Cognitive Psychology 78 (2015) 99–147

Contents lists available at ScienceDirect

Cognitive Psychology

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

Post choice information integration as a causal determinant of confidence: Novel data and a computational account

Rani Moran [a,b,*], Andrei R. Teodorescu [c], Marius Usher [a,b]

[a] School of Psychological Sciences, Tel Aviv University, Ramat Aviv, 69978, Israel
[b] Sagol School of Neuroscience, Tel Aviv University, Ramat Aviv, 69978, Israel
[c] Department of Psychological and Brain Science, Indiana University, 1101 E. 10th St., Bloomington, IN, USA

ARTICLE INFO

Article history:
Available online 8 April 2015

Keywords:
Confidence
Resolution of confidence
Decision making
Perception
Collapsing boundaries

ABSTRACT

Confidence judgments are pivotal in the performance of daily tasks and in many domains of scientific research including the behavioral sciences, psychology and neuroscience. Positive resolution i.e., the positive correlation between choice-correctness and choice-confidence is a critical property of confidence judgments, which justifies their ubiquity. In the current paper, we study the mechanism underlying confidence judgments and their resolution by investigating the source of the inputs for the confidence-calculation. We focus on the intriguing debate between two families of confidence theories. According to single stage theories, confidence is based on the same information that underlies the decision (or on some other aspect of the decision process), whereas according to dual stage theories, confidence is affected by novel information that is collected after the decision was made. In three experiments, we support the case for dual stage theories by showing that post-choice perceptual availability manipulations exert a causal effect on confidence-resolution in the decision followed by confidence paradigm. These finding establish the role of RT2, the duration of the post-choice information-integration stage, as a prime dependent variable that theories of confidence should account for. We then present a novel list of robust empirical patterns ('hurdles') involving RT2 to guide further theorizing about confidence judgments. Finally, we present a unified computational dual stage model for choice, confidence and their latencies namely, the

* Corresponding author at: Department of Psychology, Tel Aviv University, Ramat Aviv, POB 39040, Tel Aviv 69978, Israel.
E-mail addresses: rani.moran@gmail.com (R. Moran), ateodore@indiana.edu (A.R. Teodorescu), marius@post.tau.ac.il (M. Usher).

http://dx.doi.org/10.1016/j.cogpsych.2015.01.002

collapsing confidence boundary model (CCB). According to CCB, a diffusion-process choice is followed by a second evidence-integration stage towards a stochastic collapsing confidence boundary. Despite its simplicity, CCB clears the entire list of hurdles.

1. Introduction

Decision confidence is a metacognitive judgment, which enjoys a unique, dual status in the cognitive sciences. First, confidence judgments and their properties attract wide interest in their own right. For example, confidence judgments have been used to measure the calibration of subjective probabilities (i.e., the degree of correspondence between inner beliefs and objective probabilities of an event's occurrence; Lichtenstein, Fischhoff, & Phillips, 1982) and their resolution (i.e., the degree to which inner beliefs discriminate between an event's occurrence and nonoccurrence; Baranski & Petrusic, 1994). Second, confidence judgments constitute an important means for studying additional cognitive processes in a variety of fields including decision making (Koriat, 2012), psychophysical judgments (Peirce, 1877; Peirce & Jastrow, 1884), memory (e.g. Squire, Wixted, & Clark, 2007) adaptive control (Vickers, 1979), conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001) and social interactions (Shea et al., 2014). Appropriately, the neural mechanism underlying confidence is a subject of recent investigations in neuroscience (Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009).

This extensive study of confidence is motivated by its fundamental role in daily life situations. For example, the operation of control mechanisms in goal-driven behavior relies on decision-confidence, which serves as an internal, subjective form of feedback signal that can help assess progress towards one's goals. Consider a student taking a multiple-questions exam: The confidence level the student feels in the correctness of her answer will determine whether she should spend more thought on that particular item or proceed to the next one. Self-reports of confidence also play a crucial role in social interactions, where they affect the reliability attributed to information reported by others (Shea et al., 2014). The ubiquity of confidence judgments, in both daily life and the cognitive sciences, attests to the significance of understanding the psychological mechanisms that underlie them.

The critical property of decision-confidence that enables and justifies this broad usage of confidence judgments is their *positive resolution* i.e., the positive correlation between confidence and decision correctness: the higher the confidence, the more likely the decider is to be correct in his or her decision. In the current paper we examine a fundamental attribute of the mechanism underlying confidence judgments and their positive resolution. Specifically, we focus on the source of information used to calculate confidence and ask when is this information collected. According to one influential family of theories, the *single information-collection stage theories* (henceforth, single-stage theories), confidence is based on *aspects of the decision process* such as its duration, a feeling of effort or simply a different calculation involving the *same* evidence that was used to determine the decision. Importantly, such theories maintain that even if confidence is calculated 'offline' i.e., after the decision has been made, its calculation is confined by the aspects of the process that led to the decision. An opposing family of theories, the *dual information-collection stage theories* (henceforth dual-stage theories) stresses the importance of novel evidence, which is collected only after the decision has been made, as an input to the confidence-calculation process. Such theories thus deny that confidence is determined based solely on aspects of the decision process.

Our experiments were carried out to distinguish between these families of confidence theories by testing the effects of post-choice perceptual availability manipulations on the resolution of confidence judgments. To anticipate, our results provide support for dual-stage theories. We find that during the time lag between the decision and confidence responses, participants continue to collect information about the choice alternatives from the environment and this information affects the calculation of confidence judgments.

Because a post-choice (information-integration) stage exerts a causal influence on confidence, theories of confidence need to consider the duration of this stage and how it operates. Thus, the second main goal of the current paper was to identify robust empirical relations between the duration of the post-choice stage (RT2) and additional variables in our paradigm. Such empirical pattern will guide future theorizing and hence shed light on how confidence judgments are formed.

These robust empirical patterns led to the third main goal of the paper: to present a novel dual stage model that meets the challenge of accounting for the diversity of these patterns. The core assumption of our model is that confidence is generated by a collapsing boundary mechanism, which is set once the decision is met. We show that this model provides a relatively straightforward account for how confidence judgments and their latencies are determined.

We begin by describing the choice followed by confidence paradigm and the set of typical empirical patterns ('Hurdles') that it yields. We follow with a high level description of the debate between single and dual information-collection stage theories of confidence. We then discuss specific extant single and dual stage models of confidence, which are grounded in the sequential sampling framework, and we examine how they fare vis. a vis. the hurdles. With this background, we describe three variants of perceptual availability manipulations, which were designed to probe different aspects and boundary conditions of the post-choice integration hypothesis. Having found support for dual-stage theories, we extend the set of empirical hurdles with respect to the duration of the post-decision stage. Finally, we describe our novel Collapsing Confidence Boundary (CCB) model and show how it accounts for the augmented set of empirical hurdles.

1.1. The choice followed by confidence paradigm

In the current paper we focus on the choice-followed-by-confidence paradigm, in which the observer is first shown a stimulus and asked to decide which of two alternatives (A or B) it matches best. For example, in a perceptual Two-Alternative-Forced-Choice (2AFC) task, the observer may be asked whether the predominant movement direction of a cloud of dots is left or right.[1] After making the decision, the observer is asked to rate his or her confidence in the decision. This procedure allows the collection of four performance measures: (1) choice-correctness, (2) choice reaction time (RT), (3) confidence level and (4) confidence judgment reaction time (RT2).

The choice followed by confidence paradigm yields a dataset rich with dependent variables, which combine and interact to form a complex manifold of empirical relationships. This richness was summarized recently in the list of 'Empirical Hurdles' facing any quantitative confidence theory (Pleskac & Busemeyer, 2010) and provided here in the top part of Table 1 (Hurdles 1–7, see Pleskac & Busemeyer for detailed discussion. Note that Hurdles 8–10 which augment the list of hurdles are described in a later Section 3, 'RT2 in the spotlight'). We next describe the hurdles most relevant for our study (6 and 7) in more detail.

1.2. Resolution of confidence

Resolution of confidence pertains to the relationship between choice correctness and confidence. The simplest definition for *Resolution of confidence* is the difference between the mean confidence in correct choices and error choices. Hurdle 6 states that confidence judgments display a *positive* resolution to the effect that confidence in correct responses is generally higher than confidence in incorrect responses (Ariely et al., 2000; Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979). On first thought, the finding of a positive resolution of confidence might give pause. Indeed, how can observers be more confident when they are correct rather than mistaken? Doesn't this mean that the cognitive system 'knows' that it has made an error? And if so, why did it err in the first place? As we show below, theories of confidence judgment

[1] Such tasks can be carried out in two variants: Whereas in the *free-RT* variant the response time is under the full control of the participants, in the *interrogation* variant the experimenter controls the choice response time (RT) typically by issuing a response signal.

Table 1
The 'Empirical Manifold' of relationships between the variables in the choice followed by confidence paradigm.

Empirical hurdle	Description	Models
Hurdles pertaining to choice and choice latency		
1. Speed–accuracy trade-off[1,2,3]	Decision time and error rate are negatively related	BOE, pipeline, 2DSD, CCB
2. Slow/fast errors	Decision times for erroneous choices can be slower or faster than for correct choices	BOE, pipeline, 2DSD, CCB
Hurdles pertaining to confidence and confidence latency		
3. Negative relationship between confidence and difficulty[1,2,3]	Confidence decreases monotonically as the difficulty level increases	BOE, pipeline, 2DSD, CCB
4. Negative relationship between confidence and decision time[1,2,3]	During free response tasks and within experimental conditions there is a monotonically decreasing relationship between the decision time and confidence	BOE, pipeline, 2DSD, CCB
5. Positive relationship between confidence and decision time[1,2,3]	There is a monotonically increasing relationship between confidence and decision time when decision time is manipulated (e.g., different stopping points in an interrogation paradigm or between speed and accuracy conditions in free response tasks)	BOE, pipeline, 2DSD, CCB
6. Resolution of confidence[1,2,3]	Choice correctness and confidence are positively correlated	BOE, pipeline, 2DSD, CCB
7. Increased resolution in confidence with time pressure on choice[1,2,3]	When choice is made under conditions that stress speed rather than accuracy, resolution of confidence increases	pipeline, 2DSD, CCB; ***BOE predicts a decreased resolution with TP (but see Section 1.4.1.2)***
8. RT2 correlations[1,3]	RT2 is negatively correlated with (1) stimulus discriminability, (2) choice correctness and (3) confidence. RT2 is positively correlated with choice-RT	CCB; ***BOE and the pipeline model are mute with respect to RT2; The interrogation 2DSD predicts a positive correlation between RT2 and confidence***
9. Difficulty and choice accuracy interaction on confidence[1,2,3]	As the difficulty level increases confidence decreases for correct choices and increases for error choices. Thus, the overall resolution decreases.	CCB, BOE, pipeline, 2DSD
10. Difficulty and choice accuracy interaction on RT2[1,3]	As the difficulty level increases RT2 increases for correct choices and decreases for error choices	CCB; ***BOE, the pipeline model and the interrogation 2DSD are mute with respect to RT2***

Note. Hurdles 1–7 are the original list of empirical hurdles described by Pleskac and Busemeyer (2010). Here they are rearranged to facilitate the presentation. Empirical Hurdles 8–10 consist of the set of RT2 correlations, and the novel interaction findings reported here and described in Section 3. Superscripts in the 'Hurdle name' column indicate the Experiments in which each hurdle was tested. The 'Models' column lists the models (among BOE, the pipeline model, the interrogation 2DSD and CCB) that can account for each hurdle. The bold italicized text in the Models column describes problems the models face in accounting for a hurdle. It remains to be seen whether the optional-stopping 2DSD can account for the entire empirical manifold (see Footnote 11).

can account for the finding of a positive resolution while avoiding the apparent 'paradox', whereby the 'system' knows it is going to err and nevertheless still does so.

Hurdle 7 describes an additional important property of confidence-resolution: Time pressure (TP, henceforth) on choice has a *beneficiary* effect on resolution (Baranski & Petrusic, 1994; Pleskac & Busemeyer, 2010). In other words, resolution of confidence *increases* when the decision is made under instructions that stress speed rather than accuracy. As we describe below, the beneficiary TP effect on resolution is diagnostic in its ability to tease apart the two important families of single vs. dual (information-integration) stage theories of confidence (Pleskac & Busemeyer).

1.3. Single vs. dual information-collection stage theories of confidence

Many theories of confidence assume that confidence is a 'single stage' process in the sense that the information underlying both the decision and the confidence judgment is collected in a single, choice-preceding, stage. According to such theories, in order to form a decision observers must collect relevant information, but once a representation of that information is available, it also contains a 'confidence signal' on which confidence is based. For example, a common conceptualization of confidence within the context of signal-detection theories (SDT) states that decision and confidence emerge simultaneously through the scaling of the distance between the perceptual sample and the decision criterion or through the setting of multiple confidence criteria along the decision axis (e.g. Egan, Schulman, & Greenberg, 1959; Kepecs et al., 2008). The Balance of Evidence (BOE; Vickers, 1979) model comprises a second example. According to BOE, observers collect evidence in support of both choice alternatives. The decision is determined by the alternative that is the first to amass a criterion level of support, whereas confidence is determined by the difference between the amounts of evidence supporting the chosen vs. the non-chosen alternatives at the time the decision was made. (See the self-consistency model, Koriat, 2012, for a similar model.) Below, we survey additional single stage theories that are rooted in the sequential sampling framework. Remarkably, while many of these theories have been successful in accounting for *some* of the empirical hurdles (1–7), none of the extant single stage theories can clear *all* the hurdles, with the resolution hurdles comprising an especially tough challenge.

An alternative view posits that, in order to form confidence judgments, observers do not suffice with the information collected during the decisions but rather keep collecting additional information. (The Two-Stage Dynamic Signal Detection model, 2DSD, Pleskac & Busemeyer, 2010; the response-reversals model, Van Zandt & Maldonado-Molina, 2004.) According to these theories, observers take advantage of the inter-judgment time (the time between the decision and confidence responses) to continue accruing decision relevant evidence in a *post-decisional* stage. This continued evidence accrual builds on the evidence that is accumulated during the choice-stage and confidence is determined according to the evidence that was collected during both stages.

Importantly, evidence which is collected after the decision reflects the true identity of the stimulus. Thus, on error trials, post-choice evidence tends to be incongruent with the decision, whereas on correct trials it tends to be congruent with the decision. Consequently, post-choice evidence-integration provides a natural mechanism for the emergence of positive resolution. Any calculation that is sensitive to the congruency between the pre- and post-decision evidence would produce lower confidence in error responses compared to correct responses. Furthermore, below we will show that some dual stage theories can also account for the beneficiary TP effect on resolution, allowing such models to clear all the empirical Hurdles 1–7.

1.3.1. Alternative taxonomies of confidence theories and their relationship with the single vs. dual stage typology

An additional important classification-scheme of confidence theories focuses on the *timing* of confidence processing, specifically, whether confidence is processed during the decision (*decision locus)* or following the decision (*post-decision locus).* In a series of studies, Baranski and Petrusic found that confidence is being processed both during and after the decision (i.e., it is of dual loci) and that the relative proportion of decisional processing is a function of whether choice-speed or choice-accuracy is stressed, such that under accuracy stress decisional confidence processing is more substantial (Baranski & Petrusic, 1998, 2001, Petrusic & Baranski, 2003).

In contrasting the *locus of confidence* and the *number of stages* taxonomies, we note that there are types of post-choice confidence processing, which do not involve post-choice information-collection—the form of post-choice confidence processing that is assumed by dual stage-models. For example, according to BOE, the calculation of the balance (of evidence), which takes place following the choice (Vickers & Packer, 1982), merely interrogates the information that was collected by the time that the decision was made but does not involve the post-decisional collection of novel information. Single stage information collection theories can thus posit that confidence is processed only during the decision (pure decision locus), only after the decision (pure post-decision locus) or both during and after the decision. On the other hand, since dual stage theories assume post-decisional accumulation of novel information, they also necessarily assume that at least some portion of the confidence processing is of post-decisional locus. Importantly, dual stage models go beyond asserting that confidence is processed (also) after the choice in that they specify a particular form of *stimulus* processing (we revisit this issue in the General discussion Section 5.2).

There are additional important typologies of confidence theories, which for the purpose of the current paper, we subordinate to the single vs. dual-stage classification. One such typology of confidence theories pertains to the distinction between 'computational' and 'heuristic' theories. BOE, which was described above, is an example of a computational theory as confidence is *computed* directly from the perceptual information extracted from the stimulus. The time-based hypothesis (Audley, 1960; Ratcliff, 1978; Volkman, 1934; Zakay & Tuvia, 1998) on the other hand, constitutes an example of a heuristic theory. According to this hypothesis, higher decision times yield lower confidence ratings. Such a heuristic is based on the general regularities that hard decisions are both more time consuming than easy decisions and are made with lower degrees of confidence. In the heuristic approach confidence is not computed from the evidence that is extracted during the decision. Still, it is a function of some other property (in this case duration) of the decision process. Thus we can interpret the time-heuristic as a single-stage theory, in the sense that it relies on an aspect of the decision process and does not require the observer to draw novel information from the external environment after the choice has already been made, in order to determine confidence.

1.3.2. The pipeline: a minimal theory of post-choice integration

The pipeline theory (Resulaj, Kiani, Wolpert, & Shadlen, 2009) was presented in the neuroscience literature in the context of changes of mind in perceptual decisions but has not yet been applied to confidence research. According to this theory, there is a lag between the time a decision is reached and the time the response is executed due to response processing and motor execution latencies. During this interval, perceptual information continues to be available, because experimental settings involving free responses terminate stimulus presentation only once a response has been executed. This information continues to feed information-accumulation units resulting in different states at the time of the decision and the time of the response. Nonetheless, once the response is executed the perceptual gate closes and the flow of additional novel external information into the perceptual channel terminates. Thus, post decisional accumulation of novel information is strictly limited to the pipeline and hence, confidence judgments will reflect in addition to the total evidence that guided the decision, a short term (e.g. 200–300 ms) component of pipeline information.

Theoretically, the pipeline model suggests an intermediate possibility between the single and dual stage theories of confidence. In compliance with the fundamental second stage principle, in the pipeline theory confidence is determined by post decisional integration of information. On the other hand, in the spirit of single stage theories, once the decision is formed observers cease to actively seek novel external (perceptual) information. Thus, the pipeline theory could be construed as 'a minimal dual-stage theory' and can serve as an important benchmark in gauging the boundary conditions and the extent of post-choice integration. Specifically, when empirical evidence in support of post-choice integration is found, one should ask whether a pipeline theory can account for such evidence. An affirmative answer suggests that the extent of post-choice integration might have been very limited, perhaps even involuntary. A negative answer, on the other hand, implies that the post-choice stage was more extended and deliberate. In the current paper, the pipeline theory will thus serve as a yardstick, which will allow us to transcend beyond the binary yes/no question of whether post-choice integration has occurred and to probe the extent to which it has occurred.

Is confidence a single or a dual-stage process? In the following section we examine in more detail extant confidence theories and show how the empirical hurdles and especially, the positive resolution property of confidence judgments and the TP effect (Hurdles 6–7), can shed light on the single vs. dual stage debate and help to tease these families apart.

1.4. Extant models of confidence

The sequential sampling framework has been highly successful and influential in modeling choice. Naturally, most theories of the choice followed by confidence paradigm are grounded in this framework. In Appendix A, we provide a brief summary of the sequential sampling framework for binary choice, focusing on the differences between the diffusion and accumulator models (we refer readers to more extensive presentations of sequential sampling theory for additional details e.g. Gold & Shadlen, 2007; Ratcliff & McKoon, 2008; Teodorescu & Usher, 2013). We next describe some prior attempts to model confidence within this framework.

1.4.1. Single information-integration stage confidence theories

Because diffusion and accumulator based theories have offered different avenues for modeling confidence, we describe each of these paths in turn.

1.4.1.1. Diffusion based single stage confidence theories. Diffusion models accumulate evidence differences towards a threshold value θ (see Fig. 1). Thus, all decisions terminate when one alternative has an advantage of θ over the other. Importantly, any attempt to base confidence *solely* on the difference in evidence in favor of the chosen alternative is bound to fail, because it leads to the false prediction that all trials should have *identical* confidence. Such a theory could obviously not account for the increase in confidence with increase in stimulus discriminability (Hurdle 3) and for the positive resolution of confidence (Hurdle 6).

A more normative assumption is that confidence corresponds to the posterior probability of the decision, given the stream of evidence. Such a theory was proposed by Kiani and Shadlen (2009), for a task in which trials of different difficulty were intermixed within an experimental block. Importantly, when drift rate varies across trials, the diffusion mechanism accommodates for differences in posterior probability of decisions that are executed for the same threshold level of evidence. Indeed, Kiani & Shadlen showed that the likelihood ratio is a monotonically decreasing deterministic function of the accumulation time. However, when a single difficulty level exists within an experimental block the model predicts that all trials terminate with the same posterior probability for the chosen alternative and hence, it is subject to all the criticism described above.[2] Furthermore, even when difficulty levels are mixed, this model cannot account for confidence effects that survive control for RT, because confidence is a deterministic function of RT (for a given choice threshold). One such example is the ‘RT-difficulty effect’ (Baranski & Petrusic, 1994; Kiani, Corthell, & Shadlen, 2014) according to which easier stimuli generate higher levels of confidence than hard stimuli even when decision RT is controlled for.

1.4.1.2. Accumulator based single stage confidence theories. Accumulator models offer an alternative route to model confidence. This approach relies on the fact that unlike diffusion models, in which, at the time of choice, the amount of evidence favoring the chosen *over the non-chosen alternative* is constant, in accumulator models, it is variable. Capitalizing on this principle, Vickers (1979, 2001) proposed the *Balance of Evidence* hypothesis (BOE), according to which confidence is a monotonically increasing function of the difference between the amounts of evidence each counter has accumulated by the time of the decision. Paired with the BOE hypothesis, accumulator models can account for empirical Hurdles 1–6. However, in violation of Hurdle 7, BOE predicts decreased resolution with increased time pressure (Pleskac & Busemeyer, 2010). According to BOE, lower choice thresholds, induced by increased time pressure, lead to lower accumulated values at decision and thus to lower balances of evidence. Consequently, balances become similarly low for both correct and error choices, leading to a low resolution. Importantly, this conclusion should be qualified as its supportive argument assumes that TP manipulations on choice selectively influence choice thresholds. However, TP manipulation can additionally influence the balance-to-confidence mapping (Baranski & Petrusic, 1998; Pleskac & Busemeyer, 2010). Furthermore, it has recently been shown that TP manipulations can affect evidence accumulation rates (Rae, Heathcote, Donkin, Averell, & Brown, 2014). Finally, Kiani et al. (2014) have recently introduced a ‘bounded accumulation model’ according to which,

[2] When the diffusion model consists of a single difficulty level and there is no across trial variability in drift rate and starting point, the model for decision is equivalent to the normative sequential ratio probability test (SPRT; Gold & Shadlen, 2007; Moran, 2014; Wald & Wolfowitz, 1948). In SPRT the decision variable corresponds to the posterior log likelihood ratio between the choice alternatives, conditional on the observed stream of evidence, and evidence is integrated until this likelihood ratio reaches a criterion level. Thus, if confidence is based on the likelihood ratio, this normative single-difficulty model faces the same problem that we described above for diffusion to boundary.

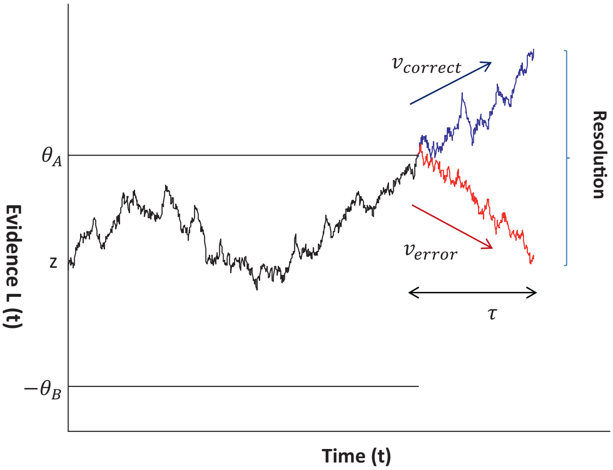

Fig. 1. The causal locus of positive resolution in 2DSD. The black trajectory, reaching the higher threshold, corresponds to a trial terminating in choice of alternative A. At this point the trial splits: If 'A' is the correct answer then during the second stage the diffuser will, on average, continue to grow (blue trace), whereas if 'A' is the wrong answer—it will tend to decline (red trace). By the end of the second stage the blue correct trace has accrued more support; hence confidence in favor of choice A will be higher than for the red erroneous choice. Positive resolution ensues. The vectors v_{correct} and v_{error} illustrate the mean post-choice drift rates for correct and erroneous trials, when A is chosen. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

confidence is a function of not only the balance of evidence, but also choice RT. Revisiting in future research the question of whether the balance of evidence (in itself or together with the decision time) can account for the beneficiary TP effect on confidence thus seems advisable.

1.4.2. Dual information-collection stages theories of confidence

Unlike single stage theories, according to dual-stage theories, people can base their confidence on additional sources of information besides the one that lead to the actual decision. This allows dual stage theories to easily account for the positive confidence-resolution. Specifically, as explained above, any information about the stimulus that is collected following the decision, correlates positively with the true stimulus and hence either supports correct decisions or counteracts erroneous ones. Thus positive resolution would be generated by any process that is sensitive to the congruency between post-decision evidence and the choice itself. In the current section we describe how specific dual-stage confidence models can account for the empirical results, particularly the resolution hurdles.

1.4.2.1. The interrogation two-stage dynamic signal detection model. Unlike the single-stage models, in the 2DSD model, the decision, formulated as a diffusion process, is only the first stage of a dual-stage process. In the second, post-decisional stage, judges continue to accumulate evidence about the alternatives. The total amount of evidence collected from the onset of the stimulus, before and after the choice, is compared with confidence criteria to determine the confidence level. In the interrogation 2DSD variant, post-decision evidence accumulation is modeled by continuing evidence accumulation for a fixed period of time, dubbed *inter-judgment time* and denoted by τ (see Fig. 1). The time of the confidence judgment, denoted t_C, is the sum of the time of the decision t_D and the inter-judgment time τ. This model further assumes that judges map the total amount of evidence $L(t_C) = L(t_D + \tau)$ into a confidence scale by setting confidence criteria at specific values of evidence, in a manner that is analogous to signal detection theory (e.g., Macmillan & Creelman, 2005). The basic idea underlying this evidence-to-confidence mapping is that higher amounts of evidence exceed higher confidence criteria and hence result in higher levels of confidence. Pleskac and Busemeyer (2010) demonstrated that the 2DSD model can account for Hurdles 1–7.

Why does 2DSD avoid the pitfalls that thwart the single stage diffusion attempts to model confidence? The crux of the answer is that whereas the amount of evidence accrued by the end of the first stage is constant (across all decisions given a decision threshold), by the end of the second post-decision stage, it is variable. 2DSD capitalizes on this property. For example, stimuli with different discriminability are indistinguishable with respect to the total relative amount of supporting evidence at the time of decision. Nevertheless, during the second stage the more discriminable stimuli will generally recruit higher amounts of evidence due to their higher drift rates. Thus, by the end of the second stage, more evidence will be accrued the more discriminable the stimulus, leading to higher confidence (Hurdle 3).

How does 2DSD account for a positive resolution and for the beneficiary TP effect? During the second information-collection stage, the post-decisional drift is congruent with correct choices and incongruent with erroneous ones, hence a positive resolution ensues (Hurdle 6, see Fig. 1). In accounting for the beneficiary time pressure effect (Hurdle 7), Pleskac and Busemeyer (2010) relied on the empirical finding that RT2 was longer under conditions that stress choice speed rather than accuracy. In 2DSD resolution emerges due to the systematic difference in second stage support for correct and error trials and hence, the longer this stage (RT2), the higher the resolution.

1.4.2.2. The pipeline model. The pipeline model (Resulaj et al., 2009) is similar to 2DSD in that it features a diffusion-like decision process.[3] However, unlike the 2DSD, in the pipeline model the post-choice stage is limited to the integration of perceptual information that is sampled in the temporal lag spanning between the formation and the execution of the choice. Importantly, this model can account qualitatively for Hurdles 1–6 in the same way as 2DSD can and can also account for the beneficiary TP effect on resolution (Hurdle 7; This is shown in Appendix B). As explained above, this model served us as a benchmark for gauging the extent of post-choice information-integration.

1.4.2.3. The optional stopping 2DSD model. In the 2DSD interrogation model inter-judgment time is treated as an *exogenous* variable. This means that rather than accounting for RT2, this model accounts for confidence and its relations with the other variables (in the choice followed by confidence paradigm) based on the observed empirical RT2. The *optional stopping* 2DSD (Pleskac & Busemeyer, 2010) is a second and more sophisticated 2DSD variant wherein, RT2 is an endogenous variable i.e., it is governed by the model mechanism, rather than fed extrinsically into the model.

In the optional stopping 2DSD, the first choice stage is identical to that of the interrogation 2DSD, a standard diffusion process, which is formulated as a Markov chain. The second stage is different though: Markers κ_j are placed along the evidence state space representing the different confidence ratings ($j = .50, .60, \ldots, 1.00$), one for each rating. As the second-stage unfolds, this Markovian-diffuser occasionally meets the markers. The 'extreme' markers corresponding to confidence .50 or 1.00 are absorbing i.e., the second stage integration is terminated and the corresponding confidence judgment is issued as soon as either of these markers is reached. When evidence passes one of the other intermediary markers (.60, .70, .80, and .90) there is a probability w_j that the judge exits and gives the corresponding confidence rating. Both the markers and the exit probabilities are free model parameters (see Pleskac & Busemeyer, 2010 for further details). Due to its ability to account for RT2, the optional stopping rule 2DSD is a promising model for all four dependent variables in the choice followed by confidence paradigm (we discuss this model further in Section 3.2).

1.5. Single vs. dual stage confidence theories: an intermediate summary

The upshot of our survey of extant confidence-theories is that presently, only dual stage theories are able account for the entire set of hurdles. This fact in itself, however, lends only *indirect* support for dual over single stage theories because it remains possible that a future single stage theory (or perhaps even an extant theory with relaxed assumptions, see our presentation of the BOE in Section 1.4.1.2) would be able to clear all hurdles. This consideration begs the necessity to conduct *direct* empirical tests of the principal assumption of dual-stage theories, i.e. post-choice information-integration.

[3] Note that the idea of a pipeline could also be implemented in the framework of accumulator models.

2. Experimental investigations of post-choice information collection

2.1. Overview of the experiments

The main purpose of our experiments was to test directly the fundamental principle distinguishing single from dual stage theories of confidence: Do people use post decision information towards forming confidence judgments? If the answer is affirmative, then is the post decisional integration short and limited to a pipeline or is it more extended in time? And if post-choice integration extends beyond a pipeline, does the duration of this stage exert a causal role on confidence?

An additional goal of the experiments was to examine in more detail the beneficiary TP effect on resolution (Hurdle 7). This finding is particularly important because it was pivotal in favoring dual over extant single stage theories and it could constrain future theories. Does the beneficiary TP effect on resolution depend on the perceptual availability of the stimulus following the decision? Does this effect empirically depend on longer RT2 in speed vs. accuracy condition (Pleskac & Busemeyer, 2010) or is the effect found even when RT2 is not longer in the speeded condition? Puzzlingly, and anticipating our results, we replicated the beneficiary TP effect even when the speeded-choice condition was not associated with a higher RT2. We thus examined whether in principle, 2DSD can predict a beneficiary TP effect even when RT2 is not shorter in the speeded choice condition. In Appendix B we present a simulation study in which we found that longer RT2 in the speed condition is not a necessary condition for the beneficiary TP effect and that a beneficiary TP effect ensues if RT2 is equal or even moderately lower in the speeded choice condition. This finding is a subtle consequence of drift rate variability and the fact that the distributions of drift rates conditional on correct and erroneous responses vary as a function of the choice threshold.

Finally, having found direct evidence supporting post-choice integration, we set out to explore additional robust relationships between RT2 and the other variables in the choice followed by confidence paradigm. Our purpose was to augment the empirical manifold with additional empirical patterns, which can provide further constraints on future theorizing of confidence.

To test these issues, we build on the *decision followed by confidence* paradigm and introduce a novel manipulation designed to experimentally control the perceptual availability of post-decisional stimulus information. As soon as the participant executed their choice (i.e. overt response), the stimulus either remained or vanished from the visual display (*remain* and *vanish* conditions respectively). Presumably, the *remain* condition should provide more favorable conditions for post-choice integration than the *vanish* condition. Granted, even when the stimulus disappears from the display, participants may be able to collect additional information by relying on their visual memory (Pleskac & Busemeyer, 2010) and/or on the information flowing through the pipeline. We assume, however, that the perceptual channel is both a more reliable (less noisy) and durable source of novel information (Sperling, 1960). Thus, by continuing to integrate evidence from the visual display (rather than from memory) participants should achieve higher levels of resolution.

Our basic philosophy in trying to distinguish empirically between single and dual-stage theories of confidence is as follows: If confidence relies on a single stage process or even on a pipeline to that effect, then all the external confidence-relevant perceptual information is available in the visual display and feeds into the perceptual channel *prior to the response*. Thus, the single stage and pipeline theories predict no sensitivity to a post-decision perceptual availability manipulation. More specifically, no difference is predicted with respect to resolution and the TP effect on resolution between a condition where the stimulus remains available on the screen (the *remain* condition) and a condition where, after the execution of a choice response, the stimulus disappears and thus is made unavailable (the *vanish* condition). Unlike single stage and pipeline theories, dual stage theories, which assume a post-choice integration stage that extends beyond the pipeline, predict a higher resolution and a stronger beneficiary TP effect for the *remain* than for the *vanish* condition, due to the improved opportunity for post-decisional integration. Thus, the perceptual availability manipulation constitutes a direct test of whether evidence is integrated following the decision (beyond a pipeline).

In addition, controlling the duration of the post-choice interval allows a direct test of whether post-choice integration has a flexible or constant time course and whether the duration of this stage exerts a causal influence on confidence. If post-choice integration is flexible (unlike a pipeline) then resolution and the beneficiary TP effect should increase with longer post decision intervals in the *remain* condition more than in the *vanish* condition.

Next, we describe our experiments in detail. To simplify the presentation, we report the results of the experiments in two 'waves'. In the following sections (Wave-1) we describe each of the experiments in turn, with a focus on analyses that are most directly related to the questions of whether post-choice integration occurs and to what extent. We defer the discussion of additional analyses pertaining to the inter-judgment time and its relations with the other variables, to a later section, 'RT2 in the spotlight' (Wave-2). This later section (Section 3) is focused on extending the empirical manifold with robust RT2 related empirical hurdles.

2.2. Experiment 1

In Exp. 1, six participants were asked to choose which of two black and white arrays, contained a larger number of black squares (Ratcliff & Smith, 2010). One array contained an equal number of black and white squares whereas the other, target array, contained a majority of black squares. To manipulate choice difficulty, the proportion of black squares in the target array varied randomly across trials between three levels. Following the choice, participants rated their confidence in the correctness of their choice on a scale between 50% and 100%. Time pressure on choice was manipulated by stressing either choice-speed or choice-accuracy across different blocks of trials. Confidence judgments though, were given under no time pressure, and feedback was given at the end of each trial using a normative measure that depends on both the accuracy of the choice and the confidence judgments (Brier, 1950). Critically, Exp. 1 contained the novel perceptual availability manipulation. As soon as the participant made a choice, these arrays either remained or vanished in a random manner from the visual display (the detailed experimental methods for all experiments are described in Appendix C).

2.2.1. Results

As a first step in the data analyses we removed trials that were likely the result of 'contaminating' processes (Ratcliff & Tuerlinckx, 2002). To minimize fast outliers, we excluded trials for which either the decision times were less than 0.2 s or the observed inter-judgment times were less than 0.15 s. To minimize slow outliers, we excluded trials where either the decision time or observed inter-judgment time was greater than 3 SDs from the mean. Finally we excluded trials that were aborted due to pressing a non-eligible key during the trial. These cutoffs eliminated on average 5.2% (min 4.1%; max 6% per participant) of the data.

Throughout the article we report statistics both at the level of individual participants and for the entire group. In some of the following analyses, standard errors for individual participants were calculated based on 10,000 nonparametric bootstrap samples and significance was examined using permutation tests with 10,001 random permutations (the use of these randomization methods is indicated in table captions). Group statistics were always based a meta-analysis, where each participant's data were treated as a separate experiment and the average statistic is calculated by weighting each participant's respective statistic by the inverse of the variance of the statistic, assuming a random effects model (Borenstein, Hedges, Higgins, & Rothstein, 2011, chap. 12; Shadish & Haddock, 1994). Broadly speaking, when given as input a series of effect estimates and standard errors of these estimates for each participant, this method outputs a group-level effect estimate, a corresponding group-level standard error, a z-statistic to test the hypothesis that the group-level effect is non-zero and a p-value.

The time pressure manipulation on choice was effective in generating a speed accuracy tradeoff. For all participants as well as for the group as a whole, both the accuracy rate and the mean decision time were smaller for the speed vs. the accuracy condition (group Accuracy: $M_{\text{speed}} = .69, M_{\text{accuracy}} = .83, z = 6.66, p < .001$; Group RT: $M_{\text{speed}} = 0.51\ s, M_{\text{accuracy}} = 1.02\ s, z = 7.06, p < .001$; the results for the individual participants are reported in Table G1). Additionally, for all participants and for the entire group the mean confidence was higher in the accuracy ($M = .89$) than in the speed condition ($M = .85, z = 7.09, p < .001$). This finding supports Hurdle 5 according to which, confidence and decision time are positively correlated. Finally, the mean inter-judgment time (RT2) was higher for the speed ($M = 0.70$ s) than the accuracy condition ($M = 0.62$ s, $z = -2.00, p < .05$). While this difference was significant for the group as a whole, it was significant for only two of the individuals (Participants 4 and 5; see Table G1). Interestingly, Participant 3 showed the opposite effect, as her RT2 was higher for accuracy trials, and for the remaining three participants no significant difference was found. The four participants for whom RT2 was not significantly larger in the accuracy condition, allowed us to examine whether the higher speed vs. accuracy RT2 is a necessary condition for the beneficial effect of time pressure on resolution (see discussion Section 2.2.2).

Next, we examined the effects of the perceptual availability manipulation: With respect to confidence, there was no significant difference between the *remain* and the *vanish* conditions ($M = .87$ for both conditions, $z = -0.71$). Additionally, RT2 was significantly larger for the *remain* ($M = 0.73$ s) than the *vanish* condition for the entire group ($M = 0.60$ s,$z = -3.21, p < .01$) and for five of the individuals. This finding offers preliminary support for the dual stage theories, if one assumes that when the stimulus remains perceptually available, participants take more time for the execution of the post-choice integration stage in an attempt to take advantage of the better post-choice integration condition. We now turn to the focal set of analyses that pertain to resolution of confidence.[4]

There are many different definitions for resolution of confidence. We thus begin with a brief description of the main measure that was used in our analyses (c.f. Nelson, 1984, for a detailed analysis and discussion about the differences between measures of resolution). To recapitulate, resolution of confidence pertains to the relation between confidence level and choice-correctness. Perhaps the simplest operative definition of this relation is given by the slope score (Yates, 1990), which is the difference between mean confidence for correct and incorrect decisions. Importantly, this measure assumes that the values of the confidence judgments emerge from the use of an interval scale. However, this assumption may be problematic in the current paradigm due to the special status of the '50%' confidence category. Because participants are not given the opportunity to report errors, they may use the 50% confidence category when they think they most likely erred. Therefore the 50% confidence category confounds 'real', 50% confidence judgments (i.e. guesses) with 'false' 50% judgments (i.e. higher than 50% confidence in an error or vice versa, lower than 50% confidence in a correct response). To mitigate such concerns, we measured resolution using an ordinal regression analysis which minimally postulates an ordinal structure of the confidence scale. To verify the robustness of our results, we also conducted analyses with additional measures, including slope, DI′, the Gamma correlation and Ag (see Appendix E for further details and analyses results). By and large, all measures yield similar results.

We calculated several effects involving resolution for each participant with a multiple probit-ordinal regression (Dobson & Barnett, 2008; Long, 1997; McCullagh & Nelder, 1990), using Matlab's 'mnrfit' function. We coded for each trial the confidence response in a variable 'CONF' with labels 1–6, corresponding to decreasing levels of confidence (from 100% to 50%). We coded the speed-accuracy tradeoff (henceforth SAT) condition (for each trial) in a variable 'TP' as 0.5 (speed) and -0.5 (accuracy) and the Perceptual Availability condition in a variable 'PA' as -0.5 (vanish) and 0.5 (remain). Finally, we defined a trial choice-correctness indicator, 'CORRECT', with values 0-error and 1-correct. We then regressed CONF on TP, PA and CORRECT (as main effects), the three double interactions and the triple interaction between these variables. We also examined the separate 'simple' effects of CORRECT, TP and their interaction for the different values of PA i.e., for vanish and remain trials. This was achieved by repeating the regression but with the variable PA′ = PA + 0.5 (vanish) or with PA ′ = PA − 0.5 (remain) replacing PA.

Table 2 displays the regression coefficients for the effects involving choice-correctness. The first data column presents the main effect of choice-correctness on confidence, i.e., the resolution. In accordance with Hurdle 6, all of the participants and the group exhibited a positive resolution. Furthermore, a positive correlation was found for the group and for all of the participants separately for vanish trials ($b_{\text{group}} = 0.94, z = 9.70, p < .001$) and for remain trials ($b_{\text{group}} = 1.26, z = 6.01, p < .001$). The second data column of Table 2 lists the perceptual availability effect on resolution (i.e., the PA$*$CORRECT interaction). At the group level, as well as for Participants 2 and 4, resolution was larger for the *remain* vs. the *vanish* condition.

[4] We also confirmed that the additional confidence-related hurdles (3–4) where replicated in our data. Because these analyses are not our main interest here, they are reported in Appendix F.1.

Table 2
Resolution of confidence effects (regression slopes) for Exp. 1.

Par	CORRECT	PA∗CORRECT	TP∗CORRECT	PA∗TP∗CORRECT
1	$0.58(0.05)^{***}$	$0.07(0.10)$	$-0.22(0.10)^{*}$	$0.02(0.21)$
2	$1.54(0.07)^{***}$	$0.78(0.13)^{***}$	$0.56(0.13)^{***}$	$0.55(0.25)^{*}$
3	$0.96(0.06)^{***}$	$0.17(0.12)$	$0.50(0.12)^{***}$	$0.08(0.24)$
4	$1.58(0.09)^{***}$	$0.79(0.18)^{***}$	$1.53(0.18)^{***}$	$0.81(0.35)^{*}$
5	$1.11(0.05)^{***}$	$0.20(0.10)$	$0.23(0.10)^{*}$	$0.22(0.20)$
6	$0.88(0.05)^{***}$	$-0.09(0.11)$	$0.23(0.11)^{*}$	$0.14(0.22)$
Group	$1.10(0.15)^{***}$	$0.30(0.14)^{*}$	$0.46(0.19)^{*}$	$0.23(0.10)^{*}$

Note. Values in parentheses are standard errors. ∗, ∗∗, ∗ ∗ ∗ indicate $p < .05$, .01, .001, respectively according to t-test for the participants and a meta-analysis for the group.

The third data column of Table 2 presents the TP effect on confidence (the TP∗CORRECT interaction). In accordance with Hurdle 7, for the group and for Participants 2–6, time pressure had a positive effect on resolution, whereby resolution was higher in the speed than in the accuracy condition. Interestingly, Participant 1 showed a negative TP effect, that is, time pressure on choice decreased her resolution (we return to this finding in the discussion below). When we examined the TP effect separately for vanish trials we found a significantly positive effect for Participants 3 and 4. For the group the trend was positive but was significant only according to a one-sided test ($b_{\text{group}} = 0.28, z = 1.87, p = .03$). For remain trials, the TP effect was positive at the group level ($b_{\text{group}} = 0.60, z = 2.43, p < .05$) and for Participants 2–5. Finally, the fourth data column in Table 2 displays the perceptual availability effect on the TP effect on resolution (the triple interaction). For the group and for Participant 2 and 4, the TP effect was larger for the *remain* than for the *vanish* condition.

2.2.2. Discussion of results

The current findings replicated all the empirical hurdles (1–7) and in particular the beneficial effect of choice-TP on resolution of confidence. Furthermore, the inclusion of the novel post-choice-perceptual availability manipulation demonstrated that perceptual events that occur following the choice affect confidence. We found an increased resolution of confidence and an increased TP effect on resolution when, following the decision, the stimulus remained on the screen rather than vanished. Notably, these findings are compatible with, and indeed are predicted, by dual stage theories such as 2DSD, which assume that confidence is based on a second evidence integration stage. When the stimulus remains on the screen rather than vanishes, the conditions for post-choice integration are more favorable since stimulus perception is not subject to decay as are mnemonic representations. Hence, both resolution and the TP effect on resolution are predicted to be larger in the *remain* condition. Furthermore, the results of Exp. 1 suggest that participants can capitalize on the better available information by increasing their post-decision integration times (RT2 was larger in the *remain* than in the *vanish* condition). In contrast, both single stage theories and the pipeline model generate confidence based *only* on information that flows into the ‘perceptual channel’ prior to the *motor* choice-response. These approaches are thus mute with respect to the effects of the post-decisional perceptual availability manipulation on resolution (and the TP effect) and on RT2.

Participant 1 constitutes an interesting exception to this rule. For her, the positive resolution effect did not differ for the *remain* vs. the *vanish* conditions. Furthermore, her resolution was higher for the accuracy than for the speed condition. Whereas this pattern of findings is not predicted by 2DSD it is predicted by BOE. Conclusions that are based on a single participant should be interpreted with caution but nonetheless, Participant 1 may indicate that there are important individual differences with respect to the number of information collection stages.

One possible interpretation of the differences between the *remain* and the *vanish* conditions is that a post-choice integration stage occurs only in the *remain* but not in the *vanish* condition. According to this explanation, when the stimulus remains available for perception following the choice, participants seize on the opportunity to collect additional information and hence, confidence is based on a dual-stage integration process. However, when the stimulus vanishes—confidence is based on a single-stage process such as BOE. Can we also infer that some form of post-choice integration is operative in the vanish condition? Interestingly, for all individuals and the group, resolution was significantly positive in the *vanish* condition, a finding that is consistent with BOE but also with the pipeline model and with the assumption that additional, beyond-pipeline information, may be extracted from visual memory. These models make different predictions with respect to the TP effect though. Whereas BOE predicts a negative TP effect, the pipeline and dual stage theories predict positive effects. With respect to the TP effect on resolution in the *vanish* condition, we found a significantly positive TP effect for Participants 3 and 4 but not at the group level. We revisit this question in Exp. 3, where we present more conclusive evidence for post-choice integration even under stricter, masking, conditions.

Finally, the results of Exp. 1 contribute to our understanding of the TP effect and the role that RT2 plays in generating this effect. Recall that prior accounts for the beneficial TP effect on resolution relied on the increased RT2 under choice-time pressure (Pleskac & Busemeyer, 2010). While for the entire group in our experiment, RT2 was longer for the speed condition, for four of the individuals the effect was slight (and consequentially non-significant) or even significantly opposite in direction (Participant 3). Still, for three of these individuals (Participants 2, 3 and 6) a positive TP effect emerged. The upshot is that an increased RT2 under choice-TP seems to be unnecessary for TP to have a beneficiary effect on resolution. Indeed, as we show in Appendix B, a beneficial TP effect on resolution could ensue from the different drift rate distribution conditional on choice correctness under TP.

2.3. Experiment 2

Exp. 2 was similar to Exp. 1 but differed in one fundamental respect (the detailed experimental methods are provided in Appendix C.2). In Exp. 2, time pressure on confidence rather than on choice was manipulated across experimental blocks. For all trials, choice was made under time pressure (we stressed choice-speed rather than accuracy to obtain higher levels of resolution). Confidence, on the other hand, was made under varying degrees of time pressure. Following their choice, the six participants were instructed to wait for an auditory signal (a short beep) before giving their confidence judgment. As soon as the auditory signal was played, participants were asked to rate their confidence quickly. Critically, across blocks the beep was issued either 300 ms ('early-confidence' condition) or 1300 ms ('late-confidence' condition) after the choice.

According to dual-stage theories, which extend accumulation of information beyond the pipeline, resolution should increase with longer inter-judgment intervals (see Fig. 1 for illustration). Thus, resolution is predicted to decrease with shorter time windows for forming confidence judgment, a prediction which we dub the *harmful confidence-TP effect on resolution*. Importantly, such a finding would also demonstrate that the duration of the inter-judgment interval exerts a *causal* influence on confidence. Furthermore, since conditions for post-choice integration are more favorable in the *remain* condition, both resolution of confidence and the harmful confidence-TP effect on resolution are predicted to be more pronounced when the stimulus remains available for the duration of the inter-judgment interval. Single stage and pipeline theories of confidence, on the other hand are again mute with respect to the perceptual availability manipulation.

2.3.1. Results

We examined the effect of the confidence delay manipulation on confidence and on RT2. Here, RT2 denotes the confidence judgment response time from the response signal (the beep). Thus RT2 is not identical to the total inter-judgment time as the latter includes also the duration between the decision and the arrival of the confidence response signal. At the group level confidence was lower for the late confidence condition ($M_{\text{early}} = .84, M_{\text{late}} = .83$, $z = -3.36$, $p < .001$) and there was no significant difference in RT2 ($M_{\text{early}} = 0.34\text{s}, M_{\text{late}} = 0.31, z = -1.19, p = .23$).[5] Additionally, we examined the effect of the perceptual availability manipulation on confidence and RT2. At the group level, there was no

[5] The confidence delay manipulation had no significant effect either on choice accuracy or on choice RT (See Table G2, which also reports individual participant-results). As in Exp. 1, in a preliminary step in the data analyses, we removed trials that were likely contaminant trials. We used the same cutoffs as in Exp. 1, which resulted in the elimination of an average 3.7% (min 3.3%; max 4.4%) of the data. Additionally, we confirmed that the other confidence-related hurdles (3–4) where replicated in our data and these results are reported in Appendix F.2.

Table 3
Resolution of confidence effects (regression slopes) for Exp. 2.

Par	Correct	PA∗Correct	TP∗Correct	PA∗TP∗Correct
2	1.94(0.06)***	0.35(0.11)**	−0.2(0.11)	−0.35(0.22)
3	1.61(0.05)***	0.27(0.09)**	−0.25(0.09)**	−0.31(0.19)
4	2.04(0.06)***	1.01(0.12)***	0.15(0.11)	−0.38(0.23)
6	1.47(0.05)***	0.37(0.09)***	−0.41(0.09)***	−0.45(0.18)*
7	1.33(0.05)***	0.28(0.09)**	−0.04(0.09)	−0.09(0.18)
9	1.18(0.05)***	0.24(0.09)**	−0.16(0.09)	−0.10(0.18)
Group	1.59(0.13)***	0.41(0.10)***	−0.16(0.08)*	−0.27(0.08)***

Note. Values in parentheses are standard errors. $*$, $**$, $***$ indicate $p < .05$, .01, .001 respectively according to t-test for the participants and a meta-analysis for the group.

significant difference in the mean confidence level between the *remain* and the *vanish* conditions ($M = .84$ for both conditions, $z = 0.004$, $p = .997$) but RT2 was longer in the *remain* vs. the *vanish* conditions ($M_{\text{remain}} = 0.33$ s,$M_{\text{vanish}} = 0.32$ s,$z = -2.15, p = .03$). We discuss this RT2 effect below.

Next, we turned to the focal set of resolution analyses. We performed an ordinal probit-regression analysis similar to Exp. 1 with a single change. Here, the variable TP coded whether the confidence beep sounded early (coded as 0.5) or late (coded as -0.5) after the choice. The first data column in Table 3 shows that in accordance with Hurdle 6, all the participants and the group displayed a positive resolution (CORRECT main effect). Notably, a positive resolution was found for the group and for all of the participants separately for vanish trials ($b_{\text{group}} = 1.38, z = 13.69, p < .001$) and for remain trials ($b_{\text{group}} = 1.80, z = 10.66, p < .001$). Furthermore, the second data column in Table 3, shows that for all participants as well as for the group, resolution was higher for the *remain* than for the *vanish* condition (a positive PA∗CORRECT interaction). The third data column of Table 3 shows that for the group (and Participants 3 and 6) resolution was lower in the early compared to the late confidence condition (a negative TP∗CORRECT interaction effect). Thus, TP on confidence is generally *harmful* with respect to resolution of confidence, supporting a causal role for RT2 in the formation of confidence judgments. However, when we examined vanish and remain trials we found that the harmful confidence-TP effect was confined to remain trials: for vanish trials the effect was virtually nil ($b_{\text{group}} = -0.03$, $z = -0.38, p = .71$), whereas for remain trials, the TP effect was negative at the group level ($b_{\text{group}} = -0.30, z = -3.19, p < .01$) and for Participants 2, 3 and 6. The fourth data column in Table 3 confirms that at the group level and for Participant 6, the confidence-TP effect was *more harmful* for the remain condition (a negative triple interaction effect).

2.3.2. Discussion of results

The results of Exp. 2 illuminate several aspects of the relationship between resolution of confidence and inter-judgment times. First, we found that unlike the beneficial effect of time pressure on choice, time pressure on confidence is harmful with respect to resolution. This was evident in the higher resolution exhibited in the late vs. early confidence conditions. Since the duration of the post-choice integration stage was manipulated in the current experiment, this finding shows that this variable exerts a causal influence on confidence judgments (and their resolution). Second, as in Exp. 1, we found that the perceptual availability manipulation had an effect on confidence responses. The resolution of confidence and the harmful influence of TP on resolution were all higher in the *remain* than in the *vanish* condition. These effects can only be accounted for, indeed predicted by, dual stage post-decisional integration theories. If resolution and the harmful confidence-TP effect on resolution result from the shorter temporal window for post decision integration in the early vs. late confidence condition, then these effects should be more pronounced when post-decisional conditions are more supportive of such integration i.e. in the *remain* condition. Interestingly, RT2 was also higher in the *remain* relative to the *vanish* condition. One speculation, which is consistent with the operation of a post-decisional integration stage, is that if participants are still perceptually engaged in extracting evidence from the stimuli when the confidence signal is issued, then disengaging from the stimulus will slowdown responding.

Admittedly, some of the effects of Experiment 2 were moderate, especially when evaluated at the level of individual participants. Note however, that the total inter-judgment times in the early-confidence condition, which consist of the sum of the 300 ms interval (between the choice and the response signal) and of RT2 are comparable to the spontaneous inter-judgment times of Exp. 1. Furthermore, we had no means to force participants to engage in thoughts about confidence for the entire duration up to the arrival of the beep, especially in the late condition, for which the total inter-judgment duration is much higher than the spontaneous durations of Exp. 1. Thus, it is likely the participants actually decided about their confidence earlier than the late beep and that they simply delayed their report until the cue. Consequentially, we find that the effects exerted on resolution of confidence by the confidence-TP manipulation, however moderate, are remarkable in demonstrating that participants did capitalize, to some extent, on the additional time that was provided in the late confidence condition. Critically, post-choice evidence-integration theories can account for the current results as long as the post choice integration time is at least moderately larger in the late confidence condition.

Finally, as in Exp. 1, we asked to what extent is post-choice integration operative in the *vanish* condition? First, resolution was positive in the *vanish* condition. These findings are consistent both with the pipeline model and with a more extended post-choice integration from visual memory. However, the pipeline model predicts no confidence-TP effect on resolution, whereas the more temporally extended assumption of information extraction from a visual mnemonic representation predicts a negative confidence-TP effect. At the group-level, the confidence-TP effect was close to zero. Thus, we cannot rule out the possibility that in the vanish condition post-choice integration was limited to the perceptual pipeline.

2.4. Experiment 3

In Exp. 3 we aimed to probe the boundary conditions of post-choice integration, by masking the stimulus once the decision was made. The design of the experiment, which included nine participants, was identical to that of Exp. 1 with a single difference. Rather than alternating between the remain and vanish conditions, the stimulus always vanished and was additionally masked immediately after the choice. (The full experimental methods are described in Appendix C.) The goal of the masking was to try and interfere with the post-decision integration stage. We reasoned that the mask will abolish, or at least severely degrade, the quality of the visual memory representation of the stimuli and the perceptual pipeline. Thus, there would be less opportunity for post-choice integration. Consequently, participants must base their confidence on evidence that was collected by the time the decision was made. Put differently, it could be argued that BOE (the only extant single-stage theory that could account for Hurdles 1–6, which were replicated in the previous experiments), only applies in situations where, following the decision, neither the stimulus nor any representations thereof are available. Thus, by masking the stimulus we aimed to provide BOE with favorable conditions to manifest. Recall, that BOE bumps into Hurdle 7 as it predicts a negative TP effect. Thus, if participants adopt a BOE strategy for forming their confidence-judgment when the stimulus is masked, we expect to find an inverted TP effect on resolution. Alternatively, finding of a beneficial TP effect under masking conditions will provide compelling evidence for the operation of post-choice integration even with minimal (post-choice) resources. Is the beneficiary TP effect resistant to masking?

2.4.1. Results

The time pressure manipulation was effective in generating a speed accuracy tradeoff. For the group, both the accuracy rate and the mean decision time were lower for the speed vs. the accuracy condition (Accuracy: $M_{\text{speed}} = .74$, $M_{\text{accuracy}} = .85$, $z = 5.46, p < .0001$; RT: $M_{\text{speed}} = 0.53s, M_{\text{accuracy}} = 0.92s, z = 5.91, p < .0001$). Additionally, in support of Hurdle 5, the mean confidence was higher in the accuracy ($M = .89$) than in the speed condition ($M = .86; z = 3.33, p < .001$). Finally, the SAT

Table 4
Resolution of confidence effects (regression slopes) for Exp. 3.

Par	CORRECT	TP∗CORRECT
2	1.56(0.10)***	0.22(0.19)
4	1.61(0.13)***	0.81(0.24)***
6	1.00(0.08)***	0.17(0.15)
7	0.82(0.10)***	−0.06(0.19)
8	0.39(0.07)***	0.25(0.15)
11	0.54(0.11)***	0.06(0.22)
12	1.33(0.10)***	0.03(0.20)
15	0.93(0.08)***	0.60(0.17)***
16	1.30(0.08)***	0.69(0.15)***
Group	1.05(0.14)***	0.31(0.10)**

Note. Values in parentheses are standard errors. ∗, ∗∗, ∗ ∗ ∗ indicate $p < .05$, .01, .001, respectively according to *t*-test for the participants and a meta-analysis for the group.

manipulation exerted no significant effect on RT2 ($M_{\text{speed}} = 0.54s, M_{\text{accuracy}} = 0.54$ s,$z = -1.21$, $p = .23$).[6]

Turning next to the focal resolution analyses, we conducted an ordinal probit-regression analysis as in Exp. 1 but without the PA variable because here, all trials were masked (so CONF was regressed only on TP, CORRECT and their interaction). Table 4 display the resolution of confidence effects. First, in accordance with Hurdle 6, the resolution was positive for all participants and for the group (a positive main effect for CORRECT). Furthermore, resolution was positive for the group and for all participants in both the speed ($b_{\text{group}} = 1.21, z = 7.27, p < .001$) and accuracy ($b_{\text{group}} = 0.89$, $z = 6.60, p < .001$) conditions. The second data column of Table 4 confirms that in accordance with Hurdle 7, for the group and for three of the Participants, the time pressure had a positive effect on resolution.

2.4.2. Discussion of results

The results of Exp. 3 replicated all the empirical hurdles (1–7) and in particular the beneficial effect of choice-TP on resolution of confidence (that was found in Exp. 1), under a more strictly reduced availability condition featuring post-choice backward masking. If masking interferes with people's ability to perform post-choice integration, we should have found a decreased TP effect, perhaps even an opposite, harmful choice-TP effect on resolution, in the case that participants had reverted to a BOE strategy. In spite of this hypothesis, we obtained a positive beneficial TP effect on resolution. These conclusions should be qualified, however, as they are based on the inability of the BOE model to predict the TP effect (see Section 1.4.1.2).

Interestingly, the TP effect in the *vanish* condition of Exp. 1 ($b_{\text{group}} = 0.28, \text{SE} = 0.15$) was nearly identical to the TP effect in the masking condition of Exp. 3 ($b_{\text{group}} = 0.31, \text{SE} = 0.10, p = 0.89$). A possible interpretation of this finding is that the perceptual pipeline survived our masking manipulation and that this pipeline feeds the post-choice integrations process to comparable extents in both the vanish (Exp. 1) and the masking conditions (Exp. 3). These results, nevertheless, attest for the robustness of post choice integration.

Finally, the results of Exp. 3 contribute to our understanding of the time pressure effect on resolution and the role that RT2 plays in generating this effect. Note that the TP effect emerged even when RT2 was not significantly larger in the speed condition. This finding converges with the findings of Exp. 1 to the conclusion that a longer RT2s is *not* a necessary condition for a beneficial TP effect.

[6] The results for the individual participants are reported in Appendix G (Table G3). As before, prior to conducting these analyses we removed trials that were likely the result of 'contaminating' processes using the same cutoffs as in Exp. 1 and 2, which resulted in the elimination of an average 4.7% (min 2.5%; max 7.5%) of the data. Additionally, we report replications of Hurdles (3–4) in Appendix F.3.

3. Further analyses: RT2 in the spotlight

The most important conclusion from our investigation thus far is that post choice integration and its duration (RT2) are *causal* determinants of confidence and that post-choice integration is extremely persistent and resistant to interference. In addition, our analyses revealed some interesting patterns involving RT2 and its relation to other variables in the choice followed by confidence paradigm. For example, we found higher RT2 in the *remain* compared with the *vanish* condition, consistent with the hypothesis that perceptual availability provides a more durable opportunity for post-choice integration.

The upshot of these results is that RT2 conveys an abundance of information with respect to the confidence formation process. In the current section we aim to confirm existing, and explore additional relations between RT2 and other variables. Extending the empirical manifold with novel, robust empirical findings involving RT2 will guide and constrain both the interpretation of the current results and future confidence theories and modeling attempts, thus extending our understanding of the confidence-forming mechanism. First, we present additional correlations of RT2 with the variables of the choice followed by confidence paradigm. Second, we present novel interaction effects on confidence and RT2. The robustness of these effects secures RT2's position as a prime dependent variable that theories of confidence should be able to account for.

3.1. Extending the empirical manifold

3.1.1. RT2 correlations

Pleskac and Busemeyer (2010) found that RT2 correlated with the choice-variables, negatively with accuracy and positively with RT. Additionally, RT2 correlated negatively with the discriminability level of the stimulus. Finally, within SAT blocks RT2 correlated negatively with confidence. These 'RT2 correlations' are summarized in Hurdle 8 (Table 1). We examined whether these correlations were replicated in our Exp. 1 and 3. First, in both experiments, for all individuals as well as for the group, RT2 was significantly shorter for correct than for error decisions (Exp1, Group: $M_{\text{correct}} = 0.62s$, $M_{\text{error}} = 0.83s, z = -4.56, p < .001$; Exp3, Group: $M_{\text{correct}} = 0.51s, M_{\text{error}} = 0.62s,\ z = -5.03, p < .001$; Results for the individuals are given in Table D1). Second, the RT2-discriminability Gamma correlation (Goodman & Kruskal, 1954; henceforth denoted Γ) was significantly negative for the group and for four (out of six) and eight (out of nine) of the individuals respectively, in Exp. 1 (Group: $\Gamma = -.1, z = -2.71, p < .01$) and Exp. 3 (Group: $\Gamma = -.12, z = -4.47, p < .001$; for participants are see Table D2). Third, Pearson's correlation between RT and RT2 was significantly positive for the group in both Exp. 1 ($\text{r} = .13, z = 3.86, p < .001$) and Exp. 3 ($\text{r} = .14, z = 2.64, p < .01$). This was also the case for all the participants in Exp. 1 and for six (out of 9) participants in Exp. 3 (see Table D2). Fourth, within SAT blocks, RT2 correlated negatively with confidence for all participants and for the group in both Exp. 1 (Group: $\Gamma = -.43, z = -6.00, p < .001$) and Exp. 3 (Group: $\Gamma = -.53, z = -7.25, p < .001$; for individual participants see Table D2).[7,8]

3.1.2. Novel interaction effects on RT2 and confidence

The finding of a causal effect of RT2 on confidence provided ample motivation for exploring the empirical manifold around confidence and RT2. We found novel interaction effects between difficulty and choice-accuracy on RT2 and confidence. These findings, which are summarized in the bottom part of Table 1, will pose additional stringent constraints on confidence-theories. To verify their robustness, we also re-analyzed the line length task of Pleskac and Busemeyer (2010).

The top and bottom panels of Fig. 2 display the mean confidence and mean RT2 respectively as functions of the discriminability level and choice accuracy. The left, middle and right panels correspond to our Exp. 1 and 3 and to the line-length task respectively. The top panels show that confidence

[7] Note that, in Exp. 1, RT2 was larger in the speed than the accuracy condition (Section 3.1). This finding however, failed to reach significance in Experiment 3, so we did not include it as part of the Hurdle list (Table 1).

[8] In Appendix H, we present RT2 results from a practice task and discuss the possibility that the findings reported in Section 3, result from difference in motor production times across confidence responses.

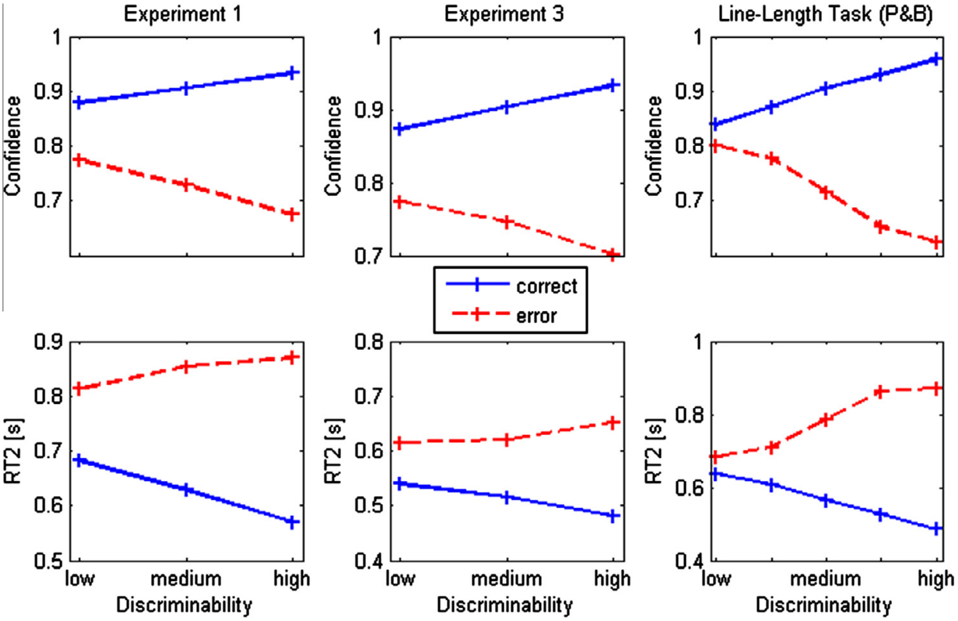

Fig. 2. The interactions between choice correctness and discriminability in Exp. 1 and 3 and in Pleskac and Busemeyer's line length task. The top panels present the interactive effect on confidence and the bottom panels—the interactive effects on RT2. The results are averaged across participants. Exp. 2 (not displayed), wherein RT2 was manipulated, yielded a similar interaction pattern for confidence.

in correct choices increases as a function of discriminability whereas, confidence in error responses decreases as a function of discriminability (Hurdle 9). Consequently, resolution improves for higher discriminability. The RT2 data revealed a 'mirror pattern' where correct RT2s speeded up with increasing discriminability whereas error RT2s-slowed down (Hurdle 10).

To evaluate statistically the interaction effect on confidence we conducted a multiple probit-ordinal regression. 'CONF' and 'CORRECT' were defined as in prior analyses and 'DISC' was a discriminability level variable, with higher values corresponding to higher discriminability level (1–3 in our Exp. 1–3 and 1–5 in the line length task). CONF was regressed on CORRECT, DISC and their interaction, which is of focal interest. We also examined the 'simple discriminability effects' for correct and error trials separately.[9] At the group level, the DISC$*$CORRECT interaction effect was positive for all experiments (Exp. 1: $b_{\text{group}} = 0.59, z = 5.59, p < .001$, Exp. 2: $b_{\text{group}} = 0.65, z = 10.11, p < .001$, Exp. 3: $b_{\text{group}} = 0.51$, $z = 4.78, p < .001$, line-length: $b_{\text{group}} = 0.62, z = 15.56, p < .001$). The simple effects of discriminability were positive for correct trials (Exp. 1: $b_{\text{group}} = 0.27, z = 4.82, p < .001$, Exp. 2: $b_{\text{group}} = 0.35, z = 9.69$, $p < .001$, Exp. 3: $b_{\text{group}} = 0.30, z = 6.64, p < .01$, line length: $b_{\text{group}} = 0.30, z = 13.06, p < .001$) but negative for error choices (Exp. 1: $b_{\text{group}} = -0.32, z = -4.57, p < .001$, Exp. 2: $b_{\text{group}} = -0.30$, $z = -7.01, p < .001$ Exp. 3: $b_{\text{group}} = -0.21, z = -2.96, p < .01$, line length: $b_{\text{group}} = -0.33, z = -10.07$, $p < .001$). The individual-participant results are reported in Table D3.

To probe the interaction effect on RT2 we conducted a multiple linear regression for RT2 on CORRECT and DISC (Exp. 2, wherein RT2 was manipulated was excluded from this analysis). For all three data sets the interaction was negative (Exp. 1: $b_{\text{group}} = -0.07, z = -3.15, p < .01$, Exp. 3: $b_{\text{group}} = -0.04, z = -3.31 p < .001$, line-length: $b_{\text{group}} = -0.09, z = -5.33, p < .001$). Considering the simple discriminability effects, for correct trials, the effect was negative in all datasets (Exp. 1: $b_{\text{group}} = -0.05, z = -3.20, p < .01$, Exp. 3: $b_{\text{group}} = -0.03, z = -4.70 p < .001$, line-length:

[9] This was achieved by repeating the regression but replacing CORRECT with CORRECT′ = CORRECT + .5 (for errors) or CORRECT′ = CORRECT − .5 (for corrects) and probing the effect of DISC.

$b_{\text{group}} = -0.04, z = -3.66, p < .001$). For error trials, in all three data sets the trend was positive, but it reached significance only for the line-length task (Exp. 1: $b_{\text{group}} = 0.02, z = 1.27, p = .20$, Exp. 3: $b_{\text{group}} = 0.01, z = 1.36 p = .17$, line-length: $b_{\text{group}} = 0.05, z = 3.17, p < .01$). The individual-participant results are reported in Table D4.

The upshot of these results is that in addition to causally influencing confidence, RT2 correlates with other variables (both dependent, such as choice-correctness and independent, such as stimulus-discriminability) in the choice-followed by confidence paradigm and it is subject to interaction influences. These findings converge on the conclusion that theories of confidence should address the immanent role that RT2 serves in shaping confidence. In the following section, we describe two alternative ways to incorporate RT2 into models of confidence, either as an exogenous or as an endogenous variable. These approaches are illustrated with respect to the 2DSD model variants. Then we continue to present a novel endogenous model of RT2.

3.2. The exogenous versus endogenous status of RT2

An *endogenous* theory for RT2 is a theory that predicts RT2 alongside choice accuracy, RT and confidence. Unlike endogenous models, in *exogenous* models RT2 is external to the model in the sense that the model 'observes' RT2 i.e., reads it from the data, rather than predicts it. This means that in order to generate predictions pertaining to confidence judgments, the empirical RT2 needs to be measured and used to inform the model (the model's predictions are a function of RT2). Alternatively, the model can make a priori predictions but only for effects that control for RT2. The two variants of 2DSD demonstrate these different approaches.

The 2DSD interrogation model treats inter-judgment time as an *exogenous* parameter. A potential limitation of such an approach is that it restricts the ability of 2DSD to make *a priori* predictions about confidence prior to measuring inter-judgment times, *even under the idealistic assumption that all the 2DSD parameters (for a given participant) are perfectly known*. Consider for example Hurdle 3, the positive correlation between confidence and stimulus discriminability. On first thought, it may seem that 2DSD predicts this relationship unequivocally. In free-RT tasks, a higher drift rate will result in an increase in the amount of evidence that is accrued during the post-decision integration stage, and hence, with all other thing being equal, confidence will increase by the end of this stage. However, not all other things are equal. Indeed, according to Hurdle 8, stimulus discriminability correlates negatively with inter-judgment time. Thus, on the one hand, easier stimuli benefit from higher drift rates in the second processing stage, which tend to yield higher levels of total evidence and hence confidence. But on the other hand, the second inter-judgment stage tends to be shorter for easier stimuli, thus decreasing the amount of total evidence. Depending on the more dominant factor, the combined effect on confidence may ultimately result in either lower or higher confidence judgments with increased stimulus discriminability. Without a theory, which specifies how RT2 is determined, this pattern can only be assessed ex-post. In a similar vein, RT2 is correlated with accuracy and with decision time and these correlations may affect the model's predictions pertaining to confidence for correct vs. errors (Hurdle 6) and to fast vs. slow choices (Hurdle 4). In summary, for exogenous models of RT2 one cannot predict confidence without considering these factors.[10] Without auxiliary assumption about RT2, the model can only make a priori predictions when RT2 is controlled for.

Thus, theories of confidence could benefit substantially by accounting endogenously for RT2. Accounting for RT2 (in addition to confidence) would allow us to test whether a theory can predict empirical effects that independent variables exert on confidence, while simultaneously capturing the variance in RT2. Furthermore, given a set of cognitive parameters, a model of RT2 can generate a priori rather than tentative predictions with respect to confidence. The optional stopping rule 2DSD is such a model (see Section 1.4.2.3). Remarkably, this model accounted for the negative relationship between inter-judgment time and confidence, which was problematic for the interrogation 2DSD (see Pleskac & Busemeyer, 2010 for further details). Thus, the 2DSD is a promising model for

[10] Notably, in fitting the interrogation 2DSD, Pleskac and Busemeyer (2010) used the mean empirical inter-judgment time in each of the SAT conditions. Thus, in the fits RT2 did not vary as a function of discriminability and choice correctness. It is thus an open question how well the model can account for the Hurdles when RT2 varies as a function of these variables as it did in the data.

all four dependent variables in the choice-followed by confidence paradigm. However, some questions, pertaining to the models' ability to account simultaneously for the entire empirical manifold, remain open for future research.[11]

4. The collapsing confidence boundary model

Our next goal was to propose a novel endogenous dual stage model for RT2, which is motivated by the various aspect of the empirical manifold. We stress that it is possible that alternative models from the extant literature may also able to account for the entire empirical manifold (see discussion in Section 5.5). Nonetheless, we decided to explore here a novel model since we believe that it provides new important insights into the nature of the confidence-generating mechanism, that it integrates confidence theories with modern models of decision making (which rely on temporally dynamic choice thresholds) and finally, that it achieves a satisfactory tradeoff in providing a unifying and yet a relatively straightforward account for the vast range of empirical hurdles. We begin with a detailed description of the model, which we dub the Collapsing Confidence Boundary (henceforth CCB) model and follow with a demonstration of how it predicts the empirical manifold.

4.1. A description of CCB

The first decision stage in CCB is identical to 2DSD: A standard diffusion model. During the second stage of processing, the diffuser, whose state is denoted $L(t)$, continues to integrate evidence, with drift rate and diffusion-noise (diffusion coefficient) identical to the first stage. At the onset of the second stage a *single* confidence boundary is placed beyond the recently crossed choice threshold. If in the first stage the upper choice boundary θ was reached then the confidence threshold is placed *above* the choice threshold whereas, if the lower choice boundary $-\theta$ was reached-the confidence threshold is placed *below* the choice threshold. Below, we refer to these events as the upper and lower choice events respectively.

The core property of the model lies in the assumption that with the passage of (the post-choice integration) time, the choice threshold collapses towards decreasing levels of choice-supportive evidence. In other words, the confidence boundary moves down or up, respectively, in the upper and lower choice events. With each collapse, the confidence boundary represents a decreasing level of confidence. Thus, at the onset of the second stage, the confidence boundary represents a confidence level of 1.00 which decreases to 0.90 after the first collapse, 0.80 after the second collapse and so on (following the fifth collapse—the boundary corresponds to confidence 0.50). The time between consecutive collapses is distributed according to a uniform distribution $U([0,\tau_j])$, where j indexes confidence levels (1.00,.90,.80,…,.60). The τ_j's are dubbed the *timer* parameter (for example, $\tau_{0.7}$ is the maximal duration that the confidence boundary 'spends' at the .70 confidence level, until it collapses to confidence .60). The *absolute* heights of the confidence threshold are $h_j, j = 1.00, 0.90, \ldots, 0.60$ where h_j decreases monotonically. Note, that in our notation these heights are measured relative to $z = 0$, the non-biased starting point of the first diffusion stage and not relative to the choice threshold (for example $h_{.90}$ is the absolute height of the confidence boundary when it corresponds to confidence .90). The mean collapse times and the heights are free model parameters. Importantly, the collapse durations are stochastically independent of each other and of the diffuser dynamics until the completion of the second stage.

In CCB, a confidence response j is given for the shortest time RT2 such that either (a) the diffuser and the confidence boundary *crossover,* i.e. $L(\mathrm{RT2}) \geqslant h_j$ (or $L(\mathrm{RT2}) \leqslant -h_j$) for the upper (lower) event and h_j is the active height of the confidence bound at time RT2; or (b) the choice boundary collapsed to

[11] First, in fitting the model the across trial variability parameters were not used, either in starting point or in drift rate. But without these parameters a diffusion model cannot account for the systematic differences between correct and error choice RTs (Hurdle 2). Furthermore, without drift rate variability and for a constant difficulty level, the second stage dependent variables (RT2 and confidence) depend only on the decision and on the second stage parameters *but not on RT*. Thus, the model can account neither for the negative RT-confidence correlations (Hurdle 5) nor for a positive RT–RT2 correlation (Hurdle 10). Second, data was collapsed into a single difficulty level, a procedure that forfeits the ability to assess whether the model accounts for difficulty related effects (Hurdles 3, 9, and 10).

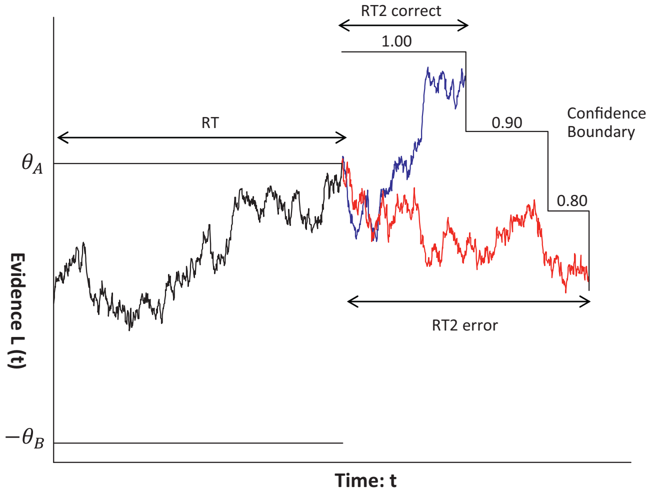

Fig. 3. Illustrative trial realizations for the CCB model. The black trace plots the first stage of a trial leading to choice of alternative A. Once, the choice is made a collapsing confidence boundary is set. The blue trace depicts a typical correct trial and the red line depicts a typical error trial (compare with Fig. 1). While the blue trace 'converges' towards the collapsing boundary, the red trace 'escapes' from the boundary. The blue trace intersects the confidence boundary as it collapses from confidence level 1.00 and hence the confidence level is .90. Similarly, the red trial terminates with confidence .70 (as the threshold collapses from confidence .80). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the .50 confidence level (in this case a confidence judgment $j = .50$ is issued immediately even if there is no cross over with the diffuser). The rationale for (b) is that .50 is the minimal eligible confidence rating, so confidence cannot diminish further. The CCB model is illustrated in Fig. 3.

According to the CCB model the confidence judgment process reflects a tradeoff between two conflicting goals of the judge: On the one hand, a high level of confidence is desired. Nevertheless, high levels of confidence mandate sufficiently high amounts of evidential support, whose accrual is time consuming. The CCB model assumes that integration time incurs costs, which may be either associated directly with the passage of time or with 'effort exertion'. The collapsing confidence boundary implements the tradeoff between the conflicting desires for high confidence and low costs by assuming that initially, participants set their threshold at a confidence level of 1.00. As time unfolds, if this target-confidence rating lacks sufficient support, the judge is willing to compromise his or her confidence, effectively reducing the second stage duration (cf. Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012 for a similar model but with collapsing *decision* boundaries).

The optional stopping 2DSD and the CCB are similar in that both models determine confidence by the final state of the diffuser in the evidence space. The fundamental differences between these models are twofold: First, whereas in 2DSD multiple 'confidence markers' are active throughout a trial, in CCB only a single height (the current state of the confidence boundary) is active at any given moment. Furthermore, whereas in 2DSD reaching an intermediate confidence marker is a stochastic termination event (integration ends with some exit probability), in CCB the confidence boundary crossover deterministically terminates the trial. In our opinion, these differences and especially the fact that crossing the threshold unavoidably terminates the trial, render the mechanism of the CCB more straightforward to understand and allow it to make clear qualitative predictions with respect to the empirical manifold, as we next show.

4.2. A combined model fitting and simulation study of a simplified CCB model

We performed a combined model fitting-simulation study to test whether the CCB model can account for the vast range of patterns in the empirical manifold. The empirical data that we used in this study was taken from the line length task of Pleskac and Busemeyer (2010). Both this task and our Experiments (1 and 3) agree with respect to all the qualitative patterns in the empirical manifold. Given this unanimity, we preferred to use the line-length task because it contains more than twice the number of trials relative to our experiments and so it yields more reliable estimates of the dependent variables.

The above presentation of CCB was highly general in that both the boundary heights and the timer parameters are free to vary with j. This yields a set of ten free parameters that specify the collapse dynamics. Model simplicity was of paramount concern for us because the simpler a model, the easier it is to understand both its operation and predictions. Thus, below we used a simplified model version, which has only three free collapse parameters. This version is obtained by assuming that the first collapse has its own timer parameters $\tau_{1.00}$ but that with each consecutive collapse, the timer parameters shrinks by a factor of two, i.e. $\tau_{1.00} = 2\tau_{.90} = 4\tau_{.80} = \cdots = 16\tau_{.60}$. This speed up between consecutive timer parameters was motivated by the observation that when confidence declines, the variance of the RT2 distribution, conditional on the confidence level, increases but in a decelerating rate. A further simplifying assumption which allowed us to reduce the number of free model parameters is that the 1.00 confidence boundary is placed at a free height $h_{1.00}$ but that with each collapse it falls by the same amount, denoted $\varDelta$ i.e., $h_{1.00} - h_{.90} = h_{.90} - h_{.80} = \cdots = h_{.70} - h_{.60} \equiv \varDelta$. Strikingly, even this highly constrained model, can predict all the qualitative patterns in the empirical manifold.

Another simplification we adopted was to assume that the second stage parameters are not influenced by the SAT manipulation. Finally we assumed that the residual time for the confidence response is a constant that is identical to the motor production time for the choice response. Notably, if evidence-integration continues while the choice-response is being produced, then $\mathrm{RT2} = t_{\mathrm{conf}} - t_m + T2_{\mathrm{ER}}$, where t_{conf} denotes the post-choice integration duration as measured from the time the choice was made (but still not produced), t_m is the choice motor production time and $T2_{\mathrm{ER}}$ is the residual time for the confidence judgment. Hence, assuming that $t_m = T2_{\mathrm{ER}}$ implies that RT2 is equal to the time spent on post-choice integration.

In our combined simulation-fitting study we used a strategy of ‘parameter separation’. Our procedure consisted of two stages: In the first *fitting* stage we fit the drift diffusion model to the group choice and RT data,[12] ignoring altogether the confidence and RT2 data. In the parametric design of our diffusion model, the choice threshold $\theta_{\mathrm{speed}}, \theta_{\mathrm{accuracy}}$ was selectively influenced by the SAT condition and the mean drift rate $v_{\mathrm{easy}}, v_{\mathrm{hard}}$ was selectively influenced by the difficulty level (but see Rae et al., 2014). In total, our diffusion model consisted of eight free parameters (see Table 5). Importantly, following the first stage the diffusion parameters were maintained at a fixed level and were not adjusted during the second simulation stage.

In the second stage we augmented the diffusion parameter set with the three ‘stage II’ parameters listed at the bottom of Table 5. These parameters were configured so that the model’s predictions with respect to response proportions and mean RT2’s would be similar to the empirical observations (see Figs. 4 and 5). We then simulated the CCB model and calculated predictions with respect to RT2 and confidence. Notably, in CCB the effects of the first stage diffusion parameters are not limited to choice and RT but also manifest in the second stage performance measures. Consequentially, the predictions with respect to confidence and RT2 can improve if the diffusion parameters are adjustable during the second stage. In allowing the diffusion parameters to be affected only by the choice data, our method is similar in spirit to the generalization criterion for model evaluation (Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Busemeyer & Wang, 2000). While our fitting procedure is highly suboptimal it was tailored to our main goal in the current study, which was not to recover the model parameters with the highest possible reliability but rather, to show that the relatively simple CCB model, can account for the vast empirical manifold in a parametric range that produces behavioral predictions, which are similar to those observed in the data.

[12] We pooled the three highest and the three lowest difficulty levels, respectively, into 2 compound difficulty levels: *hard* and *easy*. Next, for each combination of trials defined by choice correctness (correct, error) $*$ choice-stress (speed, accuracy) $*$ difficulty (hard, easy) $*$ participant we calculated the RT quantiles (.1,.3,.5,.7,.9). Finally, we averaged these quantiles across participant to obtain group data specified at the level of choice correctness $*$ stress $*$ difficulty. The group data fit was conducted using the DMAT toolbox (Vandekerckhove & Tuerlinckx, 2007, 2008).

Table 5
Parameters of the CCB simulation.

Parameter	Meaning	Value
First stage diffusion parameters		
$\theta_{\text{speed}}, \theta_{\text{accuracy}}$	Choice thresholds for the SAT conditions	0.045, 0.106
$v_{\text{easy}}, v_{\text{hard}}$	Mean drift rate for the easy and hard trials	0.110, 0.309
η	Standard deviation of drift rates across trials	0.156
σ	Diffusion coefficient	0.1*
z	Starting point	0*
s_z	Range of starting point	0.018
T_{ER}	Residual non decision time	0.377
s_T	Range of residual non decision time	0.127
Second-stage parameters		
h	Absolute height of the 1.00 confidence boundary	0.18
Δ	Collapse height	0.041
$\tau_{1.00}$	First collapse timer parameter	1.203

Note. The * indicates parameters that were fixed at a given level, rather than fit, during the first fitting stage.

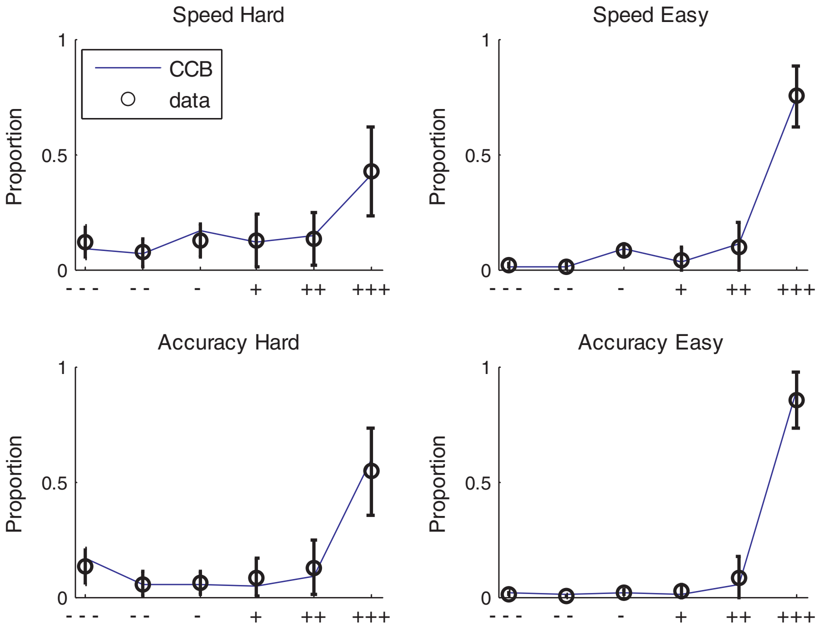

Fig. 4. Empirical and CCB predicted response-proportions for the different choice accuracy and confidence combinations. The empirical, group level results were obtained by averaging the corresponding measures across participants. The top and bottom panels correspond to the speed and accuracy SAT manipulation respectively. The left and right panels correspond to hard and easy trials respectively. Within each panel proportions, which sum to 1, are presented for correct (+) and for error choices (−). One, two and three symbols designate respectively low, medium and high confidence obtained by pooling together the .50 and .60 (low), .70 and .80 (medium) and .90 and 1.00 (high) confidence judgments. Empirical data are depicted in black 'o's and CCB predicted proportions are depicted in a blue line. The error bars correspond to 95% confidence intervals for the population measures.

4.2.1. Predictions for RT2 and confidence

The results of our study are presented in the following set of figures: Fig. 4 displays the empirical and predicted proportions of the different choice-accuracy and confidence combinations, in the various experimental conditions. The proportions predicted by CCB (in blue) track the data (black) quite

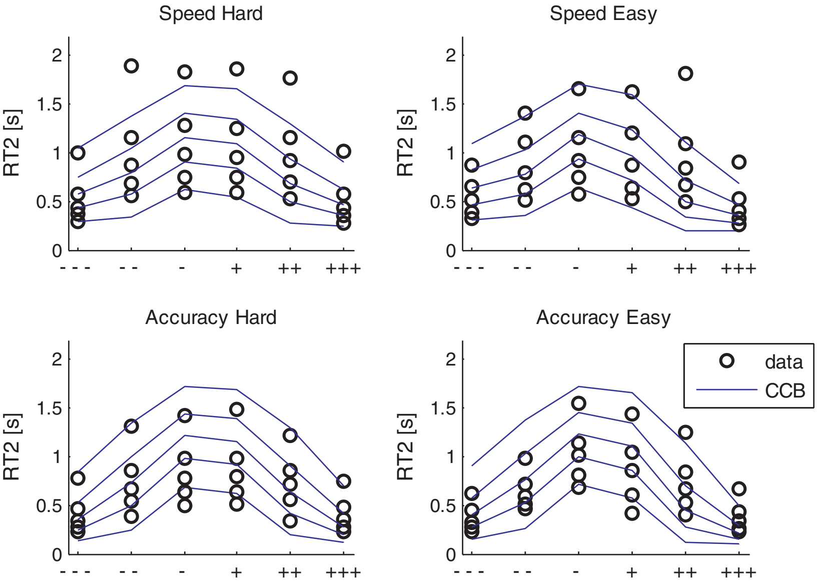

Fig. 5. Empirical and predicted RT2 quantiles. The arrangement of the figure is similar to Fig. 4.

nicely. Fig. 5 displays the RT2 empirical and predicted (10%, 30%, 50%, 70% and 90%) quantiles. While the shape of the RT2 distributions roughly follows the shape of the empirical distributions these results are admittedly ‘far from perfect’ as there are some substantial deviations between the empirical data and the predictions of CCB. In evaluating the CCB predictions, consider again the fact that we used a constrained and suboptimal procedure and a highly simplified model. Viewed from this perspective, our overall assessment is that the CCB can produce reasonable predictions for confidence and RT2. Fig. 5 also shows that CCB predicts the negative correlation between RT2 and confidence (Hurdle 8). Notably, the CCB-predicted Γ correlations between RT2 and confidence for the different SAT conditions were almost identical to the empirical correlations (CCB: $\Gamma_{\text{speed}} = -.51, \Gamma_{\text{accuracy}} = -.44$, Data: $\Gamma_{\text{speed}} = -.52$, $\Gamma_{\text{accuracy}} = -.47$).

4.2.2. CCB vis. a vis. the empirical manifold

We next describe how the CCB predicts the empirical Manifold (Table 1). To facilitate the presentation, we divided the empirical manifold into subsets of empirical patterns that are grouped around a common motif.

4.2.2.1. Hurdles involving discriminability. In CCB, the higher the drift rate (corresponding to increased levels of discriminability), the earlier the diffuser converges towards and finally crosses the confidence boundary, hence predicting a negative RT2-discriminability correlation (Hurdle 8). Furthermore, earlier crossovers entail reduced opportunity for the confidence boundary to collapse—yielding higher confidence (Hurdle 3). However, these predictions are restricted to correct trials (that comprise a majority of the trials) for which the diffuser generally converges towards the confidence boundary. For erroneous trials, the opposite pattern is predicted i.e. higher drifts result in slower and lower confidence crossovers, because for errors, the diffuser moves away from the boundary (see red trace in Fig. 3). This difference between correct and erroneous responses underlies the interaction hurdles (9 and 10; see Fig. 2). Consider a negative drift trial and an erroneous ‘upper choice’ event (i.e. the high choice boundary was reached). During the second stage the confidence boundary will collapse downwards and the diffuser will trend downwards too. The situation resembles a chase: The confidence

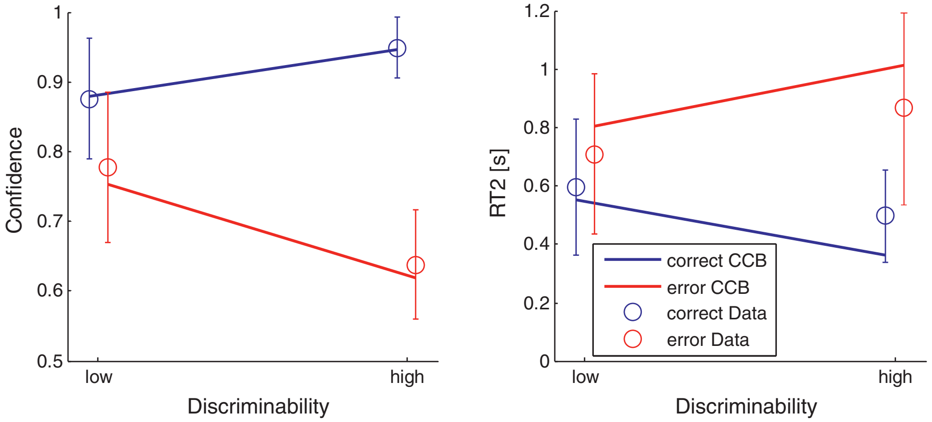

Fig. 6. Mean confidence (left panel) and RT2 (right panel) for the various difficulty levels and for correct and erroneous choices. Empirical group data are depicted with the 'o' symbols and CCB predicted proportions—with solid lines. Correct and Erroneous responses are depicted in blue and red, respectively. The error bars correspond to 95% confidence intervals for the population measures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

boundary tries to 'catch' an 'escaping' diffuser. The higher the drift rate, the faster the diffuser escapes. Thus, higher drift rates for errors predictably result in both a larger RT2 and a lower confidence. These predictions are illustrated in Fig. 6. The upshot is that CCB predicts the novel stimulus discriminability and choice interaction effects on confidence and RT2. Note also, that a consequence of the confidence interaction (left panel) is that resolution increases with stimulus discriminability (the difference in mean confidence between correct and erroneous choices is larger for high than low discriminability trials).

4.2.2.2. Relationships involving RT. Consider first, the negative RT-confidence correlation, which is found within experimental SAT conditions (Hurdle 4). Two common principles underpin the account of CCB for this effect: the drift rate variability and the continuity of the drift between the first and second integration stage. The variability in drift rate across trials results in two important consequences: First, trials with higher drifts generally yield lower RTs as the choice threshold is reached faster. Second, the higher drift carries over to the second integration stage and produces both a higher and faster confidence rating. When trials of various discriminability levels are intermixed, this negative correlation is further exacerbated. Importantly, these considerations also demonstrate that CCB predicts a positive RT–RT2 correlation (Hurdle 8).

Next, we turn to the positive RT-confidence relationship across SAT conditions (Hurdle 5). This hurdle is accounted for by the fact that, in the accuracy (relative to the speed) condition, the diffuser has to traverse a shorter distance to cross the confidence boundary. Indeed, in the accuracy vs. speed condition the first stage terminates at a higher choice-threshold. Additionally, the heights of the collapsing boundary (and the collapsing dynamics) are invariant across SAT conditions. Thus, at the onset of the second stage the diffuser and the confidence boundary are closer to each other in the accuracy condition. Consequentially the confidence crossover occurs both sooner and at a higher confidence level. These considerations account for both Hurdle 5 and for the higher RT2 in speed relative to accuracy SAT conditions as illustrated in Fig. 7.

The higher RT2 in the speeded SAT condition warrants some qualifications. In the line length task (Pleskac & Busemeyer, 2010), this pattern was found for all participants and for the group. However, in our Exp. 1 and 3 we sometimes found a significantly opposite pattern: lower RT2 in the speed condition, for four participants (Participant 3 in Exp. 1 and Participant 2, 7 and 9 in Exp. 3). Furthermore, recall from the discussion of Exp. 1, that for Participant 3 the lower RT2 in the accuracy condition was paired with a beneficial TP effect on resolution. The current

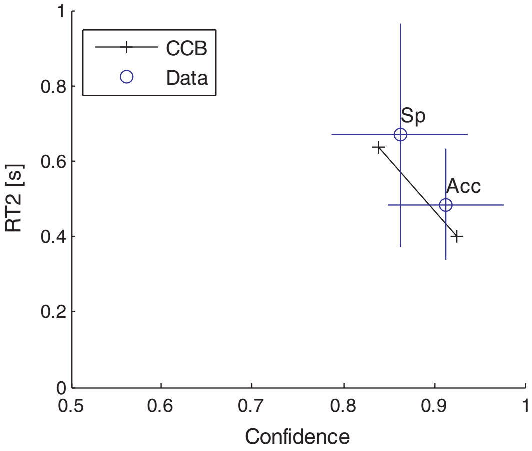

Fig. 7. Mean confidence (abscissa) vs. the mean RT2 (ordinate) for the different SAT conditions. Empirical data are presented in blue 'o's, and CCB model predictions in black '+'s. The higher asterisk and 'o' correspond to the SAT speed conditions, and the lower symbols correspond to the accuracy SAT condition. The vertical (RT2) and horizontal (confidence) error bars correspond to 95% confidence intervals for the population measures (reflecting variability across participants), calculated separately for each measure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

simplified CCB model version, in which the collapsing boundary parameters are uninfluenced by the SAT manipulation, cannot predict a lower RT2 in the accuracy condition. Nonetheless, in follow-up model simulations, we found that a more flexible CCB variant which allows the h-parameter—the initial height of the confidence boundary—to vary with the SAT manipulation (with lower values for the speeded condition, $h_{\text{speed}} < h_{\text{accuracy}}$; the rest of the parameters, the collapse height, Δ and the collapse-rates were still maintained at a fixed level across SAT conditions), can predict a lower RT2 in the accuracy condition combined with a positive TP effect (as for Participant 3 in Exp. 1). When RT2 is lower in the speed than in the accuracy SAT condition, the model can still generate a higher resolution due to drift rate variability particularly, the different distribution of second-stage drift rates, conditional on choice correctness, imposed by the different SAT conditions (see Appendix B). In conclusion, modeling the more intricate individual data patterns requires a more complex CCB variant. Such extensions are beyond the scope of the current paper, in which we focus on the more typical and robust patterns.

4.2.2.3. Resolution and the TP effects. How does the CCB account for the positive resolution of the confidence judgments (Hurdle 6)? As in 2DSD, the second stage information tends to be congruent with correct decision and incongruent with erroneous decisions. Consider for illustration the second stage following an 'upper choice' event (i.e. the positive choice threshold was reached; see also Fig. 3): For both correct and errors the confidence boundary collapses downwards. However, whereas for correct the diffuser typically traverses upwards, for errors it trends downwards. Consequentially, the average convergence rate (or the 'relative velocity' with which the diffuser and the confidence boundary move towards each other denoted $\tilde{v}_{\text{correct}}, \tilde{v}_{\text{err}}$) of the diffuser relative to the confidence boundary is higher for correct than for error choices: $\tilde{v}_{correct} > \tilde{v}_{\text{err}}$ Thus for corrects the crossover occurs sooner (Hurdle 8), and confidence is higher (Hurdle 6).

Note that strictly speaking, in CCB the confidence boundary does not move continuously but rather in 'collapsing bursts'. Nonetheless, to gain intuition with respect to the operation of the model with respect to the beneficiary TP effect (on resolution) it might help to imagine a continuous rather than a discrete 'velocity'. CCB predicts the beneficiary TP effect because the extra distance between the

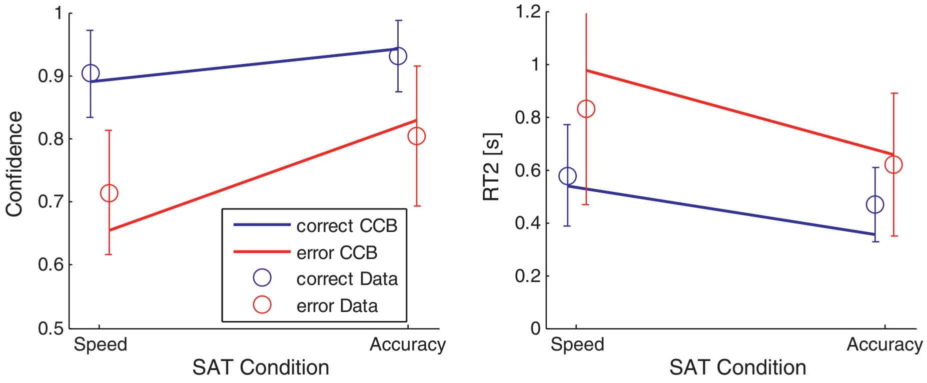

Fig. 8. Mean confidence responses (left panel) and RT2 (right panel) for the various SAT conditions and for correct and erroneous choices. Empirical data are depicted with the 'o' symbols and predicted proportions—with solid lines. Correct and Erroneous responses are depicted in blue and red respectively. The error bars correspond to 95% confidence intervals for the population measures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

diffuser and the confidence boundary (in the speed relative to the accuracy condition) is traversed in a shorter time by the higher correct (relative to error) velocity.[13] These predictions are confirmed by Fig. 8.[14]

4.2.2.4. Summary. In the current section, we presented a 'proof of concept' for the notion that dual stage theories can provide a unifying account for the many intricate aspects of the empirical manifold. Admittedly, our presentation of the CCB model was somewhat preliminary as it leaves many important questions open for futures research (see Section 5.4). Nonetheless, this presentation served our purpose, to demonstrate a (relatively) simple dual-stage model, which integrates in a coherent manner our understanding of the choice followed by confidence paradigm and which provides a plausible mechanistic theory for existing and novel empirical phenomena.

5. General discussion

The goal of the current paper was threefold: First, we presented direct empirical evidence in support of the hypothesis that in the choice followed by confidence paradigm, participants continue to accrue perceptual evidence after the decision is made and that the duration of the post-choice evidence-accrual stage causally affects their confidence judgments. The resolution of confidence and the beneficial choice-TP effect increased as a function of perceptual availability (Exp. 1). Additionally, perceptual availability increased both the resolution of confidence and the harmful—confidence TP effect (Exp. 2). Testing the boundary conditions of this phenomenon, post-choice

[13] To elaborate, define by, $\Delta_\theta = \theta_{\text{accuracy}} - \theta_{\text{speed}}$ the difference between the choice threshold in the accuracy and the speed SAT conditions. Comparing these SAT conditions, in the speed condition an additional gap of $\Delta\theta$ between the diffuser and the confidence boundary should be bridged, for a crossover to occur. The larger the boundary-diffuser convergence velocity the sooner the gap will be bridged (this time is approximated by $\frac{\Delta\theta}{\tilde{v}}$) and hence the lower the effect of the extra gap on confidence (when the gap is bridged faster the confidence boundary will collapse to a lower extent). Putting the pieces together, since the convergence velocity for corrects is larger than for errors $\tilde{v}_{\text{correct}} > \tilde{v}_{\text{err}}$, the additional $\Delta\theta$ gap in the speed condition is traversed faster for corrects $\left(\frac{\Delta\theta}{\tilde{v}_{\text{correct}}} < \frac{\Delta\theta}{\tilde{v}_{\text{err}}}\right)$. These correct vs. errors differences in RT2 differences entail a corresponding difference with respect to confidence, because faster convergences provide fewer opportunities for collapse. Thus, while confidence for corrects decreases, confidence for errors drops down to a larger extent. A beneficiary TP-resolution effect ensues (Hurdle 7).

[14] The TP and choice-correctness interaction effect on RT2, which is displayed in the right panel of Fig. 8, was significant in the line length task. We found a similar trend in our Exp. 1 but it failed to reach significance. Thus, we did not include this pattern in the empirical manifold.

integration proved surprisingly resistant to our attempt to hinder it by backward masking of the stimuli immediately after the choice (Exp. 3).

Second, we showed that the inter-judgment time (RT2) is correlated with the other variables in the choice followed by confidence paradigm. We concluded that accounting for RT2 is an important challenge for theories, which aim to explain the confidence-generation mechanism. Third, we addressed this challenge by presenting a novel theory of confidence and RT2, the CCB model. This theory, which was successful in accounting for the entire empirical manifold, provides the novel insight that confidence is determined by collapsing boundary dynamics. Below, we discuss the implications of our findings.

5.1. Single vs. dual information-collection stages theories

The Balance of Evidence (BOE) (Vickers, 1979) is arguably the most established among the family of the single-stage theories of decision confidence. Remarkably, this model is able to account for many of the empirical regularities in the decision followed by confidence paradigm (Hurdles 1–6). Despite the overall support we obtained for dual-stage models, we believe that it is premature to dismiss the BOE account. First, our results indicate that individual difference in the confidence mechanism may be at play (e.g., see our discussion of Participant 1 in Exp. 1, Section 2.2.2). Future studies are needed for assessing the possibility that BOE is a viable strategy for confidence resolution on which a portion of participants may rely. Furthermore, single stage models may be able to account for findings in other perceptual paradigms, wherein the stimulus is presented very briefly (say for 100 ms) and then masked. Indeed, in such cases pipeline and iconic memory influences may be exhausted during the first information-integration stage, and so nothing would remain to feed a second information-collection stage.

Finally, our conclusions in favor of dual-stage models accounts are so far limited to speeded perceptual-decision tasks. One speculation is that in other domains of choice, such as general knowledge questions (e.g., when participants deliberate for 30 s vs. 1 min on the question of which of two authors wrote ‘Moby Dick’), the decision stage is not mediated by integration to boundary but, rather, people terminate the decision when they exhaust the information that can be retrieved from memory (or when novel mnemonic information deteriorates in its quality). In such cases, confidence may be driven by either the balance or by the consistency (SCM; Koriat, 2012) of the evidence supporting each alternative.

5.2. Error monitoring, changes of mind and the extent of post-choice integration

Additional converging evidence for post-choice integration comes from the field of error monitoring. Error monitoring is the meta-cognitive process by which observers can detect and correct their own errors, once a choice has been made, even in the absence of explicit feedback (Rabbitt, 1966). Empirical studies suggest that this ability relies on post-choice integration of additional information, which stands against the initial erroneous choice (e.g. Jentzsch & Dudschig, 2009; Rabbitt, 2002; Rabbitt & Vyas, 1981; see Yeung & Summerfield, 2012 for a review).

Resulaj et al. (2009) provided further evidence for post-choice integration in a ‘changes of mind’ paradigm: Observes indicated the direction of a random dot motion stimulus by moving a joystick leftwards or rightwards. On some portion of the trials, the initial direction of the joystick movement was opposite to the final choice. The authors suggested that such changes of mind are caused by additional information that is integrated following the initial choice. Importantly, this post-initial choice-integration was attributed to residual information in the processing pipeline. Presumably, pipeline information can support error monitoring functions as well.

Finally, in a series of studies, which were grounded in the choice-followed by confidence paradigm, Baranski and Petrusic found evidence for *confidence processing* both during and after the decision (Baranski & Petrusic, 1998, 2001; Petrusic & Baranski, 2003). The latter, post-decisional confidence-processing, was pronounced when choice-speed rather than choice-accuracy was stressed. Importantly, as we explained in the introduction, confidence processing does not necessarily entail *stimulus processing* in the sense of seeking novel information. For example, confidence processing may be limited to a confidence calculation, which interrogates the information that was available by the time the decision was made.

By manipulating the perceptual availability of the stimulus following the choice, we were able to show directly that post-choice information integration is operative and that it extends beyond a pipeline. Notably, a pipeline is operative whether the stimulus remains on the display or disappears. Additionally, the influence of a pipeline would be exhausted prior to the early confidence condition of Exp. 2. Hence, a pipeline is insufficient in accounting for the resolution differences between the *remain* and *vanish* conditions and between the early and late confidence conditions. Whereas the effects of a short pipeline may be unintentional and unavoidable, our findings suggest that post-choice integration is driven by a deliberative act of will. To the best of our knowledge, these are the first findings that point to the existence of such an extended post-choice integration process.

5.3. Insights from CCB: collapsing boundaries as a means for effort regulation

According to the CCB model, confidence judgments reflect a tradeoff between two conflicting desires for high confidence and for low integration costs. A similar idea but with respect to choice rather than confidence, has been prominent in recent decision making models (e.g. Deneve, 2012; Drugowitsch et al., 2012; Moran, 2014; Thura, Beauregard-Racine, Fradet, & Cisek, 2012; but see Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015). The main idea that underlies these models is that when difficulty level varies across trials (as a consequence of either an objective and/or a subjective variability in discriminability) collapsing thresholds provide an efficient method to implement a tradeoff between choice accuracy and decision time (i.e., to regulate integration effort) in optimizing reward rate.

Unlike collapsing choice-thresholds models, CCB assumes that the choice is governed by a stationary threshold, whereas confidence is governed by a noisy dynamically collapsing threshold. Initially, once the decision is made, participants place a high confidence threshold corresponding to the highest confidence level. However, as time passes and the trial fails to terminate, the cognitive system infers that it is more and more likely that the current-trial drift rate is low rather than high (because easier higher drift trials tend to reach the threshold earlier on). Such an inference in turn implies that it is less likely that the decision was correct in the first place, motivating a reduction in confidence. This reduction is achieved by collapsing the confidence boundary to a lower level, allowing judges to reduce temporal costs, as a lower boundary is usually reached faster.

This description, however, begs the question of why do not thresholds already collapse during the decision stage. One speculation is that during the decision stage, a collapsing-thresholds strategy would require observers to cognitively access and maintain two collapsing thresholds, a task which may be too demanding while evidence is integrated. In CCB, on the other hand, the task is simpler because a single collapsing boundary, which is activated by the decision, has to be accessed and implemented. Observers may thus opt to execute a standard stationary-thresholds decision stage, and to regulate their effort only during the second, confidence stage.

5.4. CCB: future developments and challenges

Our focal purpose in the current research was to establish the causal influence of post-choice integration on confidence and to suggest a mechanism that can account for the vast empirical manifold. Thus some more specific and yet, highly interesting questions were left outside the scope of the paper. First, can the CCB account for the effects that the perceptual availability manipulation exerted in Exp. 1, e.g., the increased RT2, resolution and TP effect in the *remain* relative to the vanish condition? The reduced resolution and TP effect can be accounted for in CCB by assuming that the drift rate decays after the stimulus disappears in the *vanish* condition hence dampening the second-stage differences between correct and erroneous trials. However, such drift-rate decay will tend to increase RT2 by slowing down the convergence towards the confidence boundary. Thus, to account for the perceptual availability effect the model will have to assume that in addition to a drift decay, the collapse dynamics speeds up in the *vanish* condition (e.g., by setting the confidence boundary to a lower initial heights, by increasing the collapse step and/or by speeding up collapses).

Second, is the CCB also relevant in confidence-interrogation scenarios, such as our Exp. 2? One hypothesis is that when confidence is interrogated, rather than determined freely, observers do not use a collapsing confidence boundary mechanism to terminate evidence integration, but rather exhaust the entire interrogation time for the post-decision stage. When the confidence signal arrives, observers map their total evidence on a confidence scale as in the 2DSD interrogation variant. Alternatively, if integration of information is effortful and costly, as the CCB assumes (see also Drugowitsch et al., 2012) then it is reasonable that observers use a collapsing confidence boundary even when confidence is interrogated. One such possibility is that if the confidence boundary is met prior to the response-signal, participants form a *covert confidence* judgment and delay their overt report until the response-signal arrives. On the other hand, if the response deadline arrived prior to the diffuser-confidence boundary intersection, participants resort to some kind of *guessing* strategy (e.g. report 50% confidence or: report a confidence level that corresponds to level below the current height of the confidence boundary). In such cases, the harmful confidence-TP effect may result from the fact that when the confidence signal arrives early rather than late, a larger portion of the confidence judgments are produces via the guessing (rather than the covert confidence) route.

Third, it should be tested whether CCB can account for fine-grained aspects of empirical RT2 distributions. For example, as illustrated in Fig. 5, the skewness of the RT2 distribution increases as a function of confidence. While the CCB fit captured this qualitative pattern it under-predicted the skew for all levels of confidence. Future studies should test whether this shortcoming could be ameliorated by using less constrained variants of CCB or perhaps by using skewed distributions for the collapse timers (recall that here we used uniform collapse timers).

5.5. CCB and alternative dual stage models

CCB and the optional stopping 2DSD are both dual-stage diffusion based models, which aim to account for confidence and its latency. These models share an identical decision-stage and differ only with respect to the second confidence stage, specifically—their termination rule. It remains to be seen in future research, whether the optional stopping 2DSD, like CCB, can account for the entire empirical manifold.

Unlike CCB and 2DSD, the response-reversals model (Van Zandt & Maldonado-Molina, 2004) is a dual-stage accumulator-based model with Poisson counters. Following the choice, the counters continue to accrue evidence towards a second set of thresholds and confidence is determined by the balance of evidence, when this second stage terminates. Interestingly, this model cannot account for some of the patterns in the empirical manifold. For example, for a given set of accumulation rates and thresholds, RT is predicted to correlate negatively with RT2 in violation of Hurdle 8. Indeed, the faster the first choice stage terminates, the lower the count on the loser racer at the moment of choice (the count for the winner is fixed by definition to the thresholds level). Since the first stage counts serves as ‘starting points’ for the second stage, a lower first-stage count, extends the duration of the second stage by reducing statistical facilitation.

Additional future viable dual stage model for the choice followed by confidence paradigm may be based on extensions of models that successfully account for confidence judgments in an alternative popular paradigm, wherein rather than choosing and ranking confidence sequentially, observers provide choice and confidence ratings simultaneously in a single compound response. Such models include, RTCON (Ratcliff & Starns, 2009) or its successor, RTCON2 (Ratcliff & Starns, 2013) and the recent ‘bounded accumulation model’ (Kiani et al., 2014; see also Zylberberg, Barttfeld, & Sigman, 2012). Notably, these two confidence-paradigms have been found at times to yield conflicting results. For example, in violation of Hurdle 9, Kiani, Corthell and Shadlen found that confidence in erroneous choices *decreased* as a function of difficulty. Additionally, Ratcliff and Starns (2013) presented data showing negative but also flat and even positive confidence-RT correlations, suggesting that Hurdle 4 may be less robust in the simultaneous choice-confidence, as compared with the choice-followed by confidence paradigm. Such differences between the two confidence paradigms beg a fundamental question: Do both confidence paradigms measure the same psychological construct? Or perhaps, despite the face similarity, they measure two different forms of confidence? If so, what is the relationship between these ‘two confidences’? Extending confidence-models to account for findings in both paradigms could thus integrate our knowledge from two separate confidence paradigms.

One possibility to extend RTCON(2) to the choice-followed by confidence paradigm is to follow up the first diffusion stage (for choice) with a second stage akin to RTCON2, featuring a race between accumulators that correspond to the alternative eligible confidence rankings. Here, however, the challenge is to develop a theory, which will connect the first stage parameters (e.g. the choice-diffusion drift) to the second stage parameters (the confidence drifts), hence allowing the model to account for correlations between first stage variables (such as RT) and second stage variables (such as confidence).

In the bounded accumulation model of Kiani et al. (2014), the choice is based on a race between two (anti) correlated accumulators. The novelty of this model lies in the suggestion that confidence is a function of not only the balance of evidence favoring the choice, but also the decision time. Specifically, based on trial by trial correctness feedback, observers learn to associate combinations of decision time and balance of evidence with the probability of choice-correctness (i.e., confidence). One possibility to extend this model to the choice followed by confidence paradigm is to introduce a second, post-choice, higher threshold in similarity with the response-reversals model (Van Zandt & Maldonado-Molina, 2004). Confidence can then be based the balance of evidence at the end of the second stage and on the total integration time (i.e., RT + RT2).

In summary, the recent confidence literature has witnessed a flourish of interesting and promising models of confidence judgments, including the optimal-stopping 2DSD, RTCON(2), the bounded accumulation model (Kiani et al., 2014) and the current addition, CCB. We believe that the next important step is to test and compare these models with respect to benchmark data sets. Specifically, future research should examine whether extensions of RTCON(2) and the bounded accumulation model (and as mentioned above, the optional-stopping 2DSD) can account for the entire empirical manifold. Then, direct head to head comparisons between these models should be conducted to identify which of the models provide the best quantitative fit, when model complexity and flexibility are controlled for. Hopefully, neurophysiological measures may also help in probing more directly the neural mechanism associated with confidence in the various models.

5.6. Confidence qua meta-cognition

In terms of signal detection theories, the choice followed by confidence paradigm elicits responses of two different types: choice—'type 1' and confidence—'type 2' (Galvin, Podd, Drga, & Whitmore, 2003). In type-1 tasks the objective is to detect the presence or absence of a signal. Hence, observers attempt to discriminate between objective external events. In type-II tasks, on the other hand, observers are asked about their belief in the correctness of the type-I response. Here, observers distinguish between one's own subjective states: a feeling of being correct or wrong. Thus, the referent of a type II judgment is an internal rather than an external event. The type-I and type-II tasks are thus considered to measure cognition (about external events) and meta-cognition (i.e. cognition about internal events) respectively.

Logically, a type-II belief could be reduced to a belief about the external presence of a signal. In other words, beliefs about choice-correctness may be formed by reconsidering the external evidence with respect to signal presence. However, logical reductionism does not necessarily translate to psychological identity. Whereas in type-I judgments, observers seek the external environment for relevant evidence, in a type II judgments they may search their own cognitive system for evidence with respect to correctness (e.g. was the decision quick and easy to arrive at?).

Current dual stage models such as CCB and 2DSD provide an interesting perspective with respect to the relationship between cognitive and metacognitive judgments. Unlike single stage theories such as BOE, which rely solely on internal representations to derive confidence (the post-decisional evaluation of the balance of the evidence between the competing alternatives), 2DSD and CCB continue to probe the external world, seeking for novel information. Notably, the post-decisional stage updates the same mental representation (the activation of a diffuser) that was generated during the choice stage. Hence, according to these models, in experimental designs such as those employed here, metacognition may be construed as an extension of cognition, rather than a qualitatively different process. Understanding the interplay between the cognitive and the metacognitive processes in a variety of confidence paradigms is an important issue for future research.

Acknowledgments

This research was funded by the German-Israel Foundation (158/2011) and by the Israeli Science Foundation (Grant 743/12). Andrei R. Teodorescu was also funded by the Fulbright Scholar Program. We thank Timothy Pleskac and Jerome Busemeyer for sharing their empirical data and Matlab code. We also thank Christopher Summerfield and Konstantinos Tsetsos for helpful discussions and Michael Brusovansky, Leor Brecher and Luda Gelfand for their assistance in running the experiments.

Appendix A. Sequential sampling: a brief summary

In describing the sequential sampling framework for decision making, we focus on two main approaches: random walk/diffusion and accumulator models (Ratcliff & McKoon, 2008; Teodorescu & Usher, 2013). In random walk/diffusion theory, as each time interval Δt passes after stimulus $S_i(i = A, B)$ is presented, judges consider a sample of information and transform it into evidence favoring one alternative over the other. Observers maintain a single tally, which corresponds to the total accumulated evidence $L(t)$. Evidence is accumulated until the first time t_D that the total $L(t)$ reaches an upper threshold θ or a lower threshold $-\theta$.[15]

One limiting aspect of this simple diffusion model is that when participants are unbiased, RTs for correct and for error choices predictably follow the same distributions. Nevertheless, the empirical reality is that hard and easy tasks often yield slow and fast errors respectively (Hurdle 2). To account for such findings the diffusion model has been augmented with two types of across trial variability. The first, starting point variability (Laming, 1968) enables the model to account for 'fast errors'. The second, drift rate variability (Ratcliff, 1978), which is usually attributed to (across trial) fluctuations in the levels of attention, alertness or perceptual efficiency, enables the model to account for 'slow errors'.[16] The diffusion model, augmented with these trial-by-trial viabilities, has been extremely successful in accounting for choice accuracy and the forms of the RT-distributions, for both correct and incorrect responses (e.g. Ratcliff & McKoon, 2008).

An alternative sampling process of choice could be found in the family of accumulator models. When judges are asked to make a 2AFC, evidence accrues in two counters, one in support of each response alternative. The first counter to reach a threshold determines the choice. Some examples of accumulator models are the Linear Ballistic Accumulators (Brown & Heathcote, 2008), the Poisson race model (Pike, 1973; Townsend & Ashby, 1983), the Accumulator Model (Usher, Olami, & McClelland, 2002; Vickers, 1979) and the Leaky Competing Accumulator Model (Usher & McClelland, 2001). One fundamental difference between random walk\diffusion and accumulator models is that they use relative vs. absolute stopping rules, respectively (Ratcliff & Smith, 2004). In particular, whereas in diffusion models the amount of evidence favoring the chosen (over the non-chosen) alternative is constant across trials, in accumulator models it is variable. In Section 1.4, we show that this difference between diffusion and accumulator models bears significant implications regarding the possibility to model confidence based on the balance of evidence favoring the chosen over the non-chosen alternative, within the diffusion and accumulator models.

Appendix B. A simulation study: a higher RT2 in the speed vs. accuracy condition is not a necessary condition for the beneficiary TP-effect on resolution

In accounting for the beneficiary TP effect on resolution (Hurdle 7), Pleskac and Busemeyer (2010) relied on the higher RT2 in the speed vs. the accuracy condition. However, in our experiments we

[15] The theory assumes that the amounts of evidence in different samples are stochastic, independent and identically distributed. Additionally, each evidence sample, collected in a temporal duration of Δt is distributed according to a Gaussian distribution with mean $\delta\Delta t$ and variance $\sigma^2\Delta t$. The parameter δ is the drift rate, the average advantage of response A over B or the rate of evidence accumulation over time, which indexes the average strength or quality of evidence that observers are able to extract from the stimulus. The parameter σ^2 is the drift coefficient, which indexes within-trial random fluctuations.

[16] Starting point variability is incorporated into the model by assuming that the starting point $L(0)$ varies uniformly across trials. Drift rate variability is incorporated into the model by assuming that for a single difficulty level and even for a single stimulus, the drift rate δ is distributed $\sim N(\delta_0, \eta^2)$ across presentations.

found a beneficiary TP effect when RT2 was not larger (and even smaller) in the speed condition. Here we show that an increased RT2 in the speed condition is not a necessary condition for 2DSD to predict Hurdle 7.

Let us assume that the inter-judgment time is identical under speed and accuracy conditions. Resolution depends on the amounts of evidence that are accrued during the second stage for correct and error responses. These amounts of evidence in turn, are determined by the second stage drift rates. We thus examined the second stage drift rates. To facilitate the presentation, we derived a coarse quantification of the resolution effect, as a function of these drift-rates.

Assume that the correct choice is A, and denote the mean, post –decisional, drift rates for correct and error responses by v_{correct} and v_{error} respectively (see Fig. 1) and the constant post-decision integration time by τ. For correct choices, the mean evidence in support of the choice (A) by the end of the second stage is $\theta + \tau * v_{\text{correct}}$ (see blue curve in Fig. 1). Erroneous choices, on the other hand will reach the bottom criterion and thus, by the end of the second stage the diffuser state will be $-\theta + \tau * v_{\text{error}}$ (note that for the red erroneous curve in Fig. 1, B is supposed to be the correct answer, whereas here we assume that A is the correct answer. Thus the current scenario would be obtained if the black and red curves and the vector v_{error} would be reflected to the bottom threshold). However, this state reflects evidence in favor of A, now the non-chosen alternative. Since confidence is framed in relation to the chosen alternative B, the amount of supporting evidence is obtained by negation: $\theta - \tau * v_{\text{error}}$. Assuming that confidence is proportional to the total amount of evidence supporting the decision, it follows that resolution of confidence is proportional to the difference between the mean levels of evidence for correct and erroneous choices, which translates into the sum of correct and error mean drift-rates:

$$\text{resolution} \propto \theta + \tau v_{\text{correct}} - (\theta - \tau v_{\text{error}}) = \tau * (v_{\text{correct}} + v_{\text{error}}), \tag{B1}$$

In words, the resolution *rate* (i.e. the rate with which resolution grows with respect to the post-choice integration time) generated during the second stage is proportional to the summed means of the drift rates conditional on correct and error choices and also to RT2 (note that this rate is the obtained by dividing the term in Eq. (B1) by τ).

In order to understand how the beneficiary TP effect on resolution (Hurdle 7) can emerge without differences in RT2, we examined, how these second stage drifts rates $(v_{\text{error}}, v_{\text{correct}})$ are differentially distributed for correct and for error choices across different TP regimes. Thus, we simulated the first stage diffusion model under speed and under accuracy conditions. In these simulations we used the best fitting parameters from the line length tasks of Pleskac and Busemeyer (2010).

Fig. B1 displays the typical result we obtained in our simulation. The figure corresponds to Participant 3 and to the third difficulty level. The black curve shows the drift rate density distribution across all trials: the mean drift rate is $v = 0.1418$ and the standard deviation is $\eta = 0.0876$. Note, that this distribution pertains to both the speed and accuracy SAT conditions because in the Pleskac and Busemeyer fits, the drift distribution was not influenced by that manipulation. The blue and red curves correspond to the drift rate density conditional on correct and erroneous choices respectively. The 'error conditional density' (for each SAT condition) is located to the left of the 'correct conditional density' since lower drift trials are more likely than higher-drift trials to lead to error responses.

The solid and dashed curves in Fig. B1 correspond to conditions that stress choice accuracy and speed respectively. Since drift-rate variability plays a less influential role in generating errors under TP (in this condition starting point variability and within trial noise are more dominant causes of errors), drift rate distributions conditional on response type assume different shapes with and without TP. Furthermore, since errors are less associated with low drift rates under TP, the mean error drift rate under TP is higher than without TP $(M_{\text{speed}} = 0.096, M_{\text{accuracy}} = 0.052)$. The TP differences between the drift rate distributions for correct responses, on the other hand, are more minor $(M_{\text{speed}} = 0.158, M_{\text{accuracy}} = 0.157)$. It follows that the sum of the mean drift rates conditional on correct and error decisions is larger for speed than for accuracy. Indeed, this sum was ~0.25 for the speed condition and a lower ~0.21 for accuracy conditions. Recall from Eq. (B1) that this sum is the rate with which resolution is generated during the second integration stage. In conclusion, because the resolution-rate is higher under TP, resolution will be higher (Hurdle 7) even if inter-judgment time is

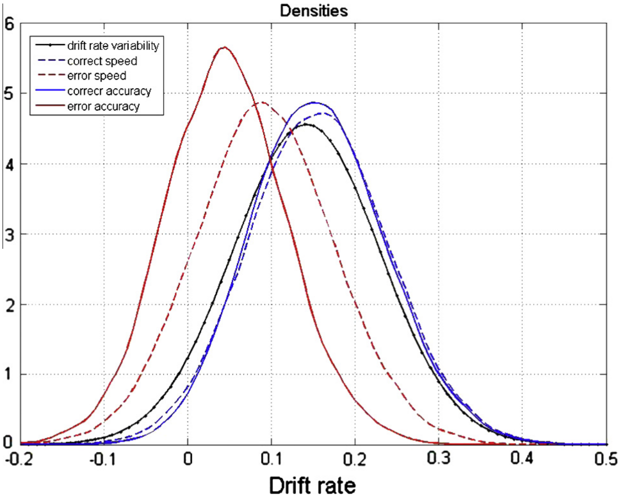

Fig. B1. Drift rate variability density curves across all trials (black) and conditional on correct (blue) and error (red) responses. Solid and dashed curves correspond to stress on choice accuracy and speed respectively. It can be clearly observed that the sum of the means for the correct and error curves, which constitutes the second stage resolution rate, is larger for the accuracy stress. Indeed, for correct responses the means are very similar whereas for error response the 'speed' mean is larger than the 'accuracy' mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

invariant with respect to the SAT regime (such as in the pipeline model). Furthermore, theoretically, due to higher resolution rate, higher resolution is predicted under speed even if under such conditions the inter-judgment time is moderately lower.

Note that the current demonstration ignored possible influences of TP on confidence mapping (Baranski & Petrusic, 1998; Pleskac and Busemeyer, 2010) and on accumulation drift rates (Rae et al., 2014). Such presumable influences may provide alternative routes to account for the beneficiary TP effect without the mediation of RT2.

Appendix C. Detailed experimental methods

C.1. Method of Experiment 1

C.1.1. Participants

The six participants (all women in the age-range of 22–25) were Tel Aviv University undergraduate psychology students. In return for their participation, they were rewarded with course credit and additionally, they were paid a performance-based reward (an amount of 30–40NIS, roughly equivalent to 8–12 $) in each of the 4–5 sessions. All the participants had normal or corrected-to normal vision.

C.1.2. Apparatus

All stimuli were presented using software programmed in Matlab with Psychtoolbox-3 (Brainard, 1997; Pelli, 1997; Kleiner, Brainard, & Pelli, 2007). This allowed for controlled presentation of graphics, instructions, event sequencing, timing, and recording of responses. Participants recorded their responses using a standard QWERTY keyboard. During the experiments participants placed their index, middle, medicinal and little fingers of the right hand on the keyboard keys 'H', 'J', 'K' and 'L' respectively. The corresponding left hand fingers were placed on the keys 'G', 'F', 'D' and 'S'. The thumbs of both hands were placed on the spacebar key. Participants used the spacebar key to indicate their readiness for the onset of the next trial and then entered their choice with either the 'L' (for choosing the right stimulus) or the 'S' (for choosing the left stimulus) keys. After making a choice, participants entered a confidence rating using the keys located between 'D' to 'K'. These keys corresponded to confidence levels 50%, 60%, ..., 100%. If the right-side stimulus was chosen then confidence increased from left to right i.e. 'D' corresponded to 50% confidence and 'K' corresponded to 100% with the intermediate keys corresponding to the intermediate confidence levels. If, on the other hand, the left-side choice was made then confidence increased from right to left i.e. 'D' corresponded to 100% confidence and 'K' corresponded to 50% confidence. Participants sat in individual sound-attenuated booths approximately 60 cm away from the screen.

C.1.3. Response entry task

Each experimental session began with a response entry practice task. On each trial participants entered a sequence of responses that simulated an expected sequence of responses on the main 2AFC experimental task. During this practice task, participants were instructed to enter a sequence of responses in the following manner: First, to initiate a trial, participant pressed the spacebar key. As soon as the spacebar was pressed the choice word 'left' or 'right' (written in Hebrew) appeared at the center of the screen and participants were instructed to press the 'S' or the 'L' key respectively. Only once the appropriate key was pressed, the choice word was replaced by one of the confidence labels 50, 60, ..., 100 and participants had to press the appropriate confidence key. A graphic confidence scale appeared below the confidence label. This scale extended from left (50) to right (100) if the 'left' choice was made or from right to left if the 'right' choice was made. The trial terminated only after participants pressed the correct confidence key. Each of the 2 (directions) $*$ 6 (confidence labels) appeared 12 times in the task for a total of 144 trials.

C.1.4. The 2AFC discrimination task

C.1.4.1. Stimuli. The basic display had a gray background. On each trial a pair of squared black and white arrays appeared on the screen. Each array was composed of 65 $*$ 65 squares where the edge size of each square was 4 pixels. Thus, the edge length of each array was 260 pixels. Vertically, theses arrays were located on the screen center. Horizontally, the array centers had an offset of 200 pixels to the left and to the right of the screen center. Each array pair consisted of one reference and one target array. In the reference array, half of the squares were white and half were black. The target array consisted of a majority of black squares whose proportion was drawn from one of the three values: 133/260 135/260, 137/260, corresponding to three different discriminability levels. The black and white arrangement of squares within each array was random.

C.1.4.2. Design and procedure. Participants completed four experimental sessions within a period of 2–3 weeks (Participant 5 participated in five sessions). Different sessions were conducted on different days. During each session participants completed two tasks. The first task was the previously described response entry task. The second task was the experimental task. During the experimental task, participants completed eight experimental blocks and two practice blocks. There were two types of blocks: Speed and accuracy blocks. During each experimental block (speed or accuracy) participants completed 72 randomly-ordered trials (3 discriminability levels $*$ 2 directions of the target array (left/right) $*$ 2 post-decisional perceptual availability (remain/vanish) $*$ 6 repetitions). For each trial a reference and target arrays were created anew, respecting the discriminability level. Each of the two practice blocks (speed or accuracy) contained only 18 trials (selected randomly from a set of 72 trials that was generated in the same way as in the experimental block). In all other respects the practice blocks were identical to the experimental blocks. The eight experimental blocks alternated in their type, emphasizing either choice speed or choice accuracy. The type of the first block (speed or accuracy) was counterbalanced for each participant across sessions. Finally, half of the participants began Session 1 with an accuracy block and half with a speed block. The first experimental block of each type was preceded by a practice block of the same type. A minimal 30sec break was enforced between consecutive experimental blocks. In total, participant completed 1152 (1440 for Participant 5) critical trials of each type (practice trials were discarded from the analysis).

Before each block of trials, participants were informed about their goal (speed or accuracy) for the upcoming block. Generally, participants were instructed to strive to achieve both response speed and accuracy but additionally, for each block they were asked to stress one over the other. For accuracy blocks, participants were instructed to enter their choice as accurately as possible even if this means that would slow down their choices. After each trial in the accuracy emphasis condition, feedback on incorrect responses was given on the feedback screen (described below) in the form of an "Error" message. The feedback screen was displayed after participants entered their confidence rating. For speed blocks participants were instructed to try and enter their choice quickly even if this policy results in a moderate compromise of accuracy. After each trial in the speed emphasis condition, feedback on responses slower than 750 ms was given in the form of a "Too Slow" message in the feedback screen. No feedback on error responses was given during the speed conditions. For both speed and accuracy blocks, participants were instructed to enter an accurate confidence rating. The instructions stressed that there is no time pressure on the confidence judgment.

An individual trial proceeded as follows. Participants were first given a preparation screen which asked them to press the spacebar key when they were ready to initiate the next trial. 250 ms after participants pressed the spacebar key, the two black and white arrays appeared to the left and right sides of the screen. Participants were instructed to choose the side with the array that contains a larger proportion of black. Once a choice was entered, a graphic confidence scale appeared below the arrays. This scale extended from left (50) to right (100) if the 'right' choice was made or from right to left if the 'left' choice was made.[17] For *remain* condition trials, the stimuli remained on the screen following the decision but for *vanish* trials, the appearance of the scale was coupled with the disappearance of the arrays from the display. Participants were instructed to judge their confidence in the correctness of the choice they just made ($50\%, 60\%, \ldots, 100\%$). After entering a confidence rating, a feedback screen appeared informing participants if they made an incorrect choice in the accuracy block or if they were too slow in the speed block. The feedback screen (which appeared after each trial) reported also the combined score (see below) for the current trial and the total accumulated score for the current block. Additionally, if either the choice or the confidence response occurred sooner than 150 ms following the stimuli or confidence scale onset respectively, the trial was aborted and a feedback screen asked participant to wait for the appropriate cue (i.e. the appearance of the arrays or of the confidence scale) before they respond. The trial was also aborted if at any stage of the trial a non-eligible key was pressed. In this case the feedback screen informed participants that they pressed the wrong key and they had to wait 4 s before they could proceed to the next trial. The feedback screen also served as the preparation screen for the next trial.

C.1.4.2.1. Combined choice-confidence accuracy score. At the beginning of the experiment, participants were told to select a confidence rating so that over the long run the proportion of correct choices for all trials assigned a given confidence rating should match the confidence rating given. Participants were reminded of this instruction before each session. This instruction is common in studies on the calibration of subjective probabilities (cf. Lichtenstein et al., 1982). As further motivation, on each trial participants earned points based on the accuracy of their choice and confidence rating according to the quadratic scoring rule (Stael von Holstein, 1970),

$$\text{points} = 100 * \left(1 - (\text{correct} - \text{confidence})^2\right) \tag{2}$$

where correct is a correct choice indicator (i.e. equal to 1 or 0, if the choice was correct or erroneous respectively) and confidence was the confidence rating entered in terms of probability of correct ($.50, .60, \ldots, 1.00$). This scoring rule is a variant of the Brier score (Brier, 1950), and as such it is a strictly proper scoring rule ensuring that participants will maximize their earnings only if they maximize their accuracy in both their choice and their confidence rating. Participants were informed of the properties of this scoring rule prior to each session and were shown a table demonstrating why it was in their

[17] One difference between the experiments of Pleskac and Busemeyer (2010) and the current experiment pertains to the response protocol. In Pleskac and Busemeyer's experiment participant made both the choice and the confidence responses using the same finger. We speculated that, perhaps (at least some part of) the increase in RT2 for the speed trials was due to a motor slowdown effect, and not due to longer post-choice integration stage. Specifically, we speculated that if the first choice is made under time pressure then the second choice with the same finger would be slower, as if the first speeded response induces a 'refractory motor period'. In our Exp. 1, to minimize interactions or carryover effects between the choice response and the confidence response, the two responses were elicited with different fingers. We hoped that the using different motor responses would weaken, perhaps even eliminate, this increased RT2 in the speeded choice for some of the participants. This may allow us to examine whether the increased RT2 is necessary for the emergence of the beneficial TP effect on resolution.

best interest to accurately report their choice and confidence rating. To enforce time pressure during the speed conditions, the points earned were cut in half if a choice exceeded the deadline of 750 ms. For every 1.600 points participants earned 1NIS (approximately 0.25$).

C.2. Method of Experiment 2

C.2.1. Participants

The six participants were Tel Aviv University undergraduate psychology students. Four of the participants participated also in Exp. 1 (Participants 2, 3, 4, 6). The two other Participants of Exp. 1 were unavailable and hence two new participants were recruited (Exp. 2 was conducted six months after Exp. 1). The two new participants (denoted Participants 7 and 9) were women in the same range age and were awarded with course credit. Participants, continuing from Exp. 1 needed no course credit, so instead they were reimbursed with double payments. Each participant carried out five sessions, each during approximately one hour.

C.2.2. Apparatus, the 2AFC discrimination task and stimuli

The Apparatus, the 2AFC discrimination task and the stimuli were identical to those used in Exp. 1.

C.2.3. Design and procedure

The design and procedure were similar in many respects to those of Exp. 1 so we describe only the differences. Participants completed five experimental sessions within a period of 3–4 weeks. In Exp. 2 all experimental blocks stressed choice speed rather than accuracy (as in the speed blocks of Exp. 1). There were two types of experimental blocks, early and late confidence interrogation blocks. Both the experimental and practice blocks were created and arranged (interleaved) in the same way as in Exp. 1, the only change being that the two block types were early vs. late confidence interrogation rather than choice speed vs. accuracy stress.

With respect to choice, participants were given the same instructions as in the speed blocks of Exp. 1. However, the instructions with respect to confidence differed as participants were asked to enter their confidence judgments only after an auditory cue (a short beep) is delivered.

The sequence of events on an individual trial is similar to that of Exp. 1 up to the presentation of the confidence scale. Once the scale was presented, participants were instructed to continue thinking about their confidence judgment until they hear the beep. The beep sounded either 300 ms or 1300 ms after the onset of the confidence scale in the early and late conditions respectively. After entering a confidence rating, feedback was given if the choice was too slow (more than 750 ms from stimuli onset) or if the confidence rating was too slow (more than 1 s from the beep onset) and the amounts of earned points for the trial was cut by half. Additionally, if either the choice or the confidence response occurred earlier than 150 ms after the stimuli or the beep onset respectively, the trial was aborted and a feedback screen asked participant to wait for the appropriate cue before they respond.

C.3. Method of Experiment 3

C.3.1. Participants

The nine participants were Tel Aviv University undergraduate psychology students. Three of the participants participated also in Exp. 1 and 2 (Participants 2, 4, 6) and one participant participated only in Exp. 2 (Participant 7). Six new participants were recruited but one was excluded as she failed to exceed chance performance in the speed condition. The five remaining novel participants (denoted Participants 8, 11, 12, 15 and 16) were in the same age-range and were awarded as in Exp. 1. Participants, continuing from Exp. 1 or 2 needed no course credit, so instead they were reimbursed with double payments. Each participant carried out two sessions, each during approximately one hour.

C.3.2. Apparatus, the 2AFC discrimination task and stimuli

The Apparatus, the 2AFC discrimination task and the stimuli were identical to those used in Exp. 1.

C.3.3. Design and procedure

The design and procedure were similar in many respects to those of Exp. 1 so we describe only the differences. Participants completed two experimental sessions within a period of one week. The sequence of events on an individual trial is similar to that of Exp. 1 up to the choice execution. Once the choice was executed black and white chess-board masking patterns replaced location of the two arrays and the confidence scale appeared. Thus, the *remain* and *vanish* conditions of Exp. 1 were replaced with a single masking condition.

Appendix D. Individual participant data pertaining to Hurdles 8–10

D.1. Hurdle 8: the RT2 correlations

See Tables D1 and D2.

Table D1
RT2 for correct and erroneous choices in Exp. 1 and 3.

Exp. 1			Exp. 3		
Par	RT2 error	RT2 correct	Par	RT2 error	RT2 correct
1	0.61(0.01)	0.55(0.00)***	2	0.53(0.01)	0.47(0.00)***
2	1.35(0.03)	1.01(0.01)***	4	0.84(0.03)	0.54(0.01)***
3	0.41(0.00)	0.39(0.00)***	6	0.42(0.01)	0.37(0.00)***
4	1.34(0.04)	0.72(0.01)***	7	0.48(0.01)	0.43(0.00)***
5	0.73(0.01)	0.61(0.01)***	8	0.60(0.01)	0.54(0.01)**
6	0.58(0.01)	0.45(0.01)***	11	0.39(0.01)	0.36(0.00)*
			12	0.68(0.02)	0.44(0.01)***
			15	0.80(0.02)	0.72(0.01)**
			16	0.89(0.02)	0.70(0.01)***
Group	0.83(0.10)	0.62(0.06)***	Group	0.62(0.05)	0.51(0.03)***

Note. RT2 is measured in seconds. Values in parentheses are standard errors. $*, **, ***$ indicate $p < .05, .01, .001$ respectively according to a z-test for individual or a meta-analysis for the group.

Table D2
Correlations between RT2 and discriminability, RT and confidence for Exp. 1 and 3.

Exp1				Exp3			
Par	Discriminability	RT	Confidence	Par	Discriminability	RT	Confidence
1	.01(0.02)	.19(0.02)***	−.20(0.02)***	2	−.24(0.03)***	.14(0.03)***	−.53(0.02)***
2	−.26(0.02)***	.23(0.02)***	−.51(0.02)***	4	−.27(0.03)***	.23(0.03)***	−.79(0.01)***
3	−.03(0.02)	.06(0.02)***	−.36(0.02)***	6	−.07(0.03)*	−.02(0.03)	−.36(0.02)***
4	−.15(0.02)***	.16(0.02)***	−.69(0.01)***	7	−.10(0.03)***	.40(0.03)***	−.31(0.02)***
5	−.12(0.02)***	.08(0.02)***	−.44(0.01)***	8	−.07(0.03)*	.29(0.03)***	−.65(0.02)***
6	−.08(0.02)***	.03(0.02)*	−.39(0.01)***	11	−.07(0.03)*	.07(0.03)***	−.51(0.05)***
				12	−.11(0.03)***	−.02(0.03)	−.83(0.02)***
				15	−.05(0.03)	.21(0.03)***	−.24(0.02)***
				16	−.12(0.03)***	−.04(0.03)	−.53(0.02)***
Group	−.10(0.04)**	.13(0.03)**	−.43(0.07)***	Group	−.12(0.03)***	.14(0.05)**	−.53(0.07)***

Note. The table presents Γ correlations for discriminability and confidence and Pearson's correlations for RT. Values in parentheses are standard errors, which for the individual participants were calculated based on bootstrapping. $*, **, ***$ designate $p < .05, .01, .001$, respectively based on permutation tests for the individuals and a meta-analysis for the group.

D.2. Hurdles 9–10: the discriminability * accuracy interactions

See Tables D3 and D4.

Table D3
The interaction between choice correctness and stimulus discriminability on confidence and the simple effects of discriminability for correct and error choices separately, for Exp. 1–3.

Participant	CORRECT*DISC	DISC(correct)	DISC(error)
101	0.31(0.06)***	0.18(0.03)***	−0.12(0.05)*
102	0.80(0.08)***	0.46(0.03)***	−0.34(0.07)***
103	0.48(0.07)***	0.13(0.04)**	−0.35(0.06)***
104	1.04(0.09)***	0.36(0.04)***	−0.68(0.08)***
105	0.57(0.06)***	0.32(0.03)***	−0.25(0.05)***
106	0.34(0.07)***	0.14(0.04)***	−0.20(0.06)***
Group (Exp. 1)	0.59(0.10)***	0.27(0.06)***	−0.32(0.07)***
202	0.88(0.07)***	0.49(0.03)***	−0.39(0.07)***
203	0.70(0.06)***	0.34(0.04)***	−0.36(0.04)***
204	0.79(0.07)***	0.37(0.04)***	−0.42(0.06)***
206	0.55(0.05)***	0.32(0.03)***	−0.23(0.04)***
207	0.51(0.06)***	0.25(0.03)***	−0.26(0.05)***
209	0.47(0.06)***	0.32(0.03)***	−0.15(0.04)***
Group (Exp. 2)	0.65(0.06)***	0.35(0.04)***	−0.30(0.04)***
302	0.85(0.13)***	0.51(0.05)***	−0.34(0.12)**
304	0.97(0.15)***	0.45(0.05)***	−0.52(0.14)***
306	0.66(0.09)***	0.29(0.05)***	−0.37(0.08)***
307	0.02(0.13)	0.13(0.04)**	0.11(0.12)
308	0.44(0.09)***	0.23(0.05)***	−0.20(0.07)**
311	−0.14(0.15)	0.09(0.07)	0.23(0.13)
312	0.86(0.12)***	0.39(0.07)***	−0.47(0.09)***
315	0.57(0.1)***	0.35(0.04)***	−0.22(0.09)*
316	0.36(0.09)***	0.27(0.05)***	−0.09(0.08)
Group (Exp. 3)	0.51(0.11)***	0.30(0.05)***	−0.21(0.07)**

Note. The table displays the coefficients of a multiple probit-ordinal regression. Values in parentheses are standard errors. *, **, * * * indicate p < .05, .01, .001 respectively according to t-test for the participants and a meta-analysis for the group.

Table D4
The interaction between choice correctness and stimulus discriminability on RT2 and the simple effects of discriminability for correct and error choices separately, for Exp. 1–3.

Participant	CORRECT*DISC	DISC(correct)	DISC(error)
101	−0.02(0.01)	0.00(0.01)	0.02(0.01)*
102	−0.16(0.04)***	−0.16(0.02)***	−0.01(0.04)
103	−0.01(0.01)	0.00(0.00)	0.01(0.01)
104	−0.28(0.04)***	−0.09(0.02)***	0.19(0.04)***
105	−0.03(0.02)	−0.06(0.01)***	−0.03(0.02)
106	−0.03(0.02)	−0.02(0.01)**	0.01(0.01)
Group (Exp. 1)	−0.07(0.02)**	−0.05(0.02)**	0.02(0.01)
302	−0.02(0.01)*	−0.03(0.00)***	0.00(0.01)
304	−0.10(0.04)**	−0.08(0.01)***	0.02(0.03)
306	−0.01(0.01)	−0.01(0.00)	0.01(0.01)
307	−0.01(0.01)	−0.01(0.00)**	−0.01(0.01)
308	−0.09(0.02)***	−0.03(0.01)**	0.05(0.02)**
311	0.01(0.02)	−0.01(0.01)	−0.02(0.01)
312	−0.10(0.02)***	−0.04(0.01)***	0.06(0.02)**
315	−0.06(0.04)	−0.02(0.02)	0.04(0.03)
316	−0.04(0.02)*	−0.04(0.01)***	0.01(0.02)
Group (Exp. 3)	−0.04(0.01)***	−0.03(0.01)***	0.01(0.01)

Note. The table displays the coefficients of a multiple linear regression. Values in parentheses are standard errors. *, **, * * * indicate p < .05, .01, .001 respectively according to t-test for the participants and a meta-analysis for the group.

Appendix E. Resolution analysis with additional measures

E.1. A brief description of the resolution measure

To recapitulate, resolution of confidence pertains to the relation between choice-accuracy and confidence. Perhaps the simplest operative definition of this relation is given by the slope scores (Yates, 1990), which are the difference between mean confidence for correct and incorrect decisions (Eq. (E1)). Another relevant statistic, the scatter score (Yates), is defined as a weighted estimate of the variance of confidence judgments for the correct and incorrect choices, where the averaging weights are proportional to the prevalence of correct and error responses (Eq. (E2)). When a comparison of resolution of confidence between experimental conditions is conducted, increased slope scores

Table E1
Resolution of confidence measures as function of the SAT and the perceptual availability manipulations of Exp. 1.

Par	1	2	3	4	5	6	Group
SLOPE							
speed	**0.06(0.01)**	**0.21(0.01)**	**0.19(0.01)**	**0.3(0.01)**	**0.22(0.01)**	**0.15(0.01)**	**0.19(0.04)**
accuracy	**0.08(0.01)**	**0.14(0.01)*****	**0.10(0.01)*****	**0.09(0.02)*****	**0.18(0.01)*****	**0.11(0.01)***	**0.12(0.02)***
DI′							
speed	**0.47(0.07)**	**1.86(0.11)**	**1.07(0.09)**	**2.31(0.15)**	**1.10(0.07)**	**0.93(0.08)**	**1.28(0.23)**
accuracy	**0.66(0.09)**	**1.41(0.13)****	**0.77(0.10)***	**0.98(0.19)*****	**0.99(0.08)**	**0.79(0.09)**	**0.92(0.1)***
Γ							
speed	**0.31(0.04)***	**0.85(0.02)**	**0.67(0.03)**	**0.90(0.01)**	**0.68(0.03)**	**0.58(0.03)**	**0.67(0.07)**
accuracy	**0.46(0.05)**	**0.73(0.04)****	**0.50(0.05)****	**0.60(0.07)*****	**0.61(0.03)**	**0.49(0.04)**	**0.57(0.04)***
A_g							
speed	**0.12(0.02)***	**0.37(0.01)**	**0.22(0.02)**	**0.37(0.01)**	**0.26(0.01)**	**0.23(0.02)**	**0.26(0.04)**
accuracy	**0.17(0.02)**	**0.29(0.02)*****	**0.15(0.02)****	**0.19(0.03)*****	**0.23(0.02)**	**0.19(0.02)***	**0.20(0.02)***
SLOPE							
remain	**0.07(0.01)**	**0.24(0.01)**	**0.17(0.01)**	**0.30(0.02)**	**0.22(0.01)**	**0.12(0.01)**	**0.19(0.03)**
vanish	**0.07(0.01)**	**0.13(0.01)*****	**0.14(0.01)**	**0.23(0.01)****	**0.19(0.01)**	**0.15(0.01)**	**0.15(0.02)**
DI′							
remain	**0.60(0.08)**	**2.15(0.14)**	**1.10(0.09)**	**2.61(0.20)**	**1.18(0.08)**	**0.83(0.09)**	**1.39(0.23)**
vanish	**0.58(0.07)**	**1.30(0.10)*****	**0.88(0.09)**	**1.87(0.14)****	**0.99(0.08)**	**0.96(0.08)**	**1.08(0.15)***
Γ							
remain	**0.41(0.04)**	**0.85(0.02)**	**0.65(0.04)**	**0.89(0.02)**	**0.70(0.03)**	**0.53(0.04)**	**0.68(0.07)**
vanish	**0.39(0.04)**	**0.78(0.03)***	**0.57(0.04)**	**0.81(0.03)***	**0.63(0.03)**	**0.58(0.03)**	**0.63(0.06)****
A_g							
remain	**0.16(0.02)**	**0.38(0.01)**	**0.21(0.02)**	**0.35(0.02)**	**0.27(0.01)**	**0.20(0.02)**	**0.26(0.03)**
vanish	**0.15(0.02)**	**0.30(0.02)****	**0.18(0.02)**	**0.33(0.02)**	**0.23(0.01)**	**0.23(0.02)**	**0.24(0.03)**
TP slope							
remain	−0.02(0.02)	**0.10(0.03)**	**0.10(0.03)**	**0.24(0.03)**	**0.06(0.02)**	**0.05(0.02)**	**0.09(0.03)**
vanish	−0.02(0.02)	**0.04(0.02)**	**0.08(0.03)**	**0.19(0.03)**	0.02(0.03)	0.03(0.02)	**0.05(0.03)***
TP DI′							
remain	−0.12(0.15)	**0.69(0.30)**	0.37(0.20)	**1.64(0.40)**	0.29(0.16)	0.24(0.18)	**0.40(0.17)**
vanish	−0.25(0.16)	0.22(0.19)	0.23(0.18)	**1.23(0.32)**	−0.05(0.16)	0.03(0.17)	0.17(0.15)*
TP Γ							
remain	−0.13(0.09)	**0.17(0.06)**	**0.18(0.08)**	**0.27(0.08)**	**0.12(0.06)**	0.14(0.08)	**0.13(0.05)**
vanish	−0.16(0.09)	0.05(0.07)	**0.16(0.09)**	**0.36(0.14)**	0.02(0.07)	0.04(0.07)	0.07(0.05)
TP A_g							
remain	−0.05(0.03)	**0.12(0.04)**	**0.08(0.04)**	**0.16(0.05)**	**0.06(0.03)**	0.06(0.04)	**0.07(0.03)**
vanish	−0.06(0.04)	0.03(0.04)	0.06(0.03)	**0.20(0.05)**	0.00(0.03)	0.03(0.04)	0.03(0.03)

Note. Values in parentheses are standard errors, which for participants were calculated based on bootstrapping (except for slope based measures). Bold numbers indicate difference from 0 using $p < .05$ (two sided). $*$, $**$, $* * *$ indicate the condition (speed vs. accuracy or remain vs. vanish) in which the relevant statistic was smaller using $p < .05$, .01, .001 (two sided), respectively. Statistical inferences are based on permutation tests for the participants (except for z-tests, for slope based measures) or on a meta-analysis for the group.

may be paired with an increase in scatter. This increase in scatter may detract from the increase in slope, in terms of a judge's resolution (Wallsten, Budescu, Erev, & Diederich, 1997; Yates & Curley, 1985). This motivates the use of a standardized measure of resolution dubbed *DI′* and defined in Eq. (E3)(Wallsten et al.).

$$\text{slope} = \overline{\text{conf}}_{\text{correct}} - \overline{\text{conf}}_{\text{incorrect}} \tag{E1}$$

$$\text{scatter} = \frac{n_{\text{correct}}\text{var}(\text{conf}_{\text{correct}}) + n_{\text{incorrect}}\text{var}(\text{conf}_{\text{incorrect}})}{n_{\text{correct}} + n_{\text{incorrect}}} \tag{E2}$$

$$\text{DI}' = \frac{\text{slope}}{\sqrt{\text{scatter}}} \tag{E3}$$

Table E2

Resolution of confidence measures for the different confidence and perceptual availability manipulations in Exp. 2.

Par	2	3	4	6	7	9	Group
SLOPE							
Early confidence	**0.26(0.01)**	**0.22(0.01)***	**0.31(0.01)**	**0.21(0.01)*****	**0.19(0.01)**	**0.19(0.01)**	**0.23(0.02)**$**$
Late confidence	**0.28(0.01)**	**0.25(0.01)**	**0.31(0.01)**	**0.26(0.01)**	**0.20(0.01)**	**0.21(0.01)**	**0.25(0.02)**
DI′							
Early confidence	**2.04(0.12)**	**1.58(0.09)**	**2.24(0.14)**	**1.28(0.07)****	**1.38(0.09)**	**1.14(0.07)**	**1.60(0.15)*****
Late confidence	**2.33(0.12)**	**1.76(0.09)**	**2.24(0.13)**	**1.62(0.08)**	**1.48(0.09)**	**1.30(0.08)**	**1.78(0.15)**
Γ							
Early confidence	**0.85(0.02)***	**0.78(0.02)***	**0.89(0.01)**	**0.70(0.02)*****	**0.65(0.03)**	**0.64(0.03)**	**0.75(0.04)****
Late confidence	**0.91(0.01)**	**0.84(0.02)**	**0.89(0.01)**	**0.80(0.02)**	**0.72(0.03)**	**0.69(0.02)**	**0.81(0.03)**
A_g							
Early confidence	**0.37(0.01)***	**0.32(0.01)**	**0.35(0.01)***	**0.30(0.01)*****	**0.28(0.02)**	**0.27(0.01)**	**0.32(0.02)*****
Late confidence	**0.41(0.01)**	**0.35(0.01)**	**0.39(0.01)**	**0.36(0.01)**	**0.32(0.01)**	**0.30(0.01)**	**0.35(0.02)**
SLOPE							
remain	**0.29(0.01)**	**0.26(0.01)**	**0.36(0.01)**	**0.26(0.01)**	**0.21(0.01)**	**0.22(0.01)**	**0.26(0.02)**
vanish	**0.24(0.01)****	**0.21(0.01)*****	**0.26(0.01)*****	**0.22(0.01)****	**0.18(0.01)***	**0.18(0.01)****	**0.21(0.01)*****
DI′							
remain	**2.42(0.13)**	**1.84(0.10)**	**2.81(0.18)**	**1.57(0.08)**	**1.58(0.09)**	**1.34(0.08)**	**1.90(0.18)**
vanish	**1.98(0.11)****	**1.51(0.09)***	**1.80(0.11)*****	**1.33(0.07)***	**1.29(0.09)***	**1.10(0.07)***	**1.49(0.13)*****
Γ							
remain	**0.90(0.02)**	**0.84(0.02)**	**0.94(0.01)**	**0.79(0.02)**	**0.72(0.03)**	**0.71(0.02)**	**0.82(0.04)**
vanish	**0.86(0.02)**	**0.78(0.02)***	**0.83(0.02)*****	**0.71(0.02)***	**0.65(0.03)**	**0.62(0.03)***	**0.74(0.04)*****
A_g							
remain	**0.41(0.01)**	**0.35(0.01)**	**0.39(0.01)**	**0.35(0.01)**	**0.32(0.01)**	**0.30(0.01)**	**0.35(0.02)**
vanish	**0.38(0.01)**	**0.31(0.01)****	**0.35(0.01)**	**0.31(0.01)***	**0.28(0.02)**	**0.26(0.01)***	**0.32(0.02)*****
TP slope							
remain	−**0.04(0.02)**	−**0.04(0.02)**	−0.03(0.02)	−**0.07(0.02)***	−0.01(0.02)	−0.03(0.02)	−**0.04(0.01)****
vanish	0.00(0.02)	−0.01(0.02)	0.03(0.02)	−0.02(0.02)	0.00(0.02)	−0.01(0.02)	0.00(0.01)
TP DI′							
remain	−0.41(0.26)	−**0.38(0.19)**	−0.28(0.37)	−**0.58(0.16)***	−0.19(0.18)	−0.22(0.16)	−**0.35(0.08)****
vanish	−0.15(0.22)	0.00(0.17)	0.19(0.22)	−0.14(0.14)	−0.02(0.17)	−0.11(0.15)	−0.06(0.07)
TP Γ							
remain	−0.05(0.03)	−**0.10(0.03)**	−0.01(0.02)	−**0.15(0.04)**	−0.08(0.06)	−0.05(0.05)	−**0.07(0.02)**
vanish	−**0.07(0.04)**	−0.01(0.04)	0.03(0.05)	−0.05(0.05)	−0.08(0.07)	−0.03(0.06)	−0.02(0.02)
TP A_g							
remain	−0.04(0.02)	−**0.05(0.02)**	−**0.05(0.02)**	−**0.08(0.02)**	−0.04(0.03)	−0.03(0.03)	−**0.05(0.01)***
vanish	−0.04(0.02)	−0.01(0.03)	−0.01(0.03)	−0.03(0.02)	−0.03(0.03)	−0.02(0.03)	−**0.02(0.01)**

Note. Values in parentheses are standard errors, which for participants were calculated based on bootstrapping (except for slope based measures). Bold numbers indicate difference from 0 using $p < .05$ (two sided). $*$, $**$, $***$ indicate the condition (early vs. late confidence or remain vs. vanish) in which the relevant statistic was smaller using $p < .05, .01, .001$ (two sided), respectively. Statistical inferences are based on permutation tests for the participants (except for *z*-tests, for slope based measures) or on a meta-analysis for the group.

Table E3
Resolution of confidence measures as function of the SAT and the perceptual availability manipulations of Exp. 3.

	Slope		DI′		Γ		A_g	
Par	Speed	Accuracy	Speed	Accuracy	Speed	Accuracy	Speed	Accuracy
2	**0.23(0.02)**	**0.19(0.02)**	**1.83(0.19)**	**1.74(0.21)**	**0.81(0.04)**	**0.78(0.05)**	**0.34(0.02)**	**0.32(0.03)**
4	**0.29(0.02)**	**0.15(0.04)**$\ast\ast\ast$	**2.28(0.22)**	**1.42(0.35)***	**0.87(0.02)**	**0.69(0.08)*****	**0.38(0.02)**	**0.26(0.05)****
6	**0.17(0.02)**	**0.14(0.02)**	**1.07(0.11)**	**0.99(0.14)**	**0.60(0.05)**	**0.53(0.06)**	**0.25(0.02)**	**0.22(0.03)**
7	**0.10(0.02)**	**0.11(0.02)**	**0.86(0.15)**	**0.90(0.18)**	**0.44(0.08)**	**0.49(0.09)**	**0.17(0.03)**	**0.20(0.04)**
8	**0.07(0.02)**	**0.04(0.01)**	**0.42(0.10)**	**0.27(0.10)**	**0.34(0.06)**	**0.18(0.07)**	**0.12(0.02)**	**0.06(0.03)**
11	**0.04(0.01)**	**0.03(0.01)**	**0.42(0.13)**	**0.37(0.16)**	**0.41(0.10)**	**0.44(0.12)**	**0.07(0.02)**	**0.07(0.02)**
12	**0.19(0.02)**	**0.14(0.03)**	**1.22(0.13)**	**1.50(0.27)**	**0.75(0.04)**	**0.80(0.05)**	**0.24(0.02)**	**0.25(0.03)**
15	**0.17(0.01)**	**0.10(0.02)****	**1.03(0.09)**	**0.69(0.13)***	**0.67(0.05)**	**0.44(0.07)****	**0.28(0.02)**	**0.19(0.03)***
16	**0.25(0.02)**	**0.15(0.02)*****	**1.60(0.14)**	**1.09(0.14)*****	**0.78(0.03)**	**0.60(0.06)****	**0.34(0.02)**	**0.25(0.03)****
Group	**0.17(0.03)**	**0.12(0.02)*****	**1.17(0.18)**	**0.96(0.16)***	**0.64(0.05)**	**0.56(0.07)***	**0.24(0.04)**	**0.20(0.03)****

Note. Values in parentheses are standard errors, which for participants were calculated based on bootstrapping (except for slope). Bold numbers indicate difference from 0 using $p < .05$ (two sided). $\ast$, $\ast\ast$, $\ast\ast\ast$ indicate the condition (speed vs. accuracy) in which the relevant statistic was smaller using $p < .05$, .01, .001 (two sided), respectively. Statistical inferences are based on permutation tests for the participants (except for z-tests, for slope based measures) or on a meta-analysis for the group.

Importantly, these measures assume that the values of the confidence judgments emerge from the use of an interval scale. As explained in the main text (Section 2.2.1), this assumption may be problematic. Thus, we augmented our set of analyses by additionally using two measures that merely postulate an ordinal structure of the confidence scale. The first measure is the Goodman and Kruskal Γ correlation (Goodman & Kruskal, 1954; henceforth we refer to it simply as the Γ correlation). The second, is the area that is locked between the type-II ROC curve and the diagonal, denoted by A_g. A_g is calculated as follows: For each confidence category $c = 50\%, 60\%, \ldots, 100\%$ we plot the proportion of correct choices that yielded confidence of at least c (ordinate) against the proportion of error choices that yielded confidence of at least c (abscissa). Next, we add the two points $(0,0)$ and $(1,1)$ to the plot and we connect consecutive points with straight lines to obtain the type-II ROC curve. A positive resolution is evident by an 'above diagonal' tendency of the curve. Formally, this trend is gauged by subtracting 0.5, the area below the non-resolution diagonal, from the area of the curve.

E.2. Results

The top part of Table E1 displays the resolution of confidence measures for the different SAT conditions in Exp. 1. The middle section of Table E1 lists the resolution of confidence measures for the *remain* and *vanish* conditions. Finally, the bottom section of Table E1 lists the TP effect on resolution for the *remain* and *vanish* conditions, which was calculated by subtracting the resolution in the choice-accuracy condition from the resolution in the speed condition. Table E2, refers to Exp. 2 and was calculated similarly with the single change that the choice-speed and accuracy conditions were replaced with the early and late confidence conditions, respectively. Finally, Table E3 displays the resolution of confidence measures for the different SAT conditions in Exp. 3. The results for each of the measures are similar to the results, which were reported in the main text, based on multiple ordinal regressions.

Appendix F. Replicating empirical Hurdles 3–4 in Experiments 1–3

F.1. Experiment 1

First, to test the relationship between confidence and stimulus discriminability, we calculated for each participant the Γ correlation (Goodman & Kruskal, 1954) between confidence judgments and stimulus discriminability. The first data row in Table F1 displays these correlations. For all participants as well as for the entire group, the correlation between confidence judgments and stimulus discriminability was significantly positive replicating Hurdle 3. Second, we tested the relationship between confidence and RT *within speed and accuracy blocks separately*. Thus, for each participant and for each SAT condition we calculated the Γ correlation of the confidence judgments and decision RT. We then

Table F1
Γ correlations between confidence and discriminability and between confidence and decision-RT in Exp. 1.

Par	1	2	3	4	5	6	Group
Disc	.13(0.02)***	.39(0.02)***	.07(0.03)*	.26(0.03)***	.21(0.02)***	.14(0.03)***	.20(0.05)***
RT	−.15(0.02)***	−.30(0.02)***	−.13(0.02)***	−.31(0.02)***	−.02(0.02)	−.09(0.02)***	−.17(0.05)***

Note. Values in parentheses are standard errors, which for participants were calculated based on bootstrapping. $*, **, ***$ indicate $p < .05, .01, .001$ (two-sided), respectively based on permutation tests for the individuals and a meta-analysis for the group.

averaged these correlations across SAT conditions to obtain single confidence-RT correlation estimates, which are displayed in the second row of Table F1. For five of participants as well as for the whole group, the correlation between confidence judgments and decision RTs was significantly negative replicating Hurdle 4.

F.2. Experiment 2

The first data row in Table F2 lists the Γ correlation between confidence and stimulus discriminability. For all participants as well as for the entire group, the correlation was significantly positive in accord with Hurdle 3. The second data row in Table F2 displays the Γ correlation between confidence and choice-RT. For four of the participants, this correlation was significantly negative as predicted by Hurdle 4. However, for Participant 2 the correlation was positive and the negativity for the group reached significance only according to a one-sided test $(z = -1.851, p = .032)$.

F.3. Experiment 3

As shown in Table F3, for the group and for all the participants, the correlation between confidence and discriminability was significantly positive replicating Hurdle 3 $(\Gamma = .25, z = 6.44, p < .0001)$. Additionally, the correlation between confidence and RT was negative, replicating Hurdle 4 $(\Gamma = -.19, z = -4.53, p < .0001)$. This negativity was significant for all but two participants (12 and 16).

Table F2
Γ correlation between confidence and discriminability and between confidence and RT for Exp. 2.

Par	2	3	4	6	7	9	Group
Disc	.41(0.02)***	.15(0.02)***	.30(0.03)***	.16(0.02)***	.22(0.02)***	.22(0.02)***	.24(0.04)***
RT	−.22(0.02)***	.06(0.02)**	−.15(0.02)***	.02(0.02)	−.11(0.02)***	−.06(0.02)***	−.08(0.04)

Note. Values in parentheses are standard errors, which for participants were calculated based on bootstrapping. $*, **, ***$ indicate $p < .05, .01, .001$ (two-sided), respectively based on permutation tests for the individuals and a meta-analysis for the group.

Table F3
Γ correlations between confidence and discriminability and between confidence and decision-RT in Exp. 3.

Par	Confidence-discriminability	Confidence-RT
2	.44(0.04)***	−.34(0.03)***
4	.42(0.04)***	−.36(0.03)***
6	.18(0.03)***	−.06(0.03)*
7	.17(0.04)***	−.20(0.02)***
8	.12(0.04)**	−.20(0.03)***
11	.22(0.07)**	−.26(0.05)***
12	.23(0.05)***	−.04(0.04)
15	.27(0.03)***	−.24(0.02)***
16	.21(0.04)***	−.02(0.03)
Group	.25(0.04)***	−.19(0.04)***

Note. Values in parentheses are standard errors, which for participants were calculated based on bootstrapping. $*, **, ***$ indicate $p < .05, .01, .001$ (two-sided), respectively based on permutation tests for the individuals and a meta-analysis for the group.

Appendix G. Tables for individual participants data

See Tables G1–G3.

Table G1
Accuracy rate, mean decision time, mean confidence and mean inter-judgment time for each participant and for the group for the speed and accuracy conditions and for the remain and vanish condition of Exp. 1.

Par	1	2	3	4	5	6	Group
ACC							
speed	0.65(0.01)***	0.79(0.01)***	0.69(0.01)***	0.73(0.01)***	0.61(0.01)***	0.67(0.01)***	0.69(0.03)***
accuracy	0.80(0.01)	0.88(0.01)	0.80(0.01)	0.94(0.01)	0.80(0.01)	0.80(0.01)	0.84(0.03)
RT							
speed	0.56(0.00)***	0.50(0.00)***	0.49(0.00)***	0.48(0.00)***	0.45(0.00)***	0.56(0.00)***	0.51(0.01)***
accuracy	0.83(0.01)	0.82(0.01)	0.70(0.01)	1.89(0.04)	0.98(0.01)	0.95(0.01)	1.02(0.07)
CONF							
speed	0.80(0.00)***	0.88(0.00)***	0.87(0.01)***	0.88(0.01)***	0.79(0.01)***	0.86(0.01)***	0.85(0.02)***
accuracy	0.84(0.00)	0.91(0.00)	0.92(0.00)	0.95(0.00)	0.85(0.01)	0.89(0.00)	0.89(0.02)
RT2							
speed	0.57(0.01)	1.08(0.02)	0.38(0.00)***	1.02(0.02)	0.70(0.01)	0.48(0.01)	0.70(0.08)
accuracy	0.57(0.01)	1.05(0.02)	0.41(0.00)	0.63(0.01)***	0.59(0.01)***	0.50(0.01)	0.62(0.06)*
CONF							
remain	0.82(0.00)	0.89(0.00)	0.9(0.01)	0.92(0.00)	0.81(0.01)	0.88(0.00)	0.87(0.02)
vanish	0.82(0.00)	0.89(0.00)	0.9(0.01)	0.91(0/00)	0.83(0.01)	0.87(0.00)***	0.87(0.02)
RT2							
remain	0.59(0.01)	1.27(0.02)	0.39(0.00)	0.88(0.02)	0.72(0.01)	0.52(0.0‘)	0.73(0.09)
vanish	0.55(0.01)***	0.87(0.01)***	0.40(0.00)	0.77(0.02)***	0.58(0.01)***	0.45(0.0‘)***	0.60(0.06)**

Note. Decision time and inter-judgment time were measured in seconds. Values in parentheses are standard errors. $*$, $**$, $* * *$ indicates the condition (speed vs. accuracy or remain vs. vanish) in which the relevant statistic was smaller with $p < .05$, .01, .001 (two-tailed) respectively, according to a *z*-test (for participants) or a meta-analysis for the group.

Table G2
Accuracy rate, mean decision time, mean confidence and mean confidence time for each participant and for the group for the early and late confidence conditions and mean confidence and mean confidence time for each participant and for the group in the remain and vanish conditions for Exp. 2.

Par	2	3	4	6	7	9	Group
Accuracy							
Early conf	0.81(0.01)	0.70(0.01)	0.79(0.01)	0.62(0.01)	0.74(0.01)	0.69(0.01)	0.73(0.03)
Late conf	0.80(0.01)	0.62(0.01)***	0.77(0.01)	0.64(0.01)	0.75(0.01)	0.67(0.01)	0.71(0.03)
RT							
Early conf	0.45(0.00)***	0.45(0.00)	0.50(0.00)	0.51(0.00)	0.52(0.00)	0.50(0.00)	0.49(0.01)
Late conf	0.47(0.00)	0.41(0.00)***	0.51(0.00)	0.5(0.00)**	0.48(0.00)***	0.44(0.00)***	0.47(0.01)
Confidence							
Early conf	0.86(0.00)	0.88(0.00)	0.90(0.00)	0.77(0.01)	0.81(0.00)	0.82(0.00)	0.84(0.02)
Late conf	0.86(0.00)	0.86(0.01)**	0.87(0.01)***	0.77(0.01)	0.8(0.00)	0.82(0.00)	0.83(0.02)***
RT2							
Early conf	0.36(0.00)	0.29(0.00)***	0.40(0.00)	0.33(0.00)***	0.35(0.00)	0.30(0.00)	0.34(0.02)
Late conf	0.29(0.00)***	0.33(0.00)	0.29(0.00)***	0.37(0.00)	0.33(0.00)***	0.25(0.00)***	0.31(0.02)
Confidence							
Remain	0.86(0.00)	0.86(0.01)***	0.89(0.01)	0.77(0.01)	0.81(0)	0.82(0.01)	0.84(0.02)
Vanish	0.86(0.00)	0.88(0.00)	0.88(0.00)*	0.77(0.01)	0.80(0)	0.83(0)	0.84(0.02)
RT2							
Remain	0.34(0.00)	0.31(0.00)	0.35(0.00)	0.36(0.00)	0.34(0.00)	0.28(0.00)	0.33(0.01)
Vanish	0.32(0.00)***	0.31(0.00)*	0.35(0.00)	0.34(0.00)***	0.34(0.00)	0.28(0.00)	0.32(0.01)*

Note. Decision time and inter-judgment time were measured in seconds. Values in parentheses are standard errors. $*$, $**$, $* * *$ indicates the condition (speed vs. accuracy or remain vs. vanish) in which the relevant statistic was smaller with $p < .05$, .01, .001 (two-tailed) respectively, according to a *z*-test (for participants) or a meta-analysis for the group.

Table G3
Accuracy rate, mean decision time, mean confidence and mean inter-judgment time for each participant and for the group for the speed and accuracy conditions of Exp. 3.

Par	Accuracy		RT		Confidence		RT2	
	Speed	Accuracy	Speed	Accuracy	Speed	Accuracy	Speed	Accuracy
2	0.82(0.02)**	0.88(0.01)	0.51(0.00)***	0.64(0.00)	0.88(0.01)**	0.91(0.01)	0.48(0.00)*	0.49(0.00)
4	0.81(0.02)***	0.95(0.01)	0.54(0.00)***	1.20(0.03)	0.89(0.01)***	0.94(0.00)	0.61(0.01)	0.54(0.01)***
6	0.68(0.02)***	0.82(0.02)	0.52(0.00)***	0.89(0.01)	0.81(0.01)***	0.85(0.01)	0.40(0.00)	0.38(0.00)***
7	0.84(0.02)**	0.90(0.01)	0.53(0.00)***	0.77(0.01)	0.87(0.01)	0.85(0.01)**	0.40(0.00)***	0.48(0.01)
8	0.68(0.02)	0.68(0.02)	0.58(0.00)***	0.64(0.01)	0.88(0.01)*	0.90(0.01)	0.56(0.01)	0.56(0.01)
11	0.76(0.02)***	0.85(0.02)	0.68(0.01)***	1.06(0.02)	0.96(0.00)*	0.97(0.00)	0.34(0.01)***	0.38(0.01)
12	0.68(0.02)***	0.89(0.01)	0.43(0.00)***	0.93(0.02)	0.9(0.01)***	0.96(0.00)	0.53(0.01)	0.45(0.01)***
15	0.73(0.02)***	0.86(0.01)	0.48(0.00)***	1.31(0.02)	0.71(0.01)***	0.79(0.01)	0.80(0.02)	0.67(0.02)***
16	0.68(0.02)***	0.80(0.02)	0.48(0.00)***	0.86(0.02)	0.82(0.01)***	0.86(0.01)	0.79(0.01)	0.72(0.01)***
Group	0.74(0.02)***	0.85(0.02)	0.53(0.02)***	0.92(0.06)	0.86(0.02)***	0.89(0.02)	0.54(0.04)	0.52(0.03)

Note. Decision time and inter-judgment time were measured in seconds. Values in parentheses are standard errors. $*$, $**$, $***$ indicates the condition (speed or accuracy) in which the relevant statistic was smaller with $p < .05$, .01, .001 (two-tailed) respectively, according to a z-test (for participants) or a meta-analysis for the group.

Appendix H. RT2 in the response entry task

We analyzed the RT2 data from the response entry task. (See detailed methods in Appendix C.1.3.) Fig. H1 displays the average RT2 across participants in our three experiments. Examining the contrast between each pair of confidence levels revealed that on average, the '100' responses were faster than all other responses, and the '60' response was slower than all other response except for the '80' responses (all contrasts were conducted with a meta-analysis and significance was determined according to a two-sided $p < .05$, planed comparisons). Note that unlike the experimental tasks, correct responding was enforced and RT2 was measured until the correct response was provided (participants were instructed to take their time to familiarize with the second stage response mappings). Still, this raises the possibility that there might have been differences in motor production times across the difference confidence responses, in our experimental tasks.

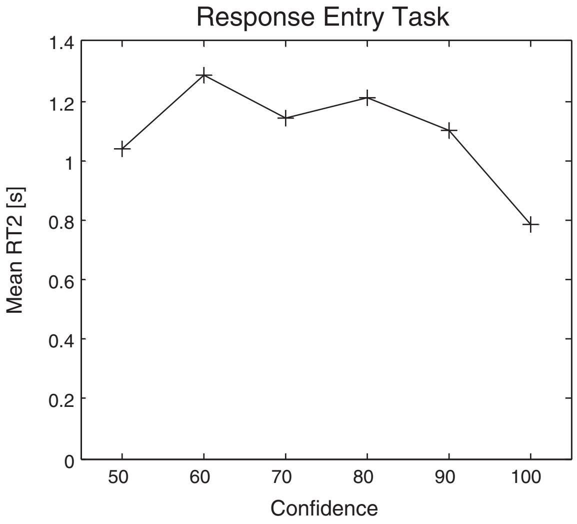

Fig. H1. Mean RT2 in the response entry task across individuals from Exp. 1–3.

We thus reexamined the RT2 correlations (Hurdle 8) in our experiments. First, the Gamma correlation between confidence and RT2 in the response entry task was negative ($\Gamma = -.20, z = -6.36, p < .001$). Importantly, however, this correlation was weaker than the same Γ correlation in either Exp. 1 ($\Gamma = -.43$) or Exp. 3 ($\Gamma = -.53; p < .001$ for both contrasts). Second, we repeated the analysis reported in Section 3.1.1, but for a 'corrected RT2', which was obtained as follows. Denote the confidence of a trial by *conf*. The corrected RT2 for each trial was calculated by subtracting from RT2, the mean RT2 across all trials with confidence *conf* (for the same participant) and adding the mean RT2 across all trials (for the participant). Notably, this correction maintains the same overall mean RT2 for each participant but obliterates any differences in mean RT2 across different confidence levels. Analyses based on the corrected RT2 across both Exp. 1 and 3 revealed that (the corrected) RT2 was still faster for correct than for error choice ($\Delta M(\text{correct} - \text{error}) = -0.11$ s,$z = -3.31, p < .001$), that the Γ correlation between RT2 and stimulus discriminability was still negative ($\Gamma = -.02, z = -1.81, p(\text{one sided}) < .05$) and that Pearson's correlation between RT2 and RT was positive ($r = .05, z = 2.37, p < .05$). These analyses suggest that the RT2 correlations remain even after controlling for possible differences in motor times across confidence responses.

Next, we used the corrected RT2 measure to reexamine Hurdle 10, the interaction between discriminability and choice-correctness on RT2. Repeating the analysis in Section 3.1.2, we found that the CORRECT$*$DISC interaction was not significant ($b = 0.00, z = 0.31, p = .75$). Thus, the control analysis did not rule out the possibility that this effect was due to differences in motor production times. Note however, that our control is highly strict in that it controls for any non-motor 'true' difference in RT2 between confidence levels. Note also that we included in the empirical manifold (Table 1) only hurdles that were replicated in the line-length task of Pleskac and Busemeyer (2010), which used a different confidence response protocol.

References

Ahn, W. Y., Busemeyer, J. R., Wagenmakers, E. J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32*(8), 1376–1402.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., et al (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*, 130–147.

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

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. (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. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology, 55*, 195–206.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1–3.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153–178.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*(1), 171–189.

Deneve, S. (2012). Making decisions with unknown sensory reliability. *Frontiers in Neuroscience, 6*.

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models*. Chapman and Hall/CRC, Taylor & Francis Group.

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.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience, 32*(11), 3612–3628.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of Acoustical Society of America, 31*, 768–773.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*(4), 843–876.

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology, 56*, 1–105.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535–574.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–769.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience, 35*(6), 2476–2484.

Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *The Quarterly Journal of Experimental Psychology, 62*(2), 209–218.

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology, 34*, 1–53.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329–1342.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759–764.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? Perception 36 ECVP abstract supplement.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*(1), 80–113.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. New York, NY: Academic Press.

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* (pp. 306–334). Cambridge, England: Cambridge University Press.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. New York, NY: Erlbaum.

McCullagh, P., & Nelder, J. A. (1990). *Generalized linear models*. New York: Chapman & Hall.

Moran, R. (2014). Optimal decision making in heterogeneous and biased environments. *Psychonomic Bulletin & Review*. http://dx.doi.org/10.3758/s13423-014-0669-3.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–141.

Peirce, C. S. (1877). Illustrations of the logic of science: The probability of induction. *The Popular Science Monthly, 12*, 705–718.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences, 3*, 73–83.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review, 10*, 177–183.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review, 80*, 53–68.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901.

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology, 71*(2), 264–272.

Rabbitt, P. (2002). Consciousness is slower than you think. *The Quarterly Journal of Experimental Psychology: Section A, 55*(4), 1081–1092.

Rabbitt, P., & Vyas, S. (1981). Processing a display even after you make a response to it. How perceptual errors can be corrected. *The Quarterly Journal of Experimental Psychology, 33*(3), 223–239.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory & Cognition, 40*, 1226–1243.

Ratcliff, R. (1978). Theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.

Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two choice reaction time. *Psychological Review, 111*, 333–367.

Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*, 70–94.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59–83.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*(3), 697–719.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*, 438–481.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature, 461*(7261), 263–266.

Shadish, W. R., & Haddock, K. C. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York, NY: Russell Sage Foundation.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18*(4), 186–193.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1–29.

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: A new perspective. *Nature Reviews Neuroscience, 8*, 872–883.

Stael von Holstein, C. (1970). Measurement of subjective probability. *Acta Psychologica, 34*, 146–159.

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review, 120*(1), 1–38.

Thura, D., Beauregard-Racine, J., Fradet, C. W., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology, 108*(11), 2912–2930.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York, NY: Cambridge University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology, 45*(6), 704–715.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*, 1011–1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*, 61–72.

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1147–1166.

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the seventeenth annual meeting of the International Society for Psychophysics* (pp. 148–153). Lengerich, Germany: Pabst.

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica, 50*(2), 179–197.

Volkman, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin, 31*, 672–673.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics, 19*(3), 326–339.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10*, 243–268.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting, 4*, 61–73.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310–1321.

Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica, 98*(1), 103–115.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience, 6*, 1–10.