if:

$$\sigma^2 < E(t(1-t)). \quad (6)$$

An immediate implication from Equation (6) is that, when judges give perfectly-calibrated confidence judgments at the item level, they will almost always exhibit a hard-easy effect at the test level. "Perfectly-calibrated at the item level" means that, for each test item, the judge reports a confidence judgment that equals her probability of choosing the correct answer. In other words, when a judge reports confidence of c for some item, she chooses the correct answer c% of the time. "Test level" means that we give a judge multiple tests, and we calculate average proportion correct and overconfidence for each test. In the language of the error model (Equation (3)), "perfect calibration at the item level" means that $\sigma^2 = 0$. Furthermore, regardless of the specific tests that a judge takes, $E(t(1-t)) \geq 0$[4]. Thus, Equation (6) is almost always satisfied.

The fact that perfect calibration at the item level leads to a hard-easy effect at the test level is unintuitive. Perfect calibration is usually defined at the test level, where, referring to Figure 1, perfectly-calibrated judges would follow the horizontal line along OC = 0. If judges are perfectly calibrated at the item level, however, they can exhibit a hard-easy effect due to the Bernoulli error (ε) inherent in $\overline{d}$. In this situation, we can write $\overline{d} = \overline{f} + \varepsilon$ and $\text{OC} = \overline{f} - \overline{d} = -\varepsilon$. Positive (negative) values of ε make $\overline{d}$ large (small) and OC small (large). This results in a hard-easy effect.[5]

The findings thus far go a long way towards rendering the hard-easy effect an uninteresting phenomenon: the effect is present for both perfect judges and awful judges. I proceed, however, to examine the ubiquity of the hard-easy effect when judges are neither perfect nor awful. This is more similar to real judges in real confidence experiments.

Example 1: Choice-50 Task

I first examine the relative magnitudes of $var(\overline{d})$ and $var(\overline{f})$ in the context of Choice-50 tasks, where judges first choose one of two alternatives and then report confidence ranging from .5-1. For these tasks, an initial distribution from which we may elect to sample the t_{jk} is the Uniform(0.5,1). With this distribution, it can be shown that $E(t(1-t)) = .167$. Therefore, to observe a hard-easy effect, the error variance must be less than .167.[6] In the context of real experiments, the .167 error variance is massive: across many contexts, Juslin et al. (2000; 2003) estimate σ^2 to be between .015 and .03. Merkle et al. (in press) and Olsson et al. (in press) both discuss the fact that this estimate is surprisingly large. These estimates are nowhere close to .167, indicating that, if judges respond in a C50 task according to the error model, they will always exhibit the hard-easy effect.

A major assumption underlying the above example was the distribution of the t_{jk}. While I assumed a Uniform(0.5,1) distribution, a more realistic distribution might involve many of the t_{jk} being clustered around .5 or 1 (see, e.g., Erev et al., 1994; Wallsten, 1996). For simplicity, I specify a discrete distribution on the t_{jk} based on that from Erev et al.:

$$P(T_{jk} = t_{jk}) = \begin{cases} .3 \text{ for } t_{jk} = .50 \\ .1 \text{ for } t_{jk} = \{.60, .70, .80, .90\} \\ .3 \text{ for } t_{jk} = 1 \\ 0 \text{ otherwise} \end{cases}$$

Under this distribution, $E(t(1-t)) = .145$. While this expectation is smaller than that of the Uniform distribution, it

[3] The unbounded error allows for the possibility that f_{jk} exceeds its bounds. As stated in the discussion and shown in Appendix B, this has no effect on the derivations in this paper.

[4] Showing this involves expressing $E(t(1-t))$ as $E(t) - E(t^2)$. For $0 \leq t \leq 1$, $E(t) \geq E(t^2)$ with equality holding only when $P(t=0) + P(t=1) = 1$. In words, $E(t(1-t)) = 0$ only when the judge is certain of every item on a test.

[5] Juslin et al. (2000) describe this type of miscalibration as "linear dependency," though they do not explicitly relate it to perfectly-calibrated judges. Furthermore, Klayman, Soll, and González-Vallejo (1999) present a method for removing linear dependency from observed confidence data.

[6] The .167 figure is an absolute lowest bound for observing a hard-easy effect. If we take into account $corr(\overline{f}, \overline{d})$ (see Proposition 1A), then we can still observe a hard-easy effect when σ^2 is greater than .167.

is still much larger than the σ^2 estimates from previous research. This means that we would still observe a hard-easy effect when the t_{jk} arise from the U-shaped distribution. In the next example, I show that these results extend to tasks where judges use a 0-1 confidence scale.

Example 2: No Choice-100 Task

While the above derivations demonstrate the ubiquity of the hard-easy effect in C50 tasks, the situation does not obviously extend to No Choice-100 (NC100) tasks. In these tasks, judges do not choose between the two alternatives. Instead, they simply give a 0-1 confidence judgment as to the truth of one specific alternative. Wallsten (1996) discusses the fact that NC100 tasks are preferable to C50 tasks because the base rate of true items is entirely in the experimenter's control. In contrast, the base rate of correct items in C50 tasks depends both on the experimenter's selection of test questions and the judge's knowledge of those questions. This can enhance any hard-easy effect that may already have been present.

Sampling the t_{jk} from a Uniform(0,1) distribution, we can calculate $E(t(1-t)) = .167$. This expectation is the same as that of the Uniform(0.5,1) distribution from the C50 task. Thus, the same arguments apply: to make the hard-easy effect disappear, we must have $\sigma^2 \geq .167$. Given that this variance is unrealistic in empirical data (Juslin et al., 2000; Merkle et al., in press), judges will always yield a hard-easy effect.

Similar to the U-shaped distribution in the C50 task, a W-shaped distribution is more realistic for the t_{jk} in an NC100 task. Such a distribution indicates a preponderance of items that participants know are true, know are false, or know nothing about. While Erev et al. (1994) employ a discrete W-shaped distribution on the t_{jk}, I use a mixture of independent Beta distributions:

$$f(t) \sim \frac{1}{2}\left(\text{Beta}(20,20) + \text{Beta}(.25,.25)\right). \tag{7}$$

The Beta distribution is characterized by two parameters (often labeled α and β), and it is bounded at 0 and 1. When $\alpha = \beta$, as is the case for both Beta distributions above, the distribution is symmetric around .5. The above mixture distribution roughly corresponds to the discrete distribution used by Erev et al., and it also lifts the restriction that judges' internal probabilities must be multiples of .10. The distribution is displayed in Figure 2.

The Beta mixture distribution of t_{jk}'s yields $E(t(1-t)) = .207$. This is again larger than realistic σ^2 values, and it is also larger than the figure derived for the Uniform distribution above. This continues the general trend of observing hard-easy effects in all realistic situations.

In an effort to find t_{jk} distributions that do not result in a hard-easy effect, I explored asymmetric Beta distributions (those for which $\alpha \neq \beta$). For example, consider the Beta(.25,1) distribution: this reflects an NC100 test where participants know that most items are false (see Figure 3). To be specific, 56% of the true probabilities (t's) lie below .1. Although this distribution is considerably different from the others I considered, it still yields $E(t(1-t)) = .089$. This is still far from the .015-.03 σ^2 range that Juslin et al. estimate from previous data.

Making the Hard-Easy Effect Disappear

Are there specific types of tests that make the hard-easy effect disappear for realistic judges? In other words, what values of α and β can we take so that $E(t(1-t)) \leq .03$? To answer this question, I derive an expression for $E(t(1-t))$ as a function of the Beta distribution parameters α and β:

$$E(t(1-t)) = \frac{\alpha^2\beta + \alpha\beta^2}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{8}$$

I then insert this expression into Equation (6), resulting in the following condition for observing a hard-easy effect:

$$\sigma^2 < \frac{\alpha^2\beta + \alpha\beta^2}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{9}$$

Three-dimensional contour plots of Equation (8) appear in Figure 4 as a function of α and β (in the plots, E stands for $E(t(1-t))$). Both plots display the same contour, but from different angles. For different values of α and β, the plots show values of σ^2 that judges must achieve to make the hard-easy effect disappear. In other words, we can horizontally slice the plots at $E = .03$. Whenever our slicing touches the contour, we have found a test (values of α and β) for which realistic judges do not exhibit a hard-easy effect.

These plots show that, to obtain $E(t(1-t)) \leq .03$, we need either α or β to be very close to 0. When α is close to zero, the majority of the t_{jk} density is close to zero. When β is close to zero, the majority of the t_{jk} density is close to one. These reflect situations where judges are certain that every item is false ($\alpha \approx 0$), certain that every item is true ($\beta \approx 0$), or certain that some items are true and other items are false ($\alpha \approx \beta \approx 0$). To summarize, the hard-easy effect will disappear only when judges are nearly certain about every question on a test. These types of tests are useless in confidence elicitation experiments: if judges are certain that every item is either true or false, then confidence judgments add no extra diagnostic information.

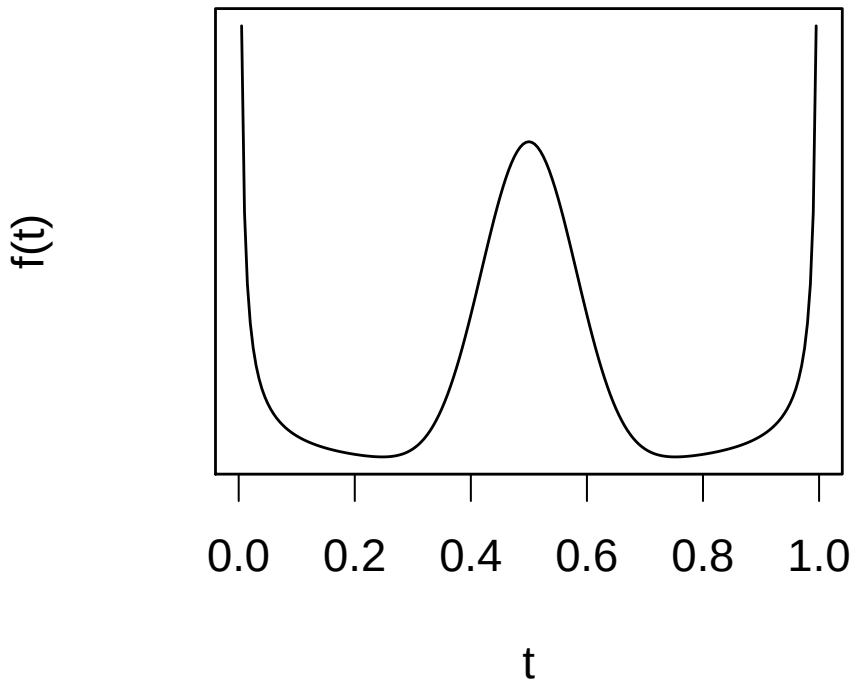
While the Beta distribution can generally account for tests of varying difficulties, it cannot immediately account for multimodal distributions. For example, the W-shaped distribution that I employed in Example 2 cannot arise from a single Beta distribution. To obtain the W-shaped distribution, I used a mixture of two symmetric Beta distributions. Under the constraints that $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$, this mixture has the form:

$$f(t) \sim \frac{1}{2}\left(\text{Beta}(\alpha_1,\beta_1) + \text{Beta}(\alpha_2,\beta_2)\right). \tag{10}$$

For such mixtures of symmetric Beta distributions, it is possible to show that:

$$E(t(1-t)) = \frac{1}{4}\left[1 - \frac{1}{4}\left(\frac{1}{(2\alpha_1+1)} + \frac{1}{(2\alpha_2+1)}\right)\right]. \tag{11}$$

Figure 2. Mixture of Beta(20,20) and Beta(.25,.25) densities.



Thus, we observe a hard-easy effect if:

$$\sigma^2 < \frac{1}{4}\left[1 - \frac{1}{4}\left(\frac{1}{(2\alpha_1+1)} + \frac{1}{(2\alpha_2+1)}\right)\right]. \qquad (12)$$

Three-dimensional contour plots of Equation (11) appear in Figure 5 as a function of α and β (in the plots, E stands for $E(t(1-t))$). Both plots display the same contour, but from different angles. Slicing the plots at $E = .03$ to resemble realistic judges, we see that there exist almost no tests for which the hard-easy effect disappears. The only tests that do make the hard-easy effect disappear are those for which $\alpha \approx \beta \approx 0$. As discussed earlier, these are unrealistic tests for which judges are certain of all items. Thus, judges exhibit a hard-easy effect for all realistic tests that arise from a mixture of symmetric Beta distributions.

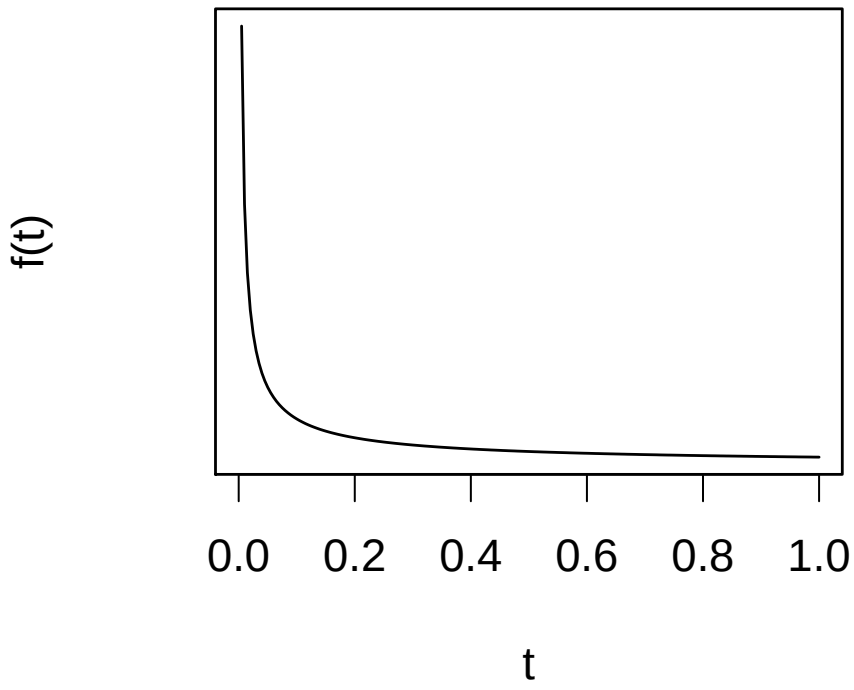
Discussion

The above analyses show that, when realistic judges respond in a confidence experiment according to the response-error model, the sufficient condition in Proposition 1 is satisfied in all realistic situations. This means that judges always exhibit a hard-easy effect in these situations. Wallsten (1996) addresses this topic in the context of a single t_{jk}: he demonstrates that, given a constant response strategy in a Choice-50 task, judges can be overconfident or underconfident depending on the magnitude of t_{jk} (leading to a hard-easy effect). The current derivations extend Wallsten's work by considering sets of t_{jk} and responses that differ for each t_{jk}. They show the extent to which the hard-easy effect is ubiquitous in both Choice-50 and No Choice-100 tasks. In addition, the mathematical derivations show that perfectly-calibrated judges (those with $\sigma^2 = 0$) and awful judges (those whose confidence judgments are unrelated to the stimulus) always exhibit the hard-easy effect.

General Discussion

In this paper, I derived necessary and sufficient conditions for observing a hard-easy effect. Using a simple error model of confidence, I then examined situations under which we could expect to observe a hard-easy effect. Across both Choice-50 tasks and No Choice-100 tasks, I showed that judges almost always exhibit hard-easy effects, even when their confidence judgments are perfectly calibrated at the item level. This result occurs in both Choice-50 and No

Figure 3. The Beta(.25,1) density.



Choice-100 tasks, and it is robust to changes in the distributions of internal probabilities (t_{jk}'s). Hard-easy effects were exhibited across Uniform distributions of internal probabilities, W-shaped mixtures of Beta distributions, and any single Beta distribution reflecting a realistic test. In the following paragraphs, I discuss the extent to which my assumptions may have influenced my results, as well as implications of my findings.

Assumptions

I employed three main assumptions to derive the results in this paper. They are:

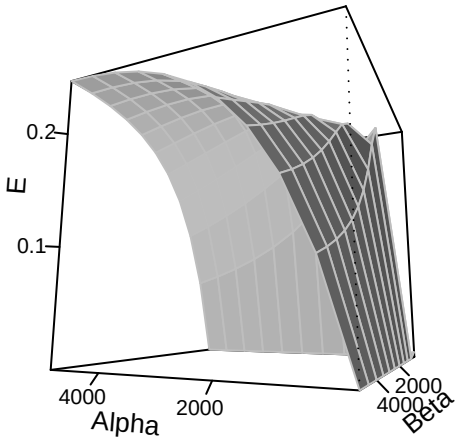
1. Judges respond according to the response-error model.
2. The t_{jk} follow a Beta distribution.
3. Overt confidence judgments (f's) are not bounded between 0 and 1.

While I have already addressed some of these assumptions, I directly discuss the impact of each of these assumptions below.

Use of the Error Model. A major assumption of my analyses is that the error model with $\sigma^2 \approx .03$ adequately describes real judges. If this assumption is grossly incorrect, then there might still exist some realistic situations where the hard-easy effect is absent. There are two main ways in which this assumption can be violated: (1) σ^2 is actually much larger than .03, or (2) the error model is simply a poor description of the confidence elicitation process.

As briefly addressed earlier in the paper, previous confidence data and analyses provide evidence that the error model is reasonable. Merkle et al. (in press) show that the model can mimic data from other confidence models such as the Decision Variable Partition Model (Ferrell & McGoey, 1980) and models that incorporate alternative-underweighting biases (McKenzie, 1997). Furthermore, Merkle et al. show that the σ^2 estimates from Juslin et al. (2000, which is similar to estimates from Björkman, 1994; Juslin et al., 1997) likely incorporate systematic biases as well as error. Thus, the error model with $\sigma^2 = .03$ yields data that are similar to empirical data (Juslin et al., 2000 report an R^2 of .99), even though judges obviously do not respond according to the model. Given that the hard-easy effect is one of observed data and is not tied to a specific confidence elicitation process, this assumption seems reasonable.

Figure 4. Contour plots of Equation (8) as a function of α and β. For the term on the right side of Equation (8) to be below .03, it is necessary that $\alpha \approx 0$ or $\beta \approx 0$.



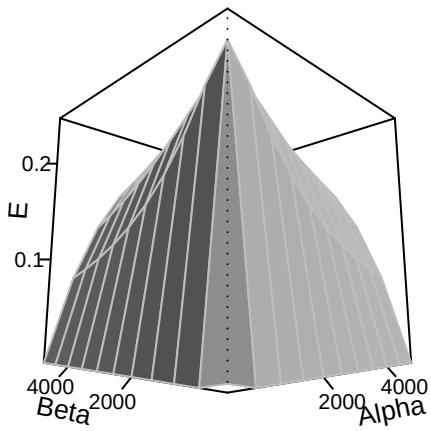
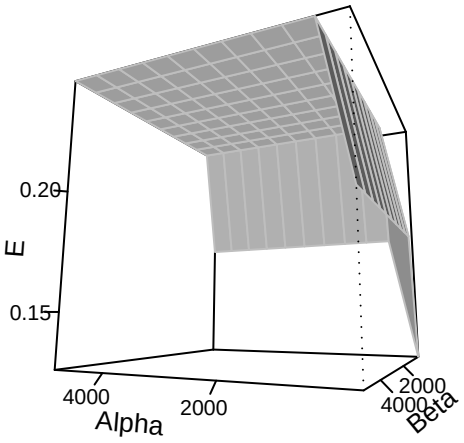
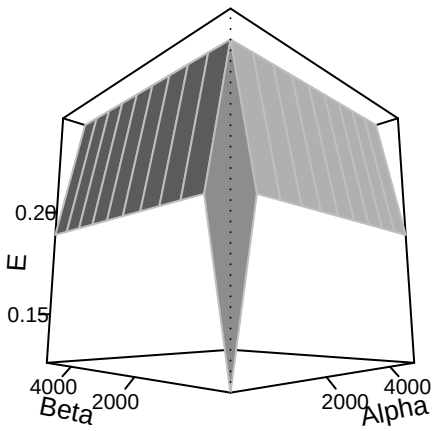


Figure 5. Contour plots of Equation (11) as a function of α and β. For the term on the right side of Equation (11) to be below .03, it is necessary that $\alpha \approx \beta \approx 0$.





Use of the Beta Distribution. The second assumption is that judges' true probabilities (t_{jk}'s) arise from a Beta distribution (although I did examine a small number of other distributions). The Beta distribution is flexible, encompassing the Uniform(0,1) distribution as well as many asymmetric distributions. Thus, many plausible t_{jk} distributions can arise from the Beta distribution. The Beta distribution cannot, however, generally account for multimodal distributions. For example, the Beta cannot immediately yield a W-shaped distribution, where judges know many test questions and are completely uncertain of others. I employed a mixture of symmetric Beta distributions to examine this situation, but I could not undertake an exhaustive analysis of all possible Beta mixture distributions. Given the other analyses, however, I would be surprised to find a Beta mixture distribution that both resembles a realistic test and fails to yield a hard-easy effect.

Unbounded Confidence Judgments. Finally, in comparing $\overline{f}$ with $\overline{d}$, I do not account for the fact that reported confidence judgments are bounded between 0 and 1. The addition of random error to each t_{jk} allows for the possibility that some confidence judgments become smaller than 0 or larger than 1. The sufficient condition that I employed for observing a hard-easy effect, however, is $var(\overline{f}) < var(\overline{d})$. If I did account for the confidence bounds in my analyses, then $var(\overline{f})$ would decrease and $var(\overline{d})$ would remain the same. Thus, the observation of a hard-easy effect in my analyses implies the observation of a hard-easy effect in bounded confidence data. While it was mathematically simpler to disregard the bounds of the confidence scale, it also resulted in conservative analyses. Further evidence supporting this claim appears in Appendix B, where I describe simulations that explicitly account for the bounds of the confidence scale.

Reversals of the Hard-easy Effect

Gigerenzer et al. (1991) discuss two studies (Keren, 1988; Ronis & Yates, 1987) that yield a total of four hard-easy effect reversals. There is nothing systematic about these two experiments: each experiment was intended to study something other than the hard-easy effect, and the original authors provide few analyses that are specifically relevant to the hard-easy effect. However, within some conditions of both studies, participants exhibited more overconfidence for an easy test than for a hard test (see Gigerenzer et al.'s Figure 10). These hard-easy reversals initially appear to invalidate the current results: I have argued that the hard-easy effect always occurs (i.e., that the covariance between proportion correct and overconfidence is always negative), while Gigerenzer et al. discuss studies where it does not occur (i.e., where the covariance appears positive).

Upon closer examination, the hard-easy effect definition employed in this paper ($cov(\overline{d}, \mathrm{OC}) < 0$) can accommodate reversals of the sort described by Gigerenzer et al. An example of this is shown within the grey box in Figure 1. The two points within the box depict a hard-easy reversal: there is greater overconfidence for the easier test. Examining the overall trend of all the points in the graph, however, there is a clear hard-easy effect. In general, two tests are insufficient to reliably assess the covariance that defines the hard-easy effect.

Verbal vs. Mathematical Definitions

The sufficient conditions derived in this paper can also be used to clarify explanations for overconfidence and the hard-easy effect. As described earlier in the paper, the sufficient condition for the hard-easy effect in Proposition 1 contains one definition of "insensitivity to task difficulty." Thus, we might accept insensitivity to task difficulty as a valid explanation for the effect. This agrees with many other researchers (e.g., Lindsay, Nilsen, & Read, 2000; Merkle & Van Zandt, 2006; Price, 1998; Sieck, Merkle, & Van Zandt, 2007; Weber & Brewer, 2004), who have discussed empirical findings related to insensitivity to task difficulty. Interestingly, however, Proposition 1 shows that insensitivity to task difficulty is not necessary for the hard-easy effect to occur.

Two further caveats can be made on the "insensitivity to task difficulty" explanation. First, the definition in Proposition 1B is data-driven: we know whether a judge is insensitive to task difficulty by directly examining whether a set of data satisfy the condition in Proposition 1B. This makes insensitivity to task difficulty more of an effect than an explanation. Other explanations for overconfidence and the hard-easy effect are more process-driven. For example, McKenzie (1997) argues that, in assessing confidence, judges "underweight the alternative:" they focus on their chosen alternative and ignore the unchosen alternative(s). There is nothing in the observed data that tells us definitively whether alternative underweighting occurred; we must rely on experimental manipulations designed to impact alternative underweighting. This type of explanation for overconfidence and the hard-easy effect is potentially more useful than that in Proposition 1B.

A second caveat is that we can define "insensitivity to task difficulty" in ways other than the definition in Proposition 1B. For example, Gigerenzer et al. (1991) define the phrase as a judge's inability to accurately gauge her performance on a test. This definition, which does not necessarily involve confidence, leads Gigerenzer et al. to conclude that their experiments "do not support the explanation of overconfidence and the hard-easy effect by assuming that subjects are insensitive to task difficulty" (p. 520). The authors specifically found that judges were relatively accurate in guessing their proportions correct across tests that varied in difficulty.

To summarize these experimental results, judges are relatively good at directly guessing their accuracy for a set of questions. Judges are relatively bad at indirectly guessing their accuracy via confidence judgments for each item. These results are consistent with the notion of separate psychological mechanisms for confidence and choice.[7] Such a

[7] This statement may appear to conflict with my analyses in this paper, where I assume a single model of confidence and choice. The distinction between a model's ability to fit data versus a model's ability to describe the confidence elicitation process becomes important here: I assume that the error model can sufficiently fit ob-

notion was discussed by Griffin and Tversky (1992), and it has recently been studied in the context of signal detection (Mueller & Weidemann, 2008) and the ASC model of confidence (Sieck et al., 2007). Furthermore, the above paragraphs demonstrate that we can define "insensitivity to task difficulty" in multiple ways, and that these different definitions can lead us to different conclusions. This demonstration generally highlights the importance of mathematical definitions for psychological phenomena: while verbal definitions are intuitive, they lack the precision to specifically define a phenomenon (e.g., Myung & Pitt, 2001).

Summary

Based on the results in this paper, I conclude that the hard-easy effect tells us nothing about the "goodness" of a judge or about the confidence elicitation process. From an empirical perspective, both perfectly-calibrated judges (at the item level) and terrible judges (those whose confidence is unrelated to the stimulus) exhibit the hard-easy effect. From a modeling perspective, the hard-easy effect occurs for all realistic judges and tests. Any model that provides a reasonable fit to observed data will exhibit the effect. Thus, the effect cannot help us learn about the confidence elicitation process because it cannot help us discriminate between potential confidence elicitation models. These results formalize and extend Wallsten's (1996) statement that the hard-easy effect and overconfidence are "not suitable for investigating basic cognitive processes" (p. 225). Instead, other confidence measures such as response distributions or base rates are necessary for the study of confidence elicitation mechanisms.

References

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*, 133-144.

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, *58*, 386-405.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Thomson Learning.

Dawes, R. M. (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (p. 327-345). Bern: Han Huber.

Dawes, R. M., & Mulford, M. (1996). The false consensus effect in overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes*, *65*, 201-211.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, *130*, 579-599.

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.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Processes*, *26*, 32-53.

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 the determinants of confidence. *Cognitive Psychology*, *24*, 411-435.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189-209.

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.

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.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, *67*, 95-119.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216-247.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance*, *20*, 159-183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 306-334). Cambridge, England: Cambridge University Press.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior*, *24*, 685-697.

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes*, *71*, 141-160.

Merkle, E. C., Sieck, W. R., & Van Zandt, T. (in press). Response error and processing biases in confidence judgment. *Journal of Behavioral Decision Making*.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*, 391-408.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of Signal Detection Theory. *Psychonomic Bulletin and Review*, *15*, 465-494.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*, 489-500.

Myung, I. J., & Pitt, M. A. (2001). Mathematical modeling. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology* (3rd ed., Vol. 4, p. 429-460). New York: John Wiley & Sons.

---

served data even if the model's explanation of the confidence elicitation process is incorrect. Thus, the true confidence elicitation process can involve separate mechanisms for confidence and choice, while still yielding calibration data similar to that of the model used in this paper.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Hoboken: Wiley.

Olsson, H., Juslin, P., & Winman, A. (in press). The role of random error in confidence judgment: Reply to Merkle, Sieck, and Van Zandt. *Journal of Behavioral Decision Making*.

Price, P. C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes*, *76*, 277-297.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*, 193-218.

Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: A cognitive contributor to overconfidence. *Organizational Behavior and Human Decision Processes*, *103*, 68-83.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201-221.

Thomson, M. E., Önkal-Atay, D., Pollock, A. C., & Macaulay, A. (2003). The influence of trend strength on directional probabilistic currency predictions. *International Journal of Forecasting*, *19*, 241-256.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York, NY: Cambridge.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes*, *65*, 220-226.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*, 156-172.

Wells, G. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*, 688-696.

Appendix A
Derivation of $var(\overline{d})$ and $var(\overline{f})$ for the Error Model

In this appendix, I derive $var(\overline{d})$ and $var(\overline{f})$ (Equations (4) and (5)) under the error model used in this paper. I first present a general version of the model, and I then discuss the specific assumptions employed for the derivations.

Let the d_{jk} be exchangeable 0/1 variables denoting whether or not a judge correctly answers item k ($k = 1, \ldots, K$) on test j ($j = 1, \ldots, J$). Each d_{jk} depends on a latent variable t_{jk} in (0,1) with distribution $g_j(t)$, such that $f(d_{jk} \mid t_{jk}) \sim$ Bernoulli(t_{jk}).

The f_{jk} depend on t_{jk} such that:

$$f_{jk} = t_{jk} + e_{jk}, \tag{13}$$

where the e_{jk} are independent and follow a distribution $h_j(e)$ with mean 0. Independent sampling of the t_{jk} from $g_j(t)$ implies that the d_{jk} are independent of one another (and also that the f_{ij} are independent of one another).

The error model described in this paper assumes specific distributions for the above model. First, it assumes that $g_j(t) \sim \text{Beta}(\alpha, \beta)\ \forall\ j$. In words, the model assumes that item difficulty follows the same distribution from test to test. This assumption is similar to experiments that randomly sample test items from a specific domain (e.g., Gigerenzer et al., 1991). Next, the model assumes that $h_j(e) \sim \text{N}(0, \sigma^2)\ \forall\ j$. This implies that the judge's response error distribution is constant across items and tests. While these assumptions simplify the derivations, they are not necessary to demonstrate the ubiquity of the hard-easy effect.

Employing the above assumptions, we can obtain the unconditional variance of $\overline{d}$ via the conditional variance identity (e.g., Casella & Berger, 2002):

$$\begin{aligned} var(\overline{d}) &= var(\frac{1}{J}\sum_j \frac{1}{K}\sum_k d_{jk}) \\ &= \frac{1}{(JK)^2}\sum_j\sum_k \left[E(var(d_{jk} \mid t_{jk})) + var(E(d_{jk} \mid t_{jk}))\right] \\ &= \frac{1}{(JK)^2}\sum_j\sum_k \left[E(t_{jk}(1-t_{jk})) + var(t_{jk})\right] \\ &= \frac{1}{JK}\left[E(t(1-t)) + var(t)\right]. \end{aligned}$$

The variance of $\overline{f}$ is:

$$\begin{aligned} var(\overline{f}) &= var(\frac{1}{J}\sum_j \frac{1}{K}\sum_k f_{jk}) \\ &= \frac{1}{(JK)^2}\sum_j\sum_k [var(t_{jk}) + var(e_{jk})] \\ &= \frac{1}{JK}\left[var(t) + \sigma^2\right]. \end{aligned}$$

When $\sigma^2 < E(t(1-t))$, $var(\overline{f}) < var(\overline{d})$. This implies a hard-easy effect (see Proposition 1).

Appendix B
Bounded Error

In this section, I show that some of the results in the paper hold when accounting for the (0,1) bounds of the confidence scale. Following Erev et al. (1994), error is added to the unbounded log-odds of t_{jk} to yield an intermediate variable x_{jk}:

$$x_{jk} = \log\left(\frac{t_{jk}}{1-t_{jk}}\right) + e_{jk}, \tag{14}$$

where $e_{jk} \sim N(0, \sigma^2)$. The intermediate variable, x_{jk}, is then transformed back to the (0,1) scale to yield f_{jk}:

$$f_{jk} = \frac{\exp(x_{jk})}{1+\exp(x_{jk})}. \tag{15}$$

Closed-form expressions for $var(\overline{f})$ are unavailable, so I conducted a simulation study to examine the situations under which a hard-easy effect is observed.

For the simulations, I allowed the two test difficulty parameters (Beta distribution parameters α and β) to vary from 0.001 to 30.001 in increments of 0.5. This specific range was chosen based on the analyses in the main text, which show that the hard-easy effect disappears only at small values of α and β. Furthermore, I allowed the error variance parameter (σ^2) to vary from 0.001 to 1.9 in increments of 0.02. The (0.001,1.9) σ^2 range on the log-odds scale approximates the (0,0.03) σ^2 range on the untransformed confidence scale.

For each combination of parameter values, I generated 1000 tests of 50 items each. In calculating $\overline{d}$ and $\overline{f}$ for each test, presence of a hard-easy effect was assessed by examining whether $cov(\overline{d}, \mathrm{OC}) < 0$. Of the 353,495 datasets generated, hard-easy effects were observed in all but 146. For all 146 datasets in which the hard-easy effect was absent, either α or β equaled 0.001. These findings support the claim that the hard-easy effect appears in all realistic situations.