Behavior Research Methods (2019) 51:40–60
https://doi.org/10.3758/s13428-018-1076-x

Combining speed and accuracy to control for speed-accuracy trade-offs(?)

Heinrich René Liesefeld[1] · Markus Janczyk[2]

Published online: 18 July 2018

Abstract

In psychological experiments, participants are typically instructed to respond as fast as possible without sacrificing accuracy. How they interpret this instruction and, consequently, which speed–accuracy trade-off they choose might vary between experiments, between participants, and between conditions. Consequently, experimental effects can appear unpredictably in either RTs or error rates (i.e., accuracy). Even more problematic, spurious effects might emerge that are actually due only to differential speed–accuracy trade-offs. An often-suggested solution is the inverse efficiency score (IES; Townsend & Ashby, 1983), which combines speed and accuracy into a single score. Alternatives are the rate-correct score (RCS; Woltz & Was, 2006) and the linear-integrated speed–accuracy score (LISAS; Vandierendonck, 2017, 2018). We report analyses on simulated data generated with the standard diffusion model (Ratcliff, 1978) showing that IES, RCS, and LISAS put unequal weights on speed and accuracy, depending on the accuracy level, and that these measures are actually very sensitive to speed–accuracy trade-offs. These findings stand in contrast to a fourth alternative, the balanced integration score (BIS; Liesefeld, Fu, & Zimmer, 2015), which was devised to integrate speed and accuracy with equal weights. Although all of the measures maintain "real" effects, only BIS is relatively insensitive to speed–accuracy trade-offs.

Keywords Speed–accuracy trade-off · Integration of errors and RTs · Integrated scoring · Task instructions · Performance strategies · Methods in experimental psychology

The ideal outcome of a behavioral experiment in many fields of experimental psychology is the predicted effect in the dependent measure of interest—usually either response times (RTs) or proportions of correct responses (PCs)—and no effect between conditions in the respective other measure (or at least an effect in the same direction). This outcome is ideal for two reasons: (a) it is easy to interpret, and (b) with just one outcome measure on which to test a hypothesis, there is no need to correct for multiple testing.

Unfortunately, it is sometimes not predictable whether participants will focus more on doing the task right or on doing it fast (i.e., which point on the speed–accuracy trade-off [SAT] continuum they will choose; for reviews, see, e.g., Heitz, 2014; Luce, 1986; Pachella, 1974; Sanders, 1998). Indeed, SATs can vary unpredictably within and across participants (e.g., Dutilh et al., 2012; Gueugneau, Pozzo, Darlot, & Papaxanthis, 2017; Liesefeld, Fu, & Zimmer, 2015), participants can adapt their SATs trial-wise and at will (e.g., Paoletti, Weaver, Braun, & van Zoest, 2015; Reuss, Kiesel, & Kunde, 2015; Voss, Rothermund, & Voss, 2004; Wickelgren, 1977), and they might even change their SATs systematically between conditions. Sometimes, of course, shifts in SATs are the phenomenon of interest (e.g., as an account of post-error slowing; Laming, 1968; see also Botvinick, Braver, Barch, Carter, & Cohen, 2001; Thura, Guberman, & Cisek, 2017), but usually SAT shifts hinder the goals of a study, probably confounding the investigated effect.

Even when the dependent measure of interest is set a priori for theoretical reasons, researchers still routinely check

✉ Heinrich René Liesefeld
Heinrich.Liesefeld@psy.lmu.de

[1] Department Psychologie, and Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Leopoldstr. 13, D-80802 Munich, Germany

[2] Department of Psychology, Eberhard Karls University of Tübingen, Tübingen, Germany

whether the respective other measure points in the same direction (e.g., shorter RTs and more correct responses/higher PC). This is because RTs and PCs pointing in opposite directions (i.e., shorter RTs and fewer correct responses/lower PC) would indicate that the observed effects may (in part) be due to SATs instead of to “real” effects. Furthermore, the (sometimes subjective) decision to interpret the one or the other measure would yield conflicting conclusions regarding the direction of the effect. Therefore, ignoring either would obviously be wrong.

These intricate situations might be resolved by using a combination of RTs and PCs, such that (a) possible SAT contributions are cancelled (or at least dramatically attenuated), while (b) ”real” effects remain in the data. There are suggestions of such combined measures in the literature, and they have been employed when the data show signs of SATs (e.g., Kunde, Pfister, & Janczyk, 2012; Kristjánsson, 2016). The goal of this article is to test whether these available measures conform to the two, just mentioned, criteria, and to arrive at well-founded recommendations as to whether or not to use them. For this purpose, we simulate pure SATs and “real” effects with the diffusion model (Ratcliff, 1978) and determine the degree to which each measure attenuates SATs and maintains “real” effects. Additionally, we will formally introduce and test a new alternative that was conceived by Liesefeld et al. (2015).

Measures to combine speed (RT) and accuracy (PC)

In the following section we describe four measures in the literature that combine RT and accuracy into a single performance measure. To make perfectly clear how the different measures are calculated, an example data set and the respective calculations are given in Table 1.

Inverse efficiency score (IES)

The most often suggested combined measure is the inverse efficiency score (IES; Townsend & Ashby, 1983), which is typically defined as mean correct RTs divided by PCs (Akhtar & Enns, 1989; Bruyer & Brysbaert, 2011):

$$IES_{i,j} = \frac{\overline{RT_{i,j}}}{PC_{i,j}} \tag{1}$$

where $\overline{RT_{i,j}}$ is participant i’s mean RT on correct-response trials in condition j and $PC_{i,j}$ is participant i’s proportion of correct responses in condition j. Although most (if not all) studies using or evaluating IES have included only correct RTs, all RTs (including those from error trials) seem to be taken into account according to the original proposal (Townsend & Ashby, 1983, p. 204). We will, however, evaluate the version without incorrect RTs, because this is the one typically reported in empirical studies. This measure can be interpreted as “the average energy consumed by the system over trials” (Townsend & Ashby, 1983, p. 204).

Rate-correct score (RCS)

An alternative suggestion is the rate-correct score (RCS; Woltz & Was, 2006):

$$RCS_{i,j} = \frac{NC_{i,j}}{\sum_{k=1}^{n_{i,j}} RT_{i,j,k}} \tag{2}$$

where $NC_{i,j}$ is participant i’s number of correct responses in condition j and the denominator reflects the total time participant i spent on trials in condition j (in other words, the sum of RTs across all $n_{i,j}$ trials of participant i in condition j). RCS “can be interpreted directly as number of correct responses per unit time” (Woltz & Was, 2006, p. 673). As is evident from a comparison of Eqs. 1 and 2 and as detailed in Appendix A, RCS is similar or even identical to the inverse of IES (see also Vandierendonck, 2018), and therefore no strong differences between the two are to be expected. Both measures also bear some similarity to *reward rate* (e.g., Balci et al., 2011; Gold & Shadlen, 2002), which instead of mere RTs takes into account all of the time between two responses (i.e., also the time between a response and the next trial).

Linear integrated speed–accuracy score (LISAS)

Vandierendonck (2017) suggested the linear integrated speed–accuracy score (LISAS), which is defined as

$$LISAS_{i,j} = \overline{RT_{i,j}} + \frac{S_{RT_{i,j}}}{S_{PE_{i,j}}} \cdot PE_{i,j} \tag{3}$$

where $\overline{RT_{i,j}}$ is participant i’s mean RT on correct-response trials in condition j and $PE_{i,j}$ is participant i’s proportion of errors $(1 - PC)$ in condition j. Note that both the mean RT and S_{RT} include only correct trials and that $S_{RT_{i,j}}$ and $S_{PE_{i,j}}$ are the across-trial sample standard deviations of participant i in condition j (i.e., with n in the denominator, and not the unbiased

Table 1 Example calculations for the inverse efficiency score (IES; Eq. 1), rate-correct score (RCS; Eq. 2), linear integrated speed–accuracy score (LISAS; Eq. 3), and balanced integration score (BIS; Eq. 4)

Subject	$\overline{RT}_c$	$\sum RT$	PC	S_{RT}	z_{PC}	z_{RT}	IES	RCS	LISAS	BIS
Group A: v = 0.3, a = 50										
1	356.8	35,684	.71	28.83	– 1.01	– 0.96	502.5	1.99	375.22	– 0.05
2	325.1	32,511	.73	45.60	– 0.86	– 1.19	445.3	2.25	352.83	0.33
3	370.6	37,064	.67	44.46	– 1.30	– 0.86	553.1	1.81	401.80	– 0.45
4	362.8	36,280	.73	30.58	– 0.86	– 0.92	497.0	2.01	381.40	0.06
5	348.5	34,853	.72	36.95	– 0.93	– 1.02	484.0	2.07	371.54	0.09
$\overline{\boldsymbol{A}}$	**352.8**	**35,278**	**.71**	**37.28**	**– 0.99**	**– 0.99**	**496.4**	**2.02**	**376.56**	**0.00**
Group B: v = 0.3, a = 200										
6	640.3	55,998	.97	269.04	0.93	1.10	660.1	1.51	687.61	– 0.18
7	634.2	60,782	.99	226.07	1.08	1.06	640.6	1.56	656.92	0.02
8	650.7	61,318	.99	274.28	1.08	1.18	657.3	1.52	678.27	– 0.10
9	584.5	60,329	.97	222.33	0.93	0.70	602.6	1.66	623.60	0.23
10	610.0	65,284	.97	189.85	0.93	0.89	628.9	1.59	643.39	0.04
$\overline{\boldsymbol{B}}$	**623.9**	**60,742**	**.98**	**236.31**	**0.99**	**0.99**	**637.9**	**1.57**	**657.96**	**0.00**

The data are from the first five simulated subjects with the respective parameters. Means are calculated per column. z standardization (as an intermediate step for calculating BIS) is done across groups; that is, ten values contribute to the mean and the standard deviation in this example). Therefore, $\overline{A}$ and $\overline{B}$ (as well as all ten individual values) of z_{PC} and z_{RT} add up to exactly 0. RT = reaction time, c = correct, PC = proportion correct, S_x = sample standard deviation of x; z_x = standardized x—that is, $\overline{z_x}$ = 0; S_{z_x} = 1. Note also the following: number of trials n = 100; proportion error (PE) = 1 – PC; $S_{PE} = \sqrt{PE \cdot (1-PE)}$; $NC = PC \cdot n$, if no time-outs occur (as in this sample data set; see note 7).

population estimate with n – 1 in the denominator).[1]

Whereas IES and RCS were constructed to have a straightforward interpretation (viz. the energy consumed on average and the number of correct responses per second of activity), the goal of LISAS is to obtain a linear integration.

Balanced integration score (BIS)

Liesefeld et al. (2015) devised a measure that focuses on giving equal weights to RTs and PCs, which we term *balanced integration score (BIS)*.[2] It is calculated by first standardizing RTs and PCs to bring them to the same scale and then subtracting one standardized score from the other:[3]

$$BIS_{i,j} = z_{PC_{i,j}} - z_{\overline{RT}_{i,j}} \tag{4}$$

with $z_{x_{i,j}} = \frac{x_{i,j}-\overline{x}}{S_x}$.

Thus, BIS is the difference in standardized mean correct RTs and PCs. The mean and sample standard deviation must be calculated over all cells that contribute relevant variance. If values were, for example, standardized per condition, all conditions would have the same mean value of 0, thus eliminating any effects. Thus, $\overline{RT}$, PC, S_{RT}, and S_{PC} are typically calculated across all observed mean RTs and all PCs from the analyzed experiment (including all subjects and all conditions).

Assessing the combined measures

Quality criteria for combined measures

As we stated above, we believe that a combined measure should ideally (a) cancel out SATs and (b) maintain "real" effects. Other criteria, however, have been put forward and are discussed in the following section.

Bruyer and Brysbaert (2011) examined how well IES worked in several empirical data sets in comparison to RTs or PCs alone. Their criterion was whether IES "clarifies matters" (p. 9). In particular, they checked (a) whether effects in RTs or PCs were preserved in IES, (b) whether new effects emerged in IES, and (c) whether IES would yield a more orderly data pattern than its constituents. After comparing the result patterns in RTs, PCs, and IES in data from several studies, they plaintively concluded: "It looks pretty much like every

[1] personal communication, André Vandierendonck, August 25, 2017

[2] To avoid any misunderstandings, we consider balanced integration to mean that both constituents (RTs and PCs) contribute equally to the combined measure. We thank André Vandierendonck for pointing out that his (2017) use of the term "balance" deviates from ours. We will come back to this point later, in the section on Balanced Integration of Measures; see also Appendix B for more information.

[3] Liesefeld et al. (2015) actually subtracted z_{PC} from z_{RT} and applied a linear transformation to bring it to the scale of RTs, so that the results could be interpreted as RTs in a hypothetical, error-free task, similar to IES. Switching the subtraction order has the same effect as multiplying BIS by – 1. Thus, the present definition of BIS is a linear transformation of the Liesefeld et al. (2015) measure and therefore does not differ in its statistical properties. We decided to adapt the definition here in order to obtain an interpretation in terms of performance above or below average, which makes it easier to discuss some of this measure's properties.

It should also be noted that Paas and Van Merriënboer (1993) developed a strikingly similar measure for the combination of mental workload—measured by rating scales, psychophysiological markers, or dual-task techniques—on the one hand, and task performance—measured by speed, accuracy, or test scores—on the other. Their work was pointed out to us in response to a conference presentation of the present study.

type of change is possible with the introduction of IES" (p. 9). This indicates that employing convenience samples of empirical data sets is not well suited to examining the suitability of a given combined measure. A promising and flexible alternative is creating artificially generated data, in which the relative contributions of SATs and "real" effects can be known.

Vandierendonck (2017) created such artificial data in order to examine properties of IES, RCS, and LISAS. He simulated pure effects on RTs or PEs and effects in both variables in the same or opposing directions and found that all three combined measures performed quite well in recovering these effects, with RCS and LISAS working better than IES.[4] Although the effects in opposing directions can be conceived as SATs,[5] his evaluation was focused on whether effects are maintained or amplified in a given combination of speed and accuracy. In contrast to this standpoint, we believe that the most important property of a combined measure should be its *insensitivity* to SATs, yielding ideally no or considerably reduced effects when the effects on RTs and PEs point in opposing directions. Only then would using the combined measure avoid interpreting spurious effects. Indeed, when patterns of RTs and PEs were opposing, the combined measures examined by Vandierendonck somewhat reduced the effects, but these SAT effects were still alarmingly large (see his Table 11 and Fig. 3d).

Interestingly, whereas Bruyer and Brysbaert (2011) advised against the use of IES, Vandierendonck's (2017) conclusion was more favorable for this particular measure (see also Vandierendonck, 2018). Both Vandierendonck (2017) and Bruyer and Brysbaert concluded that one should always examine RTs and PCs/PEs separately in addition to any combined measure. However, if such an examination is always indeed needed, it appears questionable to us whether anything is gained by examining the combined measure at all, or whether the additional analysis simply enlarges the Results section (and the alpha error).

Simulating SAT and "real" effects with a diffusion model

As was exemplified above, attempts to assess the behavior of the combined measures with empirical data have the disadvantage that the degree of SATs and the impacts of other variables on RTs and PCs (such as "real" effects) are unknown. To gain pure SATs and pure effects, we based our assessment of combined speed–accuracy measures on data artificially generated by the well-established diffusion model (Ratcliff, 1978; Ratcliff, Smith, Brown, & McKoon, 2016; Ulrich, Schröter, Leuthold, & Birngruber, 2015; Vandekerckhove & Tuerlinckx, 2007; Voss & Voss, 2007; Wagenmakers, van der Maas, & Grasman, 2007). In particular, we modeled SATs and "real" effects on RTs and PCs by variation of threshold separation and drift rate, respectively (see below for details on the model and the parameters). Furthermore, as we will demonstrate below (see the Balanced Integration of Measures section), the relative weighting of RTs and PCs depends on the accuracy level, and the different combined measures' effectiveness in canceling SATs and maintaining "real" effects varies differentially across accuracy ranges. Therefore, we densely sampled accuracies ranging from pure guessing (50%) to virtually perfect performance (100%) instead of picking only a few points from this spectrum.

Diffusion models (Ratcliff, 1978) are a class of random-walk models that have been successfully applied to modeling decision behavior and predicting RT distributions and PCs in a variety of paradigms and fields of research (for recent reviews, see Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff et al., 2016; Voss, Voss, & Lerche, 2015; Wagenmakers, 2009). The basic idea of these models (see Fig. 1 for an illustration) is that a diffusion process starts at a specified point and noisily accumulates evidence with a certain drift rate v (reflecting the strength of evidence) until one of two thresholds is exceeded. In one interpretation of the diffusion model, the upper threshold a is associated with a correct response and the lower threshold 0 represents an erroneous

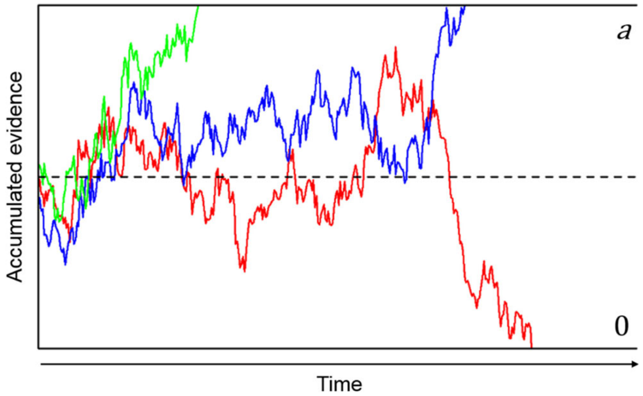


Fig. 1 Illustration of a simple diffusion model. In this example, the upper threshold a is associated with correct responses and the lower threshold, at 0, is associated with erroneous responses. Without any bias, evidence accumulation starts at $a/2$. On each time step, a fixed drift rate and random noise are added to the evidence. The green line represents a trial with a (relatively) high drift rate (hitting the threshold a rather early), whereas the blue line represents a trial with a (relatively) small drift rate (hitting the threshold a rather late). Due to random noise, it is also possible that the accumulated evidence will hit the lower threshold at 0, thus representing an erroneous response.

[4] Vandierendonck (2017) also tested several versions of a binning score (Draheim, Hicks, & Engle, 2016; Hughes, Linck, Bowles, Koeth, & Bunting, 2014), which, as Vandierendonck (2017) makes very clear, has several highly undesirable characteristics that make such measures uninteresting as integrations of RTs and PCs. Furthermore, even considerably adapted versions of the binning score failed Vandierendonck's (2017) tests. We will therefore not consider these binning scores here.

[5] Recall that any shift along the SAT continuum would result in opposing effects on RTs and PEs. Please also note that Vandierendonck's conception of SATs differs from this continuum view, and that he aimed to simulate SATs independently from the (opposing) effects on RTs and PEs. However, only the opposing effects can be interpreted as SATs, as we examine them here.

response. Although the exact starting point can theoretically vary between 0 and a, most often it is set to $a/2$, thus without any bias toward one or the other threshold. Once the accumulated evidence exceeds one of the thresholds, the correct or wrong decision is made (and a response is given, in many experimental settings). Although the drift rate v drives the diffusion process in the direction of a, noise can cause the diffusion process to reach 0, and thus an error results. Typically, the time from accumulation onset until a threshold is reached is considered the decision time, and additional processes such as perception, non-decision-related cognitive processes, and motor execution are captured by a nondecisional constant t_0. For a more complete description of parameters, see, for example, Voss et al. (2015) or Ratcliff et al. (2016).

Of particular importance for the present purposes, larger values of a will lead to longer RTs (since it takes longer to reach a threshold) but at the same time increase the likelihood of correct responses. In other words, varying a across simulated conditions can be used to induce a "pure" shift along the SAT continuum without any confound with "real" effects that would reflect between-condition differences in task difficulty (and that can be simulated by varying v or t_0).

In our simulations, a Wiener diffusion process was used (e.g., Ratcliff, 1978; Ulrich et al., 2015) in which noise is modeled as Brownian motion to which a (linear) drift function is added. We fixed the starting point at $a/2$. For the present purposes, we varied threshold separation a in order to induce an SAT, with $a \in \{5, 10, 15, \ldots, 290, 295, 300\}$. We repeated the simulation with six different drift rates, $v \in \{0.20, 0.22, \ldots, 0.30\}$, as an operationalization of "real" effects.[6] For each simulated cell, an individual value for the drift rate was sampled from a normal distribution with mean v and a standard deviation of 0.01, to induce error variance. Further error variance was induced by sampling the nondecision time t_0 for each simulated cell from a normal distribution with mean 300 and a standard deviation of 20. Within each of these 60 (threshold separation a) × 6 (drift rate v) = 360 combinations, 100 trials were simulated as a Wiener diffusion process (i.e., Brownian motion with a positive drift):

$$X(t) = B(t)\sigma + vt,$$

where $B(t)$ represents the Brownian motion at time t. In our simulation, we used $\sigma = 4$. The simulated data were organized in such a way that the data could be conceived as 100 independent between-subjects experiments with $n = 20$ participants in each cell and random variation in each participant's drift rate and nondecision time. For each simulated participant and each of the combinations of a and v, those variables were computed that were necessary in order to calculate IES, RCS, LISAS, and BIS in a subsequent step.[7] The full data set can be retrieved from https://osf.io/pyshv/

In a nutshell, we orthogonally manipulated threshold separation and drift rate and added some error variance. The variation in threshold separation emulates a variation in SATs, and the variation in drift rate emulates a variation in "real" effects. Given the goal to examine how well the various combined measures cancel SATs, the main focus in the following analyses was on the threshold separation. The drift rate was varied orthogonally for two purposes: (a) to check whether the threshold-dependent behavior of all measures generalizes across various drift rates and (b) to subsequently test how well the combined measures maintain "real" effects. Error variance was added so as to increase the comparability to real data and to avoid that even negligible effects become significant.

Effects on speed and accuracy

Figure 2 visualizes RTs and PCs of the simulated data as a function of threshold separation a and drift rate v. As desired, RTs were in a range typically observed in cognitive experiments, and importantly, responses speeded up with increasing v and slowed down with increasing a (Fig. 2, left panel). At the same time, responses were more accurate with higher values for drift rate v and with larger threshold separations a (Fig. 2, right panel). Also, PCs ranged from close to pure guessing (50%) to virtually perfect performance (100%). Thus, the data cover the whole spectrum of typically observed effects, and our variation of threshold separation a implements an SAT:

[6] Variation in threshold separation is the standard approach to simulating SATs with the diffusion model (e.g., Dutihl et al., 2012; Lerche & Voss, in press): An increase in threshold separation yields *slower* and more accurate responses (reflecting a more conservative response criterion). Drift rate was chosen to simulate "real" effects because, like threshold separation, it affects both RTs and PCs. In contrast to threshold separation, however, an increase in drift rate yields *faster* and more accurate responses (reflecting improved performance). These features make drift rate particularly interesting, and many studies have reported effects on drift rate (e.g., Germar, Schlemmer, Krug, Voss, & Mojzisch, 2014; Janczyk & Lerche, in press; Janczyk, Mittelstädt, & Wienrich, 2018; Ratcliff, Thapar, & McKoon, 2011; Schubert, Hagemann, Voss, Schankin, & Bergmann, 2015; Voss, Rothermund, & Brandtstädter, 2008; Voss et al., 2004). Although improved performance is sometimes captured by nondecision time instead of drift rate (e.g., Schmitz & Voss, 2012), nondecision time was less suited to simulating "real" effects here. This is because nondecision time influences only RTs, and therefore any manipulation of nondecision time is, by definition, best captured by pure RTs and not by any combined measure.

[7] Because typical experiments have response deadlines, a decision process taking longer than 3,200 ms was considered a timeout (adding the nondecision time of 300 ms, on average, this would correspond to a response deadline around 3,500 ms). PC was calculated on the basis of trials without timeouts. Such timeouts occurred only rarely, and a minimum of 97 (out of 100) trials were included in all respective cells (the average number was 99.99 trials).

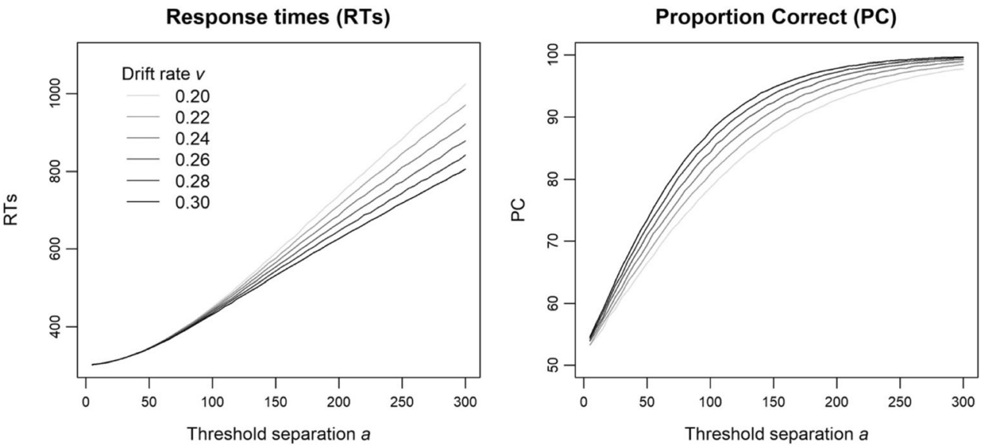


Fig. 2 RTs and PCs as a function of threshold separation *a* and drift rate *v*. For a given drift rate, both RTs and PCs increase with increasing threshold separation. This is the pattern expected in the case of a speed–accuracy trade-off.

the higher the threshold separation, the slower but more accurate the responses. The data set is complemented by a "real" effect as induced by the drift-rate manipulation. Hence, this simulated data set is well-suited for examining the properties of the combined measures.

Balanced integration of measures

In the absence of any good reason to amplify the influence on the combined measure of either RTs or PCs, it appears reasonable to give equal weight to both constituents. An operational definition of such balanced integration is that the combined measure shares as much variance with RTs as with PCs. This is achieved in BIS by design (see Appendix B). To assess the relative contributions of RTs and PCs to the other combined measures, we calculated for each combined measure M ($M \in$ {IES, RCS, LISAS}) the index I_M, as

$$I_M = \frac{r^2_{RT,M}}{r^2_{PC,M}} \tag{5}$$

where $r^2_{RT,M}$ is the squared Pearson correlation of M with RTs and $r^2_{PC,M}$ is the squared Pearson correlation of M with PCs. If RTs and PCs contribute to the same degree to the measure M, the index should take a value of $I_M \approx 1$. In contrast, $I_M < 1$ means a dominance of PC, and $I_M > 1$ means a dominance of RT, with the extreme cases of $I_M = 0$ (exclusively influenced by PC) and $I_M = \infty$ (exclusively influenced by RT). For each measure, I_M was calculated for each combination of drift rate v and threshold separation a. In particular, correlations were calculated separately for each of the 100 experiments, Fisher z-transformed, averaged across experiments, transformed back, and entered into Eq. 5.

As is evident in Fig. 3, IES, RCS, and LISAS exhibit a pattern considerably deviating from a balanced integration of RTs and PCs: For RCS and IES, the balance changes depending on a: With smaller threshold separations, the influence of RTs is

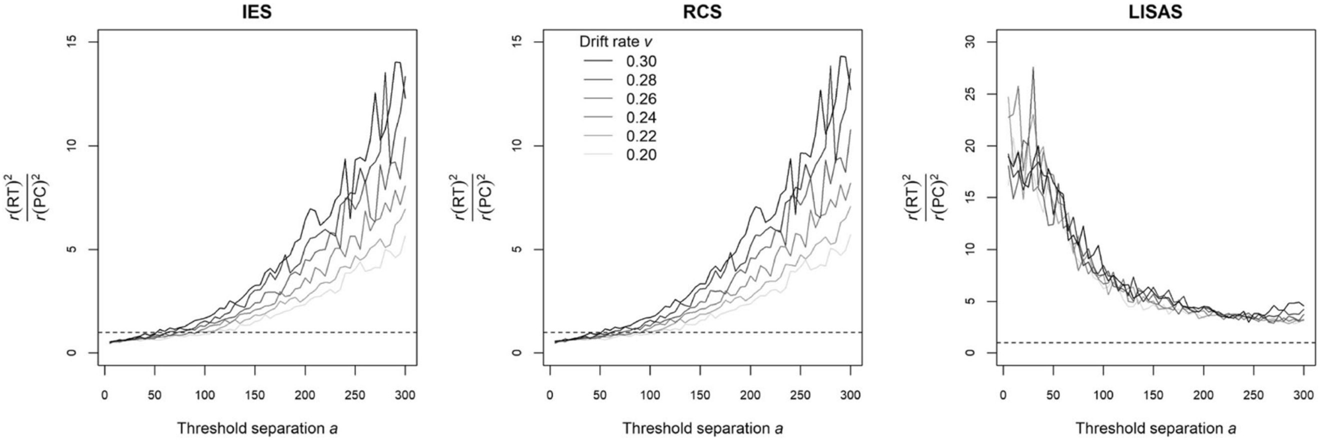


Fig. 3 Relative contributions of RTs and PCs to three of the combined measures, as a function of threshold separation *a* and drift rate *v*. The point of balanced integration is indicated by the dotted line at $I_M = 1$. Values below 1 indicate a predominance of accuracy, and values above 1 indicate a predominance of RTs. BIS is not plotted, because it integrates in a balanced manner by design, so that for all cells $I_{BIS} = 1$ (see Appendix B). Note that the *y*-axis scaling for LISAS differs from the other two measures because LISAS overemphasizes RTs to a much larger degree, and the overemphasis on PCs for low thresholds in IES and RCS would be disguised with an adapted scaling.

negligible, whereas at larger separations RTs, take over so as to affect the measures predominantly. Furthermore, the influence of RTs increases with increasing drift rates v for medium to large threshold separations a. For LISAS, the pattern is different: RTs dominate at low levels of a, and the respective influences become more balanced at higher levels (yet never reaching the point of balanced integration at $I_{LISAS} = 1$).

The behavior of the three combined measures might be problematic: Not only do these measures not weight RTs and PC in a balanced manner, but the relative influences strongly depend on the accuracy level. Thus, even if it is possible to find a weighting factor so that RTs and PCs are balanced in one experimental condition, this weighting will yield unbalanced integration in other cells of the design. Thus, none of the three measures integrates RTs and PCs in a balanced way. BIS differs in this regard, and in fact $I_{BIS} = 1$ across all levels of threshold separation a and drift rate v (see Appendix B for a formal proof).

However, whether a balanced and constant (across SATs) weighting is desirable is an open question. One advantage of the unbalanced and accuracy-dependent integration in IES and RCS could be that RTs contain much more relevant information when accuracies are close to ceiling, and accuracies contain more relevant information when RTs are very fast. Thus, in the following sections, we directly test more unambiguously desirable characteristics of any combined measure: Do they cancel SATs while maintaining “real” effects?

Testing the efficiency of the combined measures to compensate for SATs

In the worst case, an experimental manipulation would only induce a pure SAT. Analyzing either RTs or PCs would then yield spurious effects and wrong conclusions; analyzing both would yield contradictory results. In Fig. 2, for example, although threshold separation does not influence task difficulty, this parameter exerts strong effects on RTs and PCs. The combined measures should—ideally—compensate for these SATs and provide the same values irrespective of the threshold separation. Figure 4 visualizes the combined measures as a function of threshold separation a and drift rate v. From simple visual inspection, it becomes clear that none of the four measures fulfills this criterion perfectly, although the effect of threshold separation (i.e., of SATs) is clearly smaller for BIS than for all competing measures.[8]

Due to the different y-axis scalings, the sizes of the SAT effect (and of the drift-rate effect) are not directly comparable between the four panels of Fig. 4. Additionally, the figures do not contain information on the statistical error variance, and one can therefore not gauge which differences are statistically significant. To address this question directly, we ran inferential statistics on different subsets of the data to report effect sizes and how often an SAT effect was statistically significant for a given measure. We will start with t tests and continue with analyses of variance (ANOVAs).

Comparisons of two conditions using *t* tests We first concentrate on situations with two conditions in which the dependent measures are assessed by t tests. Therefore, we ran two-sample t tests comparing all combinations of threshold separation a for each of the 100 experiments while keeping drift rate v constant, thus looking at pure SATs. The results are illustrated in the form of heat maps, in which each point represents one comparison of two levels of a (as designated on the x- and y-axes). In Fig. 5, color codes the resulting values for effect size d (Cohen, 1988), and in Fig. 6, it codes the proportions of significant t tests. In both figures, the upper/left half of each heat map represents the data with drift rate $v = 0.2$, and the lower/right half the data with drift rate $v = 0.3$.

As we intended through the simulation of SATs, for RTs and PCs, the d values clearly become larger, the larger the difference between the two levels of threshold separation a (with generally larger effects on RTs). The results obtained for LISAS are very similar to those for RTs, with only slightly smaller values for d. This is not surprising, given that LISAS mainly represents RTs (see Fig. 3). The d values for IES, RCS, and BIS, in contrast, are much smaller. However, IES and RCS both also exhibit an increase in d values when small to medium values for threshold separation a are compared with large values (the orange regions far off the diagonal). That these regions are not centered at the edges (as for RTs, PC, and LISAS) results from the nonmonotonic behavior of both measures in the smaller range of threshold separation a (in particular, values with $0 < a \leq 50$ do not follow the general trend of rising [IES] or falling [RCS] with increasing levels of a; see Fig. 4).

Additional, subtle patterns become more clearly visible when we consider the proportions of significant t tests, in Fig. 6 (recall that we simulated 100 experiments for each cell). First, for RTs, PCs, and LISAS, nearly all t tests are significant, with exceptions only close to the diagonal—that is, for small differences in SAT. Another effect of the above-mentioned nonmonotonic behavior of IES and RCS (Fig. 4) becomes apparent in Fig. 6, for comparisons of

[8] As we mentioned above, there is no BIS for a single cell, and we therefore had to use the following procedure to arrive at the values in Fig. 4 (lower right panel): All size 2 k permutations with replacement of the six drift rates (= 36 possible combinations) were crossed with all size 2 combinations of threshold separation without replacement (= 1,770 possible combinations). The z standardization required to calculate BIS was then done for the resulting 63,720 possibilities of two samples for each of the 100 experiments separately. Finally, the resulting BIS values were averaged for all combinations of a and v. In other words, we calculated BIS for all possible pair-wise combinations of all Drift Rate × Threshold Separation cells and estimated a unique BIS for each cell by averaging across all BIS values for the respective cell.

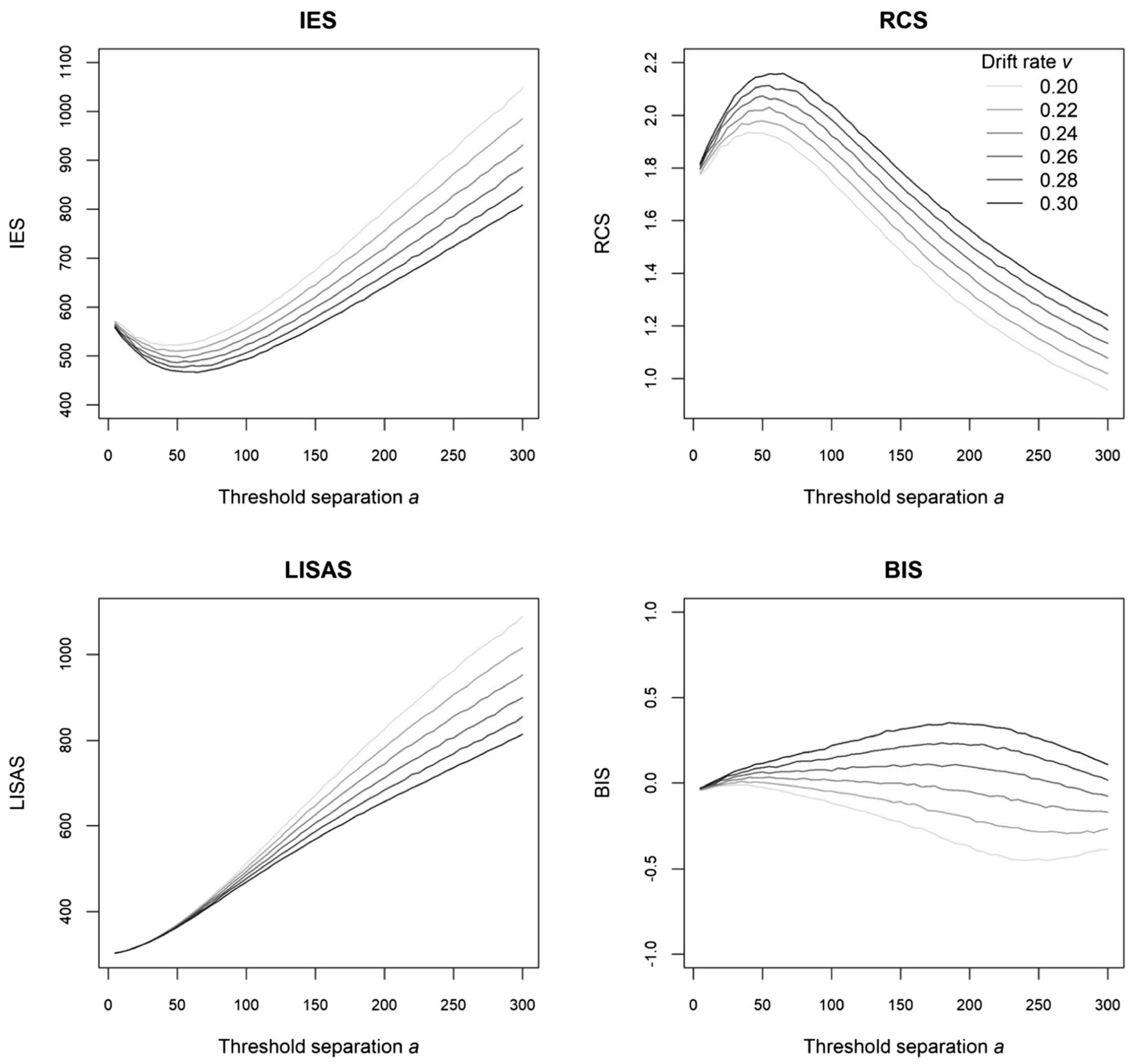


Fig. 4 IES, RCS, LISAS, and BIS as a function of threshold separation *a* (SATs) and drift rate *v* ("real" effects). All of the measures retain "real" effects (i.e., the differences between lines should be large) at reasonable levels of *a* (i.e., when there is sufficient time to accumulate evidence, so that PCs clearly deviate from guessing performance). However, effectiveness in cancelling SATs (the lines ideally should be flat) strongly differs between the various combined measures.

the smallest threshold separations to small-to-medium threshold separations: Deviating from the high incidence of significant tests in other regions, only about 20%–30% of the *t* tests are significant, yielding the yellowish stripes in the lower left corners of the graphs. This is again due to the nonmonotonic behavior of IES and RCS as a function of threshold separation, as described above and visible in Fig. 4. For BIS, large areas with no or very few significant *t* tests are apparent. Surprisingly, some tests are significant around the diagonal—that is, for small differences in SAT. Additionally, more tests are significant for the higher drift rate of $v = 0.3$, in particular when combined with high threshold separations. However, the proportion of significant tests is lower than for any competing measure for virtually all comparisons.

Comparisons of three conditions using ANOVAs The examined measures are by no means restricted to comparisons of *two* conditions (in contrast to other alternatives, such as the binning score; Draheim et al., 2016; Hughes et al., 2014). To get an impression of their behavior in more complex situations, we also considered the case of *three* conditions, as would be assessed with ANOVAs. Unfortunately, this more complex design does not allow an exhaustive examination of all possible comparisons, as was possible with pairwise comparisons. Instead, we drew three random levels (without replacement) 100 times for each experiment and drift rate *v*. Then we calculated a one-way between-subjects ANOVA on each data set. Analogous to the analyses with *t* tests above, Fig. 7 visualizes the mean effect sizes η^2 and the mean proportions of significant ANOVAs. As we expected, the ANOVA

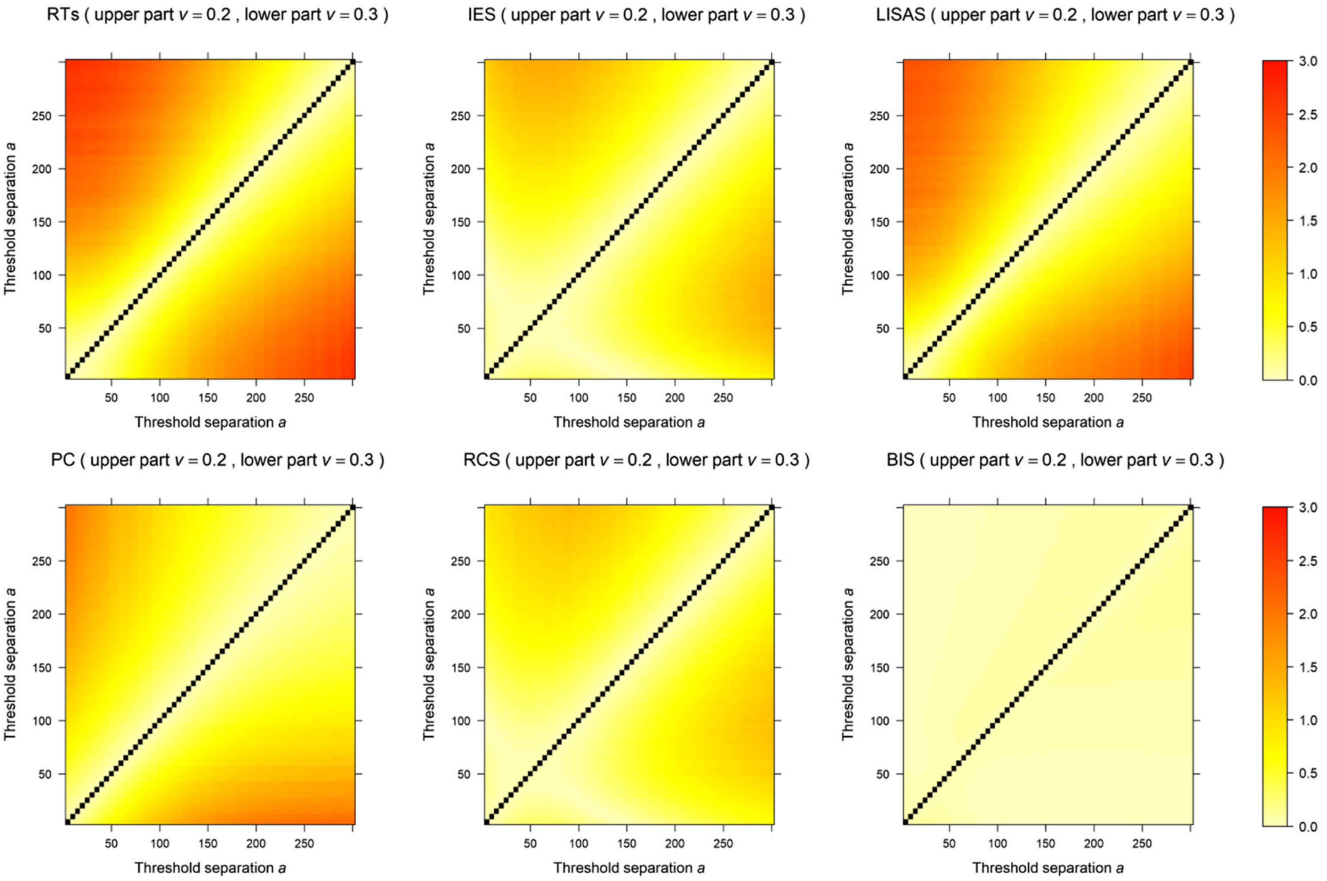


Fig. 5 Effect sizes d for pairwise comparisons of threshold separations a, reflecting SATs, at two different values of drift rate v ($v = 0.2$ and $v = 0.3$, above and below the diagonal, respectively). Each point in a panel denotes a comparison between two combinations of threshold separation (e.g., $a = 250$ vs. $a = 200$). The black diagonals indicate the absence of comparisons between a cell and itself (e.g., $a = 250$ vs. $a = 250$). Note that all measures except BIS are heavily influenced by the simulated SATs, with RTs alone and LISAS performing worst, followed by PCs. IES and RCS show lower (but still substantive, up to $d = 2$) effects of SATs. For most measures, the effect depends approximately linearly on the difference in threshold separation. IES and RCS diverge from this orderly pattern due to their highly nonlinear dependence on threshold separation, displayed in Fig. 4. The effects on BIS do not exceed $d = 0.12$.

almost always revealed high effect sizes and significant effects for RTs and PCs. However, this was also true for IES, RCS, and LISAS. For BIS, the effect sizes and proportions significant were considerably smaller. Surprisingly and in contrast to the other measures, the effect of SATs on BIS depends on the size of the "real" effect, with an increase from around 50% to 60% significant with increasing drift rate v.

Testing the efficiency of the combined measures to maintain "real" effects

One trivial explanation for the efficiency of BIS in canceling SAT effects would be that it cancels any effect. If this were the case, BIS would be useless as a combined measure. Although Fig. 4 already shows that this is unlikely, we also analyzed "real" effects, as simulated by variations in drift rate, to address this concern more formally (Figs. 8 and 9 for combinations of drift rates as assessed with t tests; Fig. 10 for averages across all combinations of three drift rates, as assessed with ANOVAs; all calculations were performed for four exemplary threshold separations with $a \in \{5, 100, 200, 300\}$). As can be seen, all combined measures nicely maintain "real" effects, and sometimes even enhance them. For very small threshold separations—where performance is close to chance (pure guessing)—variations in drift rate have no effect on any of the examined measures (see also Fig. 4). This confirms the usual recommendation not to consider data (in particular RTs) when performance is close to chance.

Discussion

In the present study, we examined the usefulness of several approaches for combining response times (RTs; speed) and proportions correct (PC; accuracy) to control for speed–accuracy trade-offs. Instead of using empirical data in which the levels of SAT are unknown, we simulated pure SATs and "real" effects without any confound, by varying threshold separation and drift

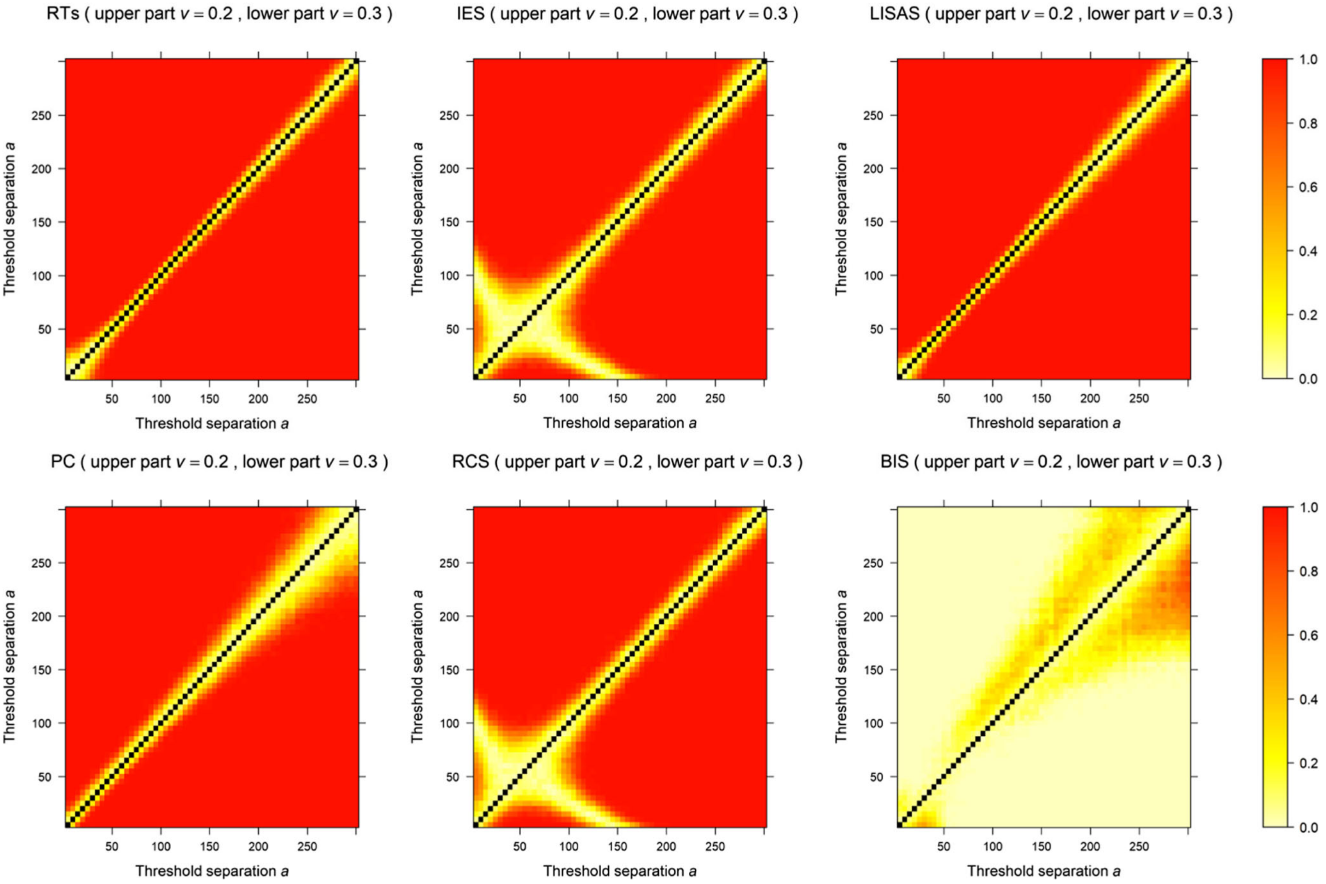


Fig. 6 Proportions of significant pairwise comparisons (two-sample *t* tests) of threshold separations *a*, reflecting SATs, at two different values of drift rate *v* ($v = 0.2$ and $v = 0.3$, above and below the diagonal, respectively). Each point in a panel denotes a comparison between two combinations of threshold separation (e.g., $a = 250$ vs. $a = 200$). The black diagonals indicate the absence of comparisons between a cell and itself (e.g., $a = 250$ vs. $a = 250$).

rate in the diffusion model (Ratcliff, 1978). Arguably, a useful combined measure should cancel any effects of differential SATs, so that it is insensitive to the specific trade-off along the SAT continuum, without markedly attenuating “real” effects. Whereas all combined measures fulfilled the latter criterion, BIS came far closer to the first criterion than the other alternatives, and it might be worth a recommendation whenever a combination of RTs and PCs is desired.

Recommendations on the use of combined measures

Advantages of combining RTs and PCs There are several potential reasons why a researcher may want to combine RTs and PCs. First, when using BIS, SATs are canceled to a large degree, thus considerably decreasing the likelihood of interpreting spurious effects that are mainly driven by SATs.

Second, combining RTs and PCs can yield a gain in statistical power in two (not necessarily mutually exclusive) situations: (a) If some participants focus more on speed and others focus more on accuracy, the effects of experimental manipulations will be distributed across RTs and PCs, and a combination of the two can potentially reconstitute the full effect (see also Hughes et al., 2014, p. 705). (b) For situations in which there is no clear theoretical reason to focus on either RTs or PCs, testing both would yield an inflation of alpha error and, thus, require an adaptation of the alpha level (i.e., with the typical level of α = .05, only tests with p < .025 could safely be considered significant). Deciding a priori to analyze BIS instead would allow for maintaining the original alpha level.

When to combine RTs and PCs Before using any combination of RTs and PCs (or any other measure), the researcher must, of course, critically ask whether this combination makes theoretical sense in the given situation. Only when the cognitive process of interest affects both RTs and PCs (see Vandierendonck, 2017, p. 654), and if a trade-off is possible—that is, when the process is more error-prone when it is speeded (like a decision)—RTs and PCs can reasonably be interpreted as the result of a common underlying process. In

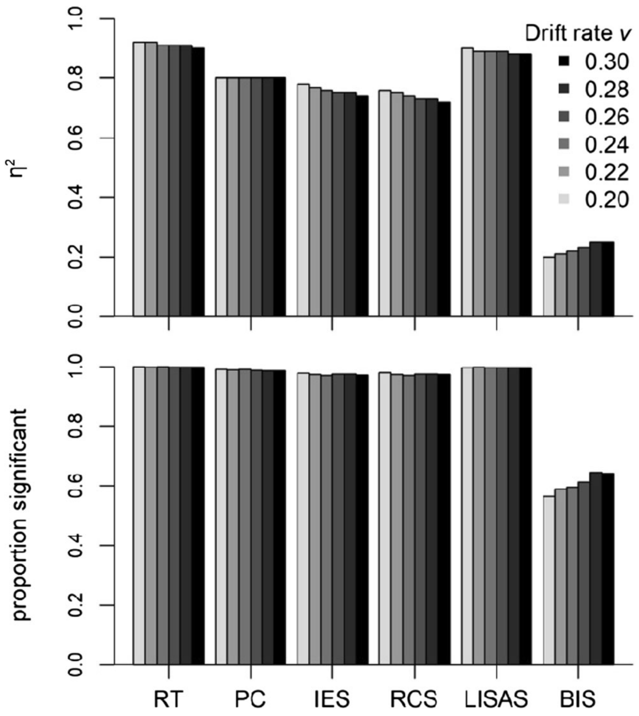


Fig. 7 Mean effect sizes η^2 and the proportions of significant one-way analyses of variance (between subjects) with three (randomly drawn) threshold separations for each measure

situations in which RTs and PCs are mainly influenced by different cognitive mechanisms, combined measures should be avoided. As an example, in the change-detection task (a typical task to assess visual working memory), participants see two subsequent arrays of objects and have to decide whether these are identical or whether one object has changed in between. In this task, the researcher might be interested in the capacity of working memory, which mainly affects PCs (when capacity is exceeded; Alvarez & Cavanagh, 2004; Luck & Vogel, 1997, 2013), or in the efficiency of the comparison between working memory entries and a test display, which mainly affects RTs (Gilchrist & Cowan, 2014; Hyun, Woodman, Vogel, Hollingworth, & Luck, 2009; Liesefeld, Liesefeld, Müller, & Rangelov, 2017). Combining RTs and PCs would confound capacity limitations and comparison efficiency, and therefore would complicate rather than clarify the interpretation of potential effects.

Risk of *p* hacking It might be tempting to check one or several combined measures whenever RTs and PCs yield a nonsignificant, but "trending," result in the same direction. Interpreting any resulting effect as confirmatory evidence would, of course, be misleading, due to the associated inflation in alpha error. It is, however, perfectly fine to decide in advance that BIS will be analyzed when the theory makes no clear predictions as to whether an effect influences RTs or PCs, or when the effect is expected to be distributed across both measures (e.g., due to inter- and intra-individual variation in SATs).

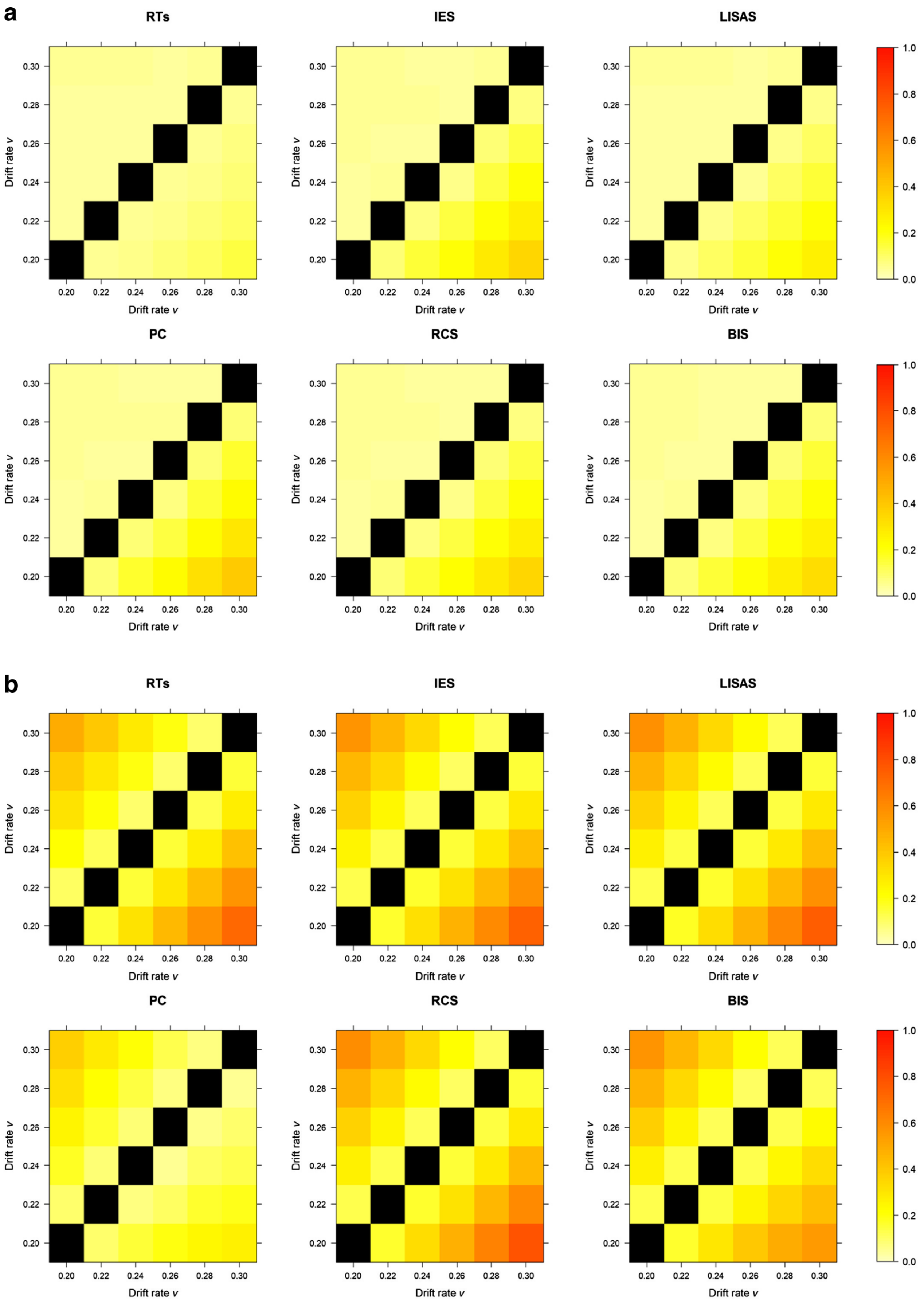
Sample dependence of BIS

The standardization across subjects and conditions involved in the calculation of BIS implies a major deviation from all previous combined measures: There is no BIS for a single cell; instead, individual values reflect whether performance was below (BIS < 0) or above (BIS > 0) the average performance (across all subjects and cells) for the respective subject in the respective cell of the design. In other words, a particular value of BIS is not only influenced by the performance of the respective subject in the respective cell, but also by the performance of the subject in other cells and the performance of the other subjects in the sample. In fact, the way the standardization is performed is the main difference to LISAS, which standardizes per subject/condition and thus provides sample-independent performance values (Vandierendonck, 2017, 2018).

The reader might wonder whether this sample dependence of BIS is problematic for its use. The answer to this question depends on the goal of calculating the measure. If the goal is to determine the *absolute performance* (i.e., without comparison to a specific group) of a particular subject in a particular task (this might sometimes apply during job recruiting or grading of academic achievements), BIS is not suited. It is, however, well suited for determining *relative performance*—that is, whether one (group of) subject(s) is *better* than another (group of) subject(s), or one condition is *more difficult* than another condition. This is exactly the type of question typically asked in (experimental) psychological research, for which the sample dependence of BIS is, consequently, unproblematic. On the contrary, concerning this type of question, BIS is often easier to interpret than the constituent measures (RTs and PCs), because it directly expresses whether a (group of) subject(s) performs above or below average (BIS > 0 or BIS < 0, respectively).

To approach this question from another vantage point, consider that statistical tests are insensitive to linear transformations such as the standardizations involved in BIS. In particular, the difference between two conditions in mean RTs will result, by definition, in the exact same *t* value as the difference of any linear transformation, if the transformation is applied uniformly to all RTs (such as a standardization with the same mean and the same standard deviation used in the calculation of BIS).[9] This feature of linear transformations also means that it is not the standardization, but the additive component (the subtraction), that does the job of controlling for SATs (as we demonstrated above).

[9] The skeptic is invited to confirm this fact on the sample data in Table 1 by calculating and comparing *t* values for the group difference in, say, RTs before (i.e., $\overline{RT_c}$) and after (i.e., z_{RT}) the standardization of RTs.


a
RTs
IES
LISAS
PC
RCS
BIS
b
RTs
IES
LISAS
PC
RCS
BIS
Drift rate v
0.20
0.22
0.24
0.26
0.28
0.30
1.0
0.8
0.6
0.4
0.2
0.0

◀ **Fig. 8** Effect sizes d for pairwise comparisons of drift rates v, reflecting "real" effects, at four different values of threshold separation a (panel **a**: $a = 5$ and $a = 100$; panel **b**: $a = 200$ and $a = 300$, above and below the diagonal, respectively). Each point (square) in a panel denotes a comparison between two combinations of drift rate (e.g., $v = 0.20$ vs. $v = 0.22$). The black diagonals indicate the absence of comparisons between a cell and itself (e.g., $v = 0.20$ vs. $v = 0.20$). Note that all combined measures reveal "real" effects as well as or better than any of the constituents (RTs or PCs) in most comparisons, and that BIS reveals these effects virtually as well as any other competitor in most comparisons, and only slightly worse than the best competitor in some comparisons (for $a = 300$, lower right areas in panel **b**, in particular).

Standardizing across different subsamples

In the present examination, RTs and PCs were standardized across all conditions and all subjects for a given test. This is, in a way, the most conservative approach, because all variance is kept (see the previous section). There might, however, be situations in which it is reasonable to remove some of the variance. When the research focus is on a Group × Treatment interaction, for example, it might be a good idea to remove the main effect of group (and the related error variance) by standardizing per individual (e.g., Bush, Hess, & Wolford, 1993; Faust, Balota, Spieler, & Ferraro, 1999) or per group.

In general, when calculating BIS, it is important to carefully ponder what shall be compared and therefore should be minimally included in the standardization. If, for example, standardization was performed separately per condition, any differences between conditions would be removed by design, and it would be impossible to detect any effects. If there is no particular reason to exclude a particular contributor of variance (as in the Group × Treatment example above), we recommend including the mean RT and PC for all subjects, groups, and conditions of the experiment, because this maximizes the data basis for calculating means and standard deviations.

Comparison to model fitting

In the present study, the combined measures were used to extract "real" effects that were simulated via manipulations of the drift-rate parameter of the diffusion model. In a way, the aim was to "recover" effects on drift rate and to ignore variations in another parameter (threshold separation). Obviously, the best way to recover any parameter of the diffusion model would be to fit the (simulated) data to the diffusion model itself. Indeed, a diffusion-model analysis of speed–accuracy data has several advantages in many situations (Forstmann et al., 2016; Ratcliff et al., 2016; Voss et al., 2015; Wagenmakers, 2009). Many, but not all, researchers would argue that major strengths of the diffusion model are that it is based on several well-validated theoretical assumptions and that its parameters are psychologically interpretable. These strengths, however, also restrict its use to specific situations, namely those in which a decision process is at the heart of the observed behavior.

BIS, in contrast, was developed on the basis of purely statistical considerations (the same is likely true for LISAS). It yields a balanced integration of RTs and PCs in any task, independent of what the task measures and which cognitive processes it involves. It would, of course, be advantageous to show for each specific task that BIS cancels SATs while maintaining "real" effects. Still, BIS was not specifically developed for decision processes and still performs quite well with data generated by a decision-process model; this gives us some confidence that BIS would perform equally well on data generated by other models/processes.

To elaborate a bit on how BIS might complement the modeling approach: Most popular models (such as the diffusion model) focus on the decision process, whereas many phenomena of interest to cognitive psychologists are captured in the residual "nondecision time" (e.g., Schmitz & Voss, 2012). Arguably, SATs can also occur in nondecision components of a task (Rinkenauer, Osman, Ulrich, Müller-Gethmann, & Mattes, 2004). For example, in a mental rotation task, Liesefeld et al. (2015) found that participants differed from each other in the time they took for performing the rotation, whereby taking less time meant that the resulting rotated representation of the original stimulus was less accurate and therefore more errors were committed; thus, it was not the decision component of the task, but the rotation component preceding the decision that was influenced by SATs.

Second, it is an empirical fact that researchers do use combined measures (e.g., Collignon et al., 2008; Gabay, Nestor, Dundas, & Behrmann, 2014; Kristjánsson, 2016; Kunde et al., 2012; Mevorach, Humphreys, & Shalev, 2006; Petrini, McAleer, & Pollick, 2010; Röder, Kusmierek, Spence, & Schicke, 2007; Spence, Kingstone, Shore, & Gazzaniga, 2001a; Spence, Shore, Gazzaniga, Soto-Faraco, & Kingstone, 2001b), and that reviewers do request the use of these measures (according to our own experiences and informal reports from colleagues), especially if RTs and PCs show opposite patterns of effects (which would indicate condition-contingent SATs). One reason for using simple combinations of RTs and PCs instead of decision models might be that authors desire a combined measure that does not rely on a specific psychological theory. This is understandable if the research focus does not lie on decision making (which is the case for most of the studies cited above) or if behavioral data are secondary to the research question (as in many neuroimaging studies employing SAT measures; e.g., Kiss, Driver, & Eimer, 2009; Küper, Gajewski, Frieg, & Falkenstein, 2017; Reeder, Hanke, & Pollmann, 2017). Thus, from a practical standpoint, there is quite some demand for combined speed–accuracy measures.

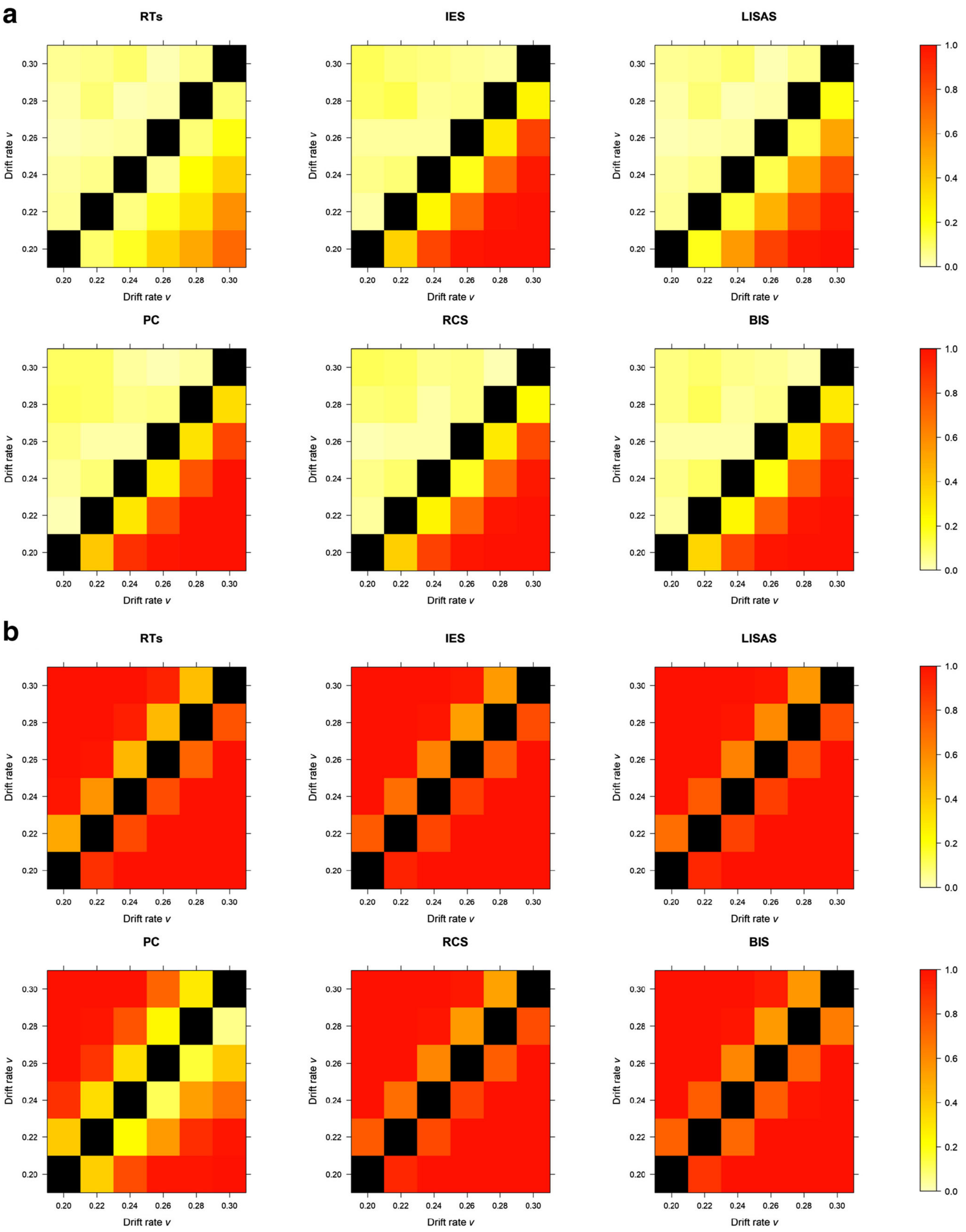

a
RTs
IES
LISAS
PC
RCS
BIS
b
RTs
IES
LISAS
PC
RCS
BIS
Drift rate v
0.20
0.22
0.24
0.26
0.28
0.30
Drift rate v
0.20
0.22
0.24
0.26
0.28
0.30
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 9 Proportions of significant pairwise comparisons of drift rates v (two-sample t tests), reflecting "real" effects, at four different values of threshold separation a (panel **a**: $a = 5$ and $a = 100$; panel **b**: $a = 200$ and $a = 300$, above and below the diagonal, respectively). Each point (square) in a panel denotes a comparison between two combinations of drift rate (e.g., $v = 0.20$ vs. $v = 0.22$). The black diagonals indicate the absence of comparisons between a cell and itself (e.g., $v = 0.20$ vs. $v = 0.20$).

Finally, BIS is easy to calculate, and therefore is potentially accessible to a wider range of researchers. Although tutorials and easy-to-use implementations and accessible tutorials for the diffusion model and other evidence accumulation models are available (e.g., Donkin, Brown, & Heathcote, 2011; Voss et al., 2015; Wagenmakers et al., 2007; Wagenmakers, van der Maas, Dolan, & Grasman, 2008), their correct application and interpretation still require considerable theoretical background. Easy-to-use code for calculating BIS in Matlab, R, and Excel can be retrieved from https://github.com/Liesefeld/BIS.

EZ-diffusion model

Another powerful, yet easy to calculate, tool for combining speed and accuracy data is the EZ-diffusion model (Wagenmakers et al., 2008; Wagenmakers et al., 2007). Based on the diffusion model (with a few simplifying assumptions), this model provides simple equations for calculating drift rate v, threshold separation a, and nondecision time t_0 on the basis of mean RTs, PCs, and the variance in RTs. In a way, the drift rate of EZ diffusion corresponds to the combined measures examined here (it would reflect our "real" effects while canceling out SATs). In addition, it provides estimates of threshold separation (and thus of the degree of SAT) and nondecision time. This model, of course, cannot be reasonably compared to the other combined measures with the present set of simulated data, because EZ diffusion is based on the same model that was used to generate the data. In fact, our simulations were designed so that they would meet all (or virtually all; see below) assumptions underlying EZ diffusion (which is not guaranteed with real data; see also the section below on Desirable Extensions of the Present Simulations).

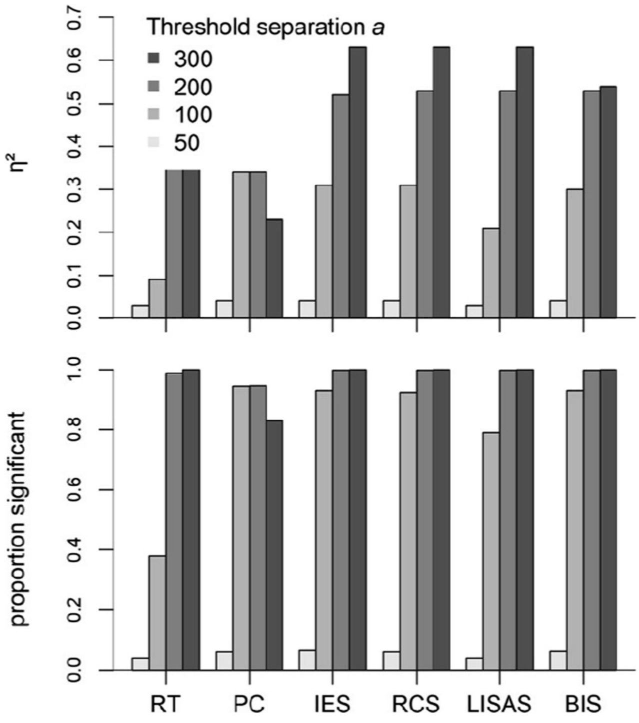


Fig. 10 Mean effect sizes η^2 and the proportions of significant one-way ANOVAs averaged across all possible combinations (without replacement) of three drift rates (between subjects) for each measure.

To illustrate the use of the EZ-diffusion model and to also validate our simulations, we extracted drift rate, threshold separation, and non-decision time from our data set using the EZ-diffusion model. The results are visualized in Fig. 11, and two interesting pieces of information are revealed: First, the results validate our simulation by showing that the parameters are recovered well in large parts. In particular, the drift-rate parameter seems mostly independent of the simulated threshold separation, but it is influenced by our "real" effects (as induced via variations in drift rate). Second, the results point to a limitation of the EZ-diffusion model whenever its assumptions are violated. The particular violation here (and elsewhere—e.g., Ratcliff, 2008) is that a trial is aborted after a while (a *response deadline*; here, around 3,500 ms), which implies that sometimes the decision process cannot finish. The larger the threshold separation is, the more often this happens, thus leading to distorted RT patterns in these cases (a few very long RTs are missing in the data; see note 7). In Fig. 11, this becomes most obvious in the overestimations of nondecision time with high threshold separations (right panel).

Additionally, it remains to be investigated whether parameter extraction using the EZ-diffusion model has advantages over the other measures in canceling SATs when data were not generated with the diffusion model but with, for example, the leaky competing accumulator model (Usher & McClelland, 2001), the linear ballistic model (Brown & Heathcote, 2008), or the fast-guess model (Ollman, 1966; see Van Ravenzwaaij & Oberauer, 2009, for related comparisons).

Outlook: Open questions and future directions

Confounds of variation in threshold separation and drift rate

This article has treated only two idealized situations: pure SATs without any "real" effects (a threshold separation variation) and pure "real" effects without any SAT (a drift-rate variation). In these situations, there is no correlation between SATs and the "real" effects across conditions (because one of the two was always kept constant). All types of combinations of these two situations are, of course, possible and likely do occur in reality.

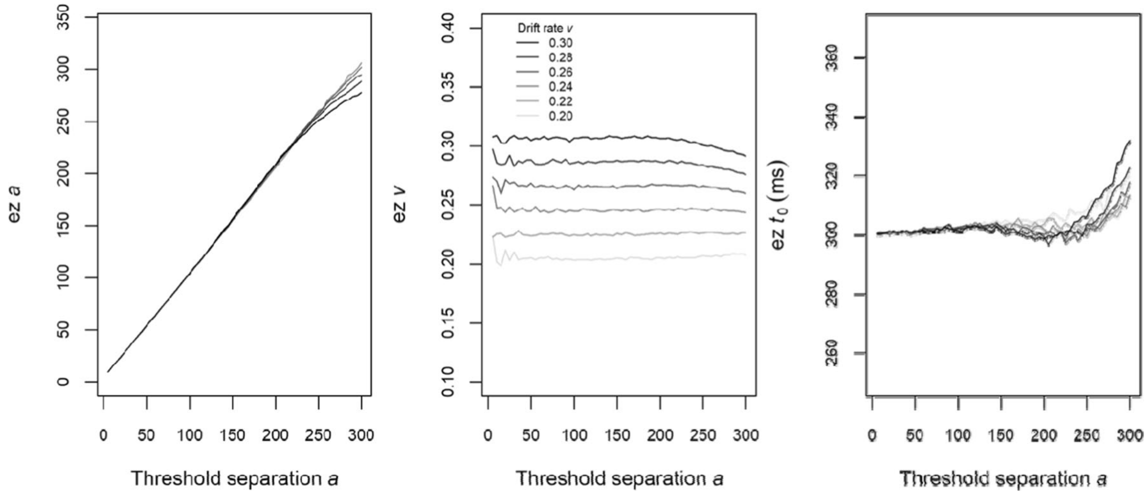


Fig. 11 Threshold separation a, drift rate v, and nondecision time t_0 as extracted from the simulated data using the EZ-diffusion model (Wagenmakers et al., 2008; Wagenmakers et al., 2007), as a function of threshold separation a and drift rate v as implemented in the simulation generating the data set. Note that the mean nondecision time was set to 300 ms in our simulations.

Unfortunately, a systematic investigation of confounds between threshold separation and drift rates is subject to a combinatory explosion (in the present case of 60 levels of threshold separation and six levels of drift rates, there are 360 possible combinations of these two parameters, yielding 64,620 pairwise comparisons) and is beyond the scope of the present article. That BIS cancels out pure SATs and leaves intact pure "real" effects is already important information, especially in light of our demonstrations that other measures dramatically fail already in the simplest situation of pure SATs. Nevertheless, preliminary explorations of this combinatory space are in line with the general pattern reported here: BIS cancels or strongly reduces effects of SATs, while typically maintaining "real" effects (see Appendix C for more details).

Desirable extensions of the present simulations Although the diffusion model is arguably one of the best validated and most established models to reflect core cognitive processes employed in a wide range of experimental tasks (Forstmann et al., 2016; Ratcliff et al., 2016; Voss et al., 2015; Wagenmakers, 2009), it is likely that data simulated by this model differ from real data in various respects. Furthermore, although variations in threshold separation are the standard approach to induce SATs, there is some debate as to whether SATs are (typically) reflected in (pure) variations of threshold separation (Lerche & Voss, in press; Rae, Heathcote, Donkin, Averell, & Brown, 2014; Rinkenauer et al., 2004; Starns, Ratcliff, & McKoon, 2012; Voss et al., 2004), casting additional doubt on the validity of the present simulations. For these reasons, future studies should strive to confirm the results reported here with alternative operationalizations of SATs—namely, by using other parameter combinations (including nondecision time t_0 and drift rate v; see, e.g., Rae et al., 2014; Rinkenauer et al., 2004) and/or other models (e.g., the leaky competing accumulator model, Usher & McClelland, 2001; the linear ballistic model, Brown & Heathcote, 2008; or the fast-guess model, Ollman, 1966), and, with some qualifications (see the Quality Criteria for Combined Measures section above), real data (see Bruyer & Brysbaert, 2011; Vandierendonck, 2018).

Other statistical tests We have rather exhaustively tested comparisons of two independent samples, have only parenthetically tested designs with more levels of a factor (three), and have excluded any multifactorial and within-subjects designs. Although there is no reason to believe that other combined measures would gain the lead in these situations, and although the results look very similar in some preliminary explorations with such designs, these assumptions should be tested carefully and systematically in future work. Similarly, it appears likely that BIS improves results in correlative approaches (see, e.g., Draheim et al., 2016; Hughes et al., 2014; Van Ravenzwaaij & Oberauer, 2009), but this topic, too, must await future validation.

Unequal weighting of RTs and PCs BIS was designed to integrate RTs and PCs in a balanced manner. There is no guarantee, however, that balanced weighting is ideal (this would be a rather surprising coincidence, in fact). Thus, future research should strive to determine which weighting of RTs and PCs is ideal in a given situation. For the meantime, an equal integration of the two constituents seems the most reasonable choice to us. If it turns out that an unequal weighting is preferable, the equal (and constant across accuracy levels) weighting is still a convenient feature of BIS, because it allows easily adapting the relative weights of RTs and PCs. This can be achieved by simply adding a weighting parameter w to Eq. 4 (with $0 < w < 1$), as in, for example:

$$BIS_{i,j} = w \cdot z_{PC_{i,j}} - (1-w) \cdot z_{\overline{RT}_{i,j}}$$

A major difficulty with such an endeavor would be to find criteria according to which one should determine w. Obviously, to try different values for w until a desired outcome (a statistically significant effect) is obtained would inflate the alpha error and must be avoided.

Transforming the constituents Close inspection of Figs. 5, 6, and 7 indicates that for BIS, SATs influence ANOVAs much more than they influence t tests (although this influence is still considerably less than for all of the alternatives). One remedy would be to avoid using ANOVAs for testing critical hypotheses and instead to focus on t tests or contrasts (which usually reflect the hypothesis of interest much better, anyway). Another alternative might be to transform RTs and PCs before entering them into Eq. 4. In particular, BIS integrates RTs and PC linearly, ignoring that RTs and PCs are typically not linearly related. Closer approximation to a linear relationship between RT and PC can be achieved by first transforming both measures. It turns out that the following transformations provide reasonable approximations to linearity[10] (but see, e.g., Lo & Andrews, 2015, for potential pitfalls of such transformations):

$$\overline{RT}'_{i,j} = \ln\left(\overline{RT}_{i,j}\right), \text{ and } PC'_{i,j} = \ln\left(\frac{1}{1-PC_{i,j}}\right)$$

Combining multiple measures BIS is in no way restricted to combining only RTs and PCs. To give an example, complex-span tasks are measures of working memory capacity that correlate highly with general intelligence. In these tasks, participants have to remember a sequence of memoranda (e.g., words) and after each memorandum a short processing task has to be solved (e.g., verification of an algebraic equation). After several memorandum–processing pairs, participants have to recall all memoranda. Usually, analyses of this type of task focus on recall performance, but it turned out that accuracy on the processing part correlates with intelligence, too (Unsworth, Redick, Heitz, Broadway, & Engle, 2009)—potentially because people do trade off memorizing and processing. BIS could be used to combine performance on both aspects of the task in order to gain a more comprehensive measure of complex-span performance.

Furthermore, BIS can combine an arbitrary number of performance measures, by simply standardizing all constituents and adding measures for which high values reflect good performance (such as PC) and subtracting measures for which high values reflect bad performance (such as RTs). To stick with the example of complex-span tasks, in addition to recall accuracy and processing accuracy, processing time could be included as a third performance measure (see Unsworth et al., 2009).

Conclusion

We have formally introduced and validated a new approach to control for speed–accuracy trade-offs, the balanced integration score (BIS), and compared it to alternative measures. This measure effectively controls for speed–accuracy trade-offs while retaining true effects. Furthermore, it is highly flexible and easy to calculate. Matlab and R code as well as an Excel sheet for calculating this measure can be retrieved from https://github.com/Liesefeld/BIS

Author note H.R.L.'s work is supported by LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative, and by the Graduate School of Systemic Neurosciences, Munich Center for Neurosciences–Brain & Mind. M.J.'s work is supported by the Institutional Strategy of the University of Tübingen (ZUK 63, German Research Council). We thank Scott Brown, Marc Brysbaert, Anna Liesefeld, Rolf Ulrich, and André Vandierendonck for helpful and constructive comments and suggestions on a previous version of the manuscript.

Appendix A: The relationship between IES and RCS

For the following discussion, consider the dataset given in Table 2. The first trials 1, …, m were responded to correctly (coded as 1), and trials $m + 1$, …, M represent error trials (coded as 0).

The argument involves three steps. Steps 1 and 2 will slightly transform IES and RCS, respectively, and Step 3 will bring both measures together and express IES as a function of RCS.

Table 2 Structure of a data set to demonstrate the relationship between IES and RCS

Trial	Correct	RT
1	1	RT_1
. . .	. . .	. . .
m	1	RT_m
m+1	0	RT_{m+1}
. . .	. . .	. . .
M	0	RT_M

[10] We thank Jochen Krebs for advice in this regard.

Step 1.

$$IES = \frac{R\bar{T}_c}{1-PE} = \frac{R\bar{T}_c}{PC} = \frac{\frac{\sum_{i=1}^{m} RT_i}{m}}{\frac{m}{M}}$$

$$= \frac{\sum\limits_{i=1}^{m} RT_i}{m} \cdot \frac{M}{m} \Rightarrow IES \cdot \frac{m}{M} = \frac{\sum\limits_{i=1}^{m} RT_i}{m}$$

Step 2.

$$RCS = \frac{m}{\sum_{i=1}^{M} RT_i} \Rightarrow \frac{1}{RCS} = \frac{\sum_{i=1}^{M} RT_i}{m} = \frac{\sum_{i=1}^{m} RT_i}{m} + \frac{\sum_{i=m+1}^{M} RT_i}{m}$$

$$\Rightarrow \frac{1}{RCS} - \frac{\sum_{i=m+1}^{M} RT_i}{m} = \frac{\sum_{i=1}^{m} RT_i}{m}$$

Step 3. Because the rightmost parts of the equations above are equal, the terms to the left of the equal sign can be equalized:

$$IES \cdot \frac{m}{M} = \frac{1}{RCS} - \frac{\sum_{i=m+1}^{M} RT_i}{m}$$

$$\Rightarrow IES = \frac{M}{m}\left(\frac{1}{RCS} - \frac{\sum_{i=m+1}^{M} RT_i}{m}\right) = \frac{M}{m \cdot RCS} - \frac{M \cdot \sum_{i=m+1}^{M} RT_i}{m^2}$$

Although this last equation shows the general case including both correct and erroneous responses, an interesting relationship becomes apparent when we assume that the proportion correct (*PC*) approaches 1 (what means that *m* approaches *M*). In this case, the minuend of the rightmost term approaches $\frac{1}{RCS}$. Because the nominator of the subtrahend contains the sum of erroneous RTs—which will approach 0—the subtrahend will disappear. Thus, in the extreme case of $PC = 1$, it follows that

$$IES = \frac{1}{RCS}$$

The same follows for any *PC* if RTs from incorrect-response trials are included in IES (as was apparently intended by Townsend & Ashby, 1983, p. 204):

$$IES = \frac{\overline{RT}}{PC} = \frac{\frac{\sum_{i=1}^{M} RT_i}{M}}{\frac{m}{M}} = \frac{\sum_{i=1}^{M} RT_i}{m} = \frac{1}{RCS}$$

Appendix B: Correlation of RTs and PCs with BIS

In the main text, we noted that for BIS the ratio given in Eq. 5 equals 1—that is, $I_{BIS} = \frac{r^2_{RT,BIS}}{r^2_{PC,BIS}} = 1$. This is true because $r_{RT,\ BIS} = -r_{PC,\ BIS}$ holds. In this appendix, we provide a formal proof for the latter relation. Because linear transformations do not change the correlation of two variables, the equation under question can be stated slightly differently:

$$r_{RT,BIS} = -r_{PC,BIS} \Leftrightarrow r_{z_{RT},BIS} = -r_{z_{PC},BIS}$$

Transformation of the latter formulation then shows the claimed equality

$$\begin{aligned}
& r_{z_{RT},BIS} = -r_{z_{PC},BIS} \\
& \Leftrightarrow r_{z_{RT},z_{PC}-z_{RT}} = -r_{z_{PC},z_{PC}-z_{RT}} \\
& \Leftrightarrow \frac{cov(z_{RT}, z_{PC}-z_{RT})}{S_{z_{RT}} \cdot S_{z_{PC}-z_{RT}}} = -\frac{cov(z_{PC}, z_{PC}-z_{RT})}{S_{z_{PC}} \cdot S_{z_{PC}-z_{RT}}} \\
& \Leftrightarrow \frac{cov(z_{RT}, z_{PC}) - cov(z_{RT}, z_{RT})}{1 \cdot S_{z_{PC}-z_{RT}}} = -\frac{cov(z_{PC}, z_{PC}) - cov(z_{PC}, z_{RT})}{1 \cdot S_{z_{PC}-z_{RT}}} \\
& \Leftrightarrow \frac{cov(z_{RT}, z_{PC}) - 1}{1 \cdot S_{z_{PC}-z_{RT}}} = -\frac{1 - cov(z_{PC}, z_{RT})}{1 \cdot S_{z_{PC}-z_{RT}}} \\
& \Leftrightarrow cov(z_{RT}, z_{PC}) - 1 = -(1 - cov(z_{RT}, z_{PC})) \\
& \Leftrightarrow cov(z_{RT}, z_{PC}) - 1 = cov(z_{PC}, z_{RT}) - 1
\end{aligned}$$

Appendix C: Variations in threshold separation *and* drift rate (preliminary analysis)

In the main article, we focused on evaluating situations in which either only threshold separation *a* or only drift rate *v* varied. These simulations allowed for full control over the degree of SATs and "real" effects, and it was possible to vary both influences independently. Empirical data, in contrast, might confound SATs and "real" effects, and it is not typically possible to tease them apart perfectly. Instead, decreasing RTs accompanied by decreasing PCs across conditions are sometimes used as a convenience criterion indicating an SAT in empirical data.

A thorough coverage of situations with concurrent variations in threshold separation and drift rate is beyond the scope of the present article. Yet, a preliminary analysis suggests that the combined measures behave similarly to what we have reported for pure effects. In particular, we randomly sampled one experiment and, from this experiment, two random subsamples 10,000 times (with differing threshold separation *a* and/or drift rate *v*). On the basis solely of mean RTs and PCs, we determined whether or not an SAT was present (decreasing RTs with decreasing PC) and calculated the proportions of significant two-sample *t* tests for all dependent measures (RTs, PCs, IES, RCS, LISAS, and BIS). The results are provided in Table 3. Without empirical evidence for an SAT, all combined measures appear to *increase* the likelihood of obtaining a significant result to more or less the same degree (when compared to RTs and PCs). In cases with empirical evidence for an SAT, the proportion of significant results is slightly attenuated for IES, RCS, and LISAS, but this desirable attenuation is much more pronounced for BIS.

Despite this (for BIS) encouraging outcome, we would point out the preliminary character of these results. Future work will be required in order to consider varying effect sizes for RT and PC differences. Furthermore, in the present analysis, the contributions of differences in threshold separation *a* and drift rate *v* are not clear, so we cannot exclude the possibility that BIS overcorrects and, thus, underestimates the true effects in these situations.

Table 3 Proportions of significant two-sample *t* tests for six different dependent measures separately for cases without and with an SAT (as indicated by opposing effects on mean RTs and PCs)

	RTs	PC	IES	RCS	LISAS	BIS
Without SAT	.85	.59	.94	.94	.95	.95
With SAT	.97	.94	.87	.87	.94	.14

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111. doi:https://doi.org/10.1111/j.0963-7214.2004.01502006.x

Akhtar, N., & Enns, J. T. (1989). Relations between covert orienting and filtering in the development of visual attention. *Journal of Experimental Child Psychology*, *48*, 315–334. doi:https://doi.org/10.1016/0022-0965(89)90008-8

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics*, *73*, 640–657. doi:https://doi.org/10.3758/s13414-010-0049-7

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652. doi:https://doi.org/10.1037/0033-295X.108.3.624

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178. doi:https://doi.org/10.1016/j.cogpsych.2007.12.002

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, *51*, 5–13. doi:https://doi.org/10.5334/pb-51-1-5

Bush, L. K., Hess, U., & Wolford, G. (1993). Transformations for within-subject designs: A Monte Carlo investigation. *Psychological Bulletin*, *113*, 566–579. doi:https://doi.org/10.1037/0033-2909.113.3.566

Cohen, J. (1988). Statistical power analysis for the behavioral sciences *(*2nd ed*.)*. Hillsdale, NJ: Erlbaum.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135. doi:https://doi.org/10.1016/j.brainres.2008.04.023

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*, 140–151. doi:https://doi.org/10.1016/j.jmp.2010.10.001

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and accuracy: The relationship between working memory capacity and task switching as a case example. *Perspectives on Psychological Science*, *11*, 133–155. doi:https://doi.org/10.1177/1745691615596990

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. J., Forstmann, B. U., & Wagenmakers, E.-J. (2012). How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology*, *56*, 208–216. doi:https://doi.org/10.1016/j.jmp.2012.04.001

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–799. doi:https://doi.org/10.1037/0033-2909.125.6.777

Forstmannn, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666. doi:https://doi.org/10.1146/annurev-psych-122414-033645

Gabay, S., Nestor, A., Dundas, E., & Behrmann, M. (2014). Monocular advantage for face perception implicates subcortical mechanisms in adult humans. *Journal of Cognitive Neuroscience*, *26*, 927–937. doi:https://doi.org/10.1162/jocn_a_00528

Germar, M., Schlemmer, A., Krug, K, Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision-making: A diffusion model analysis. *Personality and Social Psychology Bulletin*, *40*, 217–231. doi:https://doi.org/10.1177/0146167213508985

Gilchrist, A. L., & Cowan, N. (2014). A two-stage search of visual working memory: Investigating speed in the change-detection paradigm. *Attention, Perception, & Psychophysics*, *76*, 2031–2050. doi:https://doi.org/10.3758/s13414-014-0704-5

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*, 299–308. doi:https://doi.org/10.1016/S0896-6273(02)00971-6

Gueugneau, N., Pozzo, T., Darlot, C., & Papaxanthis, C. (2017). Daily modulation of the speed–accuracy trade-off. *Neuroscience*, *356*, 142–150. doi:https://doi.org/10.1016/j.neuroscience.2017.04.043

Heitz, R. P. (2014). The speed–accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150. doi:https://doi.org/10.3389/fnins.2014.00150

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods*, *46*, 702–721. doi:https://doi.org/10.3758/s13428-013-0411-5

Hyun, J., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1140–1160. doi:https://doi.org/10.1037/a0015019

Janczyk, M., & Lerche, V. (in press). A diffusion model analysis of the response–effect compatibility effect. *Journal of Experimental Psychology: General*. doi:10.1037/xge0000430

Janczyk, M., Mittelstädt, P., & Wienrich, C. (2018). Parallel dual-task processing and task-shielding in older and younger adults: Behavioral and diffusion model results. *Experimental Aging Research*, *44*, 95–116. doi:https://doi.org/10.1080/0361073X.2017.1422459

Kiss, M., Driver, J., & Eimer, M. (2009). Reward priority of visual target singletons modulates event-related potential signatures of attentional selection. *Psychological Science*, *20*, 245–251. doi:https://doi.org/10.1111/j.1467-9280.2009.02281.x

Kristjánsson, Á. (2016). The slopes remain the same: Reply to Wolfe (2016). *i-Perception*, *7*, 1–4.

Kunde, W., Pfister, R., & Janczyk, M. (2012). The locus of tool-transformation costs. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 703–714. doi:https://doi.org/10.1037/a0026315

Küper, K., Gajewski, P. D., Frieg, C., & Falkenstein, M. (2017). A randomized controlled ERP study on the effects of multi-domain cognitive training and task difficulty on task switching performance in older adults. *Frontiers in Human Neuroscience*, *11*, 184. doi:https://doi.org/10.3389/fnhum.2017.00184

Laming, D. R. J. (1968). Information theory of choice-reaction times. London, UK: Academic Press.

Lerche, V., & Voss, A. (in press). Speed–accuracy manipulation in diffusion modeling: Lack of discriminant validity of the manipulation or of the parameter estimates? *Behavior Research Methods*. doi:https://doi.org/10.3758/s13428-018-1034-7

Liesefeld, H. R., Fu, X., & Zimmer, H. D. (2015). Fast and careless or careful and slow? Apparent holistic processing in mental rotation is explained by speed–accuracy trade-offs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1140–1151. doi:https://doi.org/10.1037/xlm0000081

Liesefeld, H. R., Liesefeld, A. M., Müller, H. J., & Rangelov, D. (2017). Saliency maps for finding changes in visual scenes?. *Attention, Perception, & Psychophysics*, *79*, 2190–2201. doi:https://doi.org/10.3758/s13414-017-1383-9

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. doi:https://doi.org/10.3389/fpsyg.2015.01171

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organisation (Oxford Psychology Series, Vol. 8). New York, NY: Oxford University Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281. doi:https://doi.org/10.1038/36846

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*, 391–400. doi:https://doi.org/10.1016/j.tics.2013.06.006

Mevorach, C., Humphreys, G. W., & Shalev, L. (2006). Opposite biases in salience-based selection for the left and right posterior parietal cortex. *Nature Neuroscience*, *9*, 740–742. doi:https://doi.org/10.1038/nn1709

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*, 155–156. doi:https://doi.org/10.3758/BF03328004

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, *35*, 737–743. doi:https://doi.org/10.1177/001872089303500412

Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), Human information processing: Tutorials in performance and cognition (pp. 41–82). Hillsdale, NJ: Erlbaum.

Paoletti, D., Weaver, M. D., Braun, C., & van Zoest, W. (2015). Trading off stimulus salience for identity: A cueing approach to disentangle visual selection strategies. *Vision Research*, *113*(Pt. B), 116–124. doi:https://doi.org/10.1016/j.visres.2014.08.003

Petrini, K., McAleer, P., & Pollick, F. (2010). Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence. *Brain Research*, *1323*, 139–148. doi:https://doi.org/10.1016/j.brainres.2010.02.012

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, and Cognition*, *40*, 1226–1243. doi:https://doi.org/10.1037/a0036801

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. doi:https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychonomic Bulletin & Review*, *15*, 1218–1228. doi:https://doi.org/10.3758/PBR.15.6.1218

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*, 260–281. doi:https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General*, *140*, 464–487. doi:https://doi.org/10.1037/a0023810

Reeder, R. R., Hanke, M., & Pollmann, S. (2017). Task relevance modulates the representation of feature conjunctions in the target template. *Scientific Reports*, *7*, 4514. doi:https://doi.org/10.1038/s41598-017-04123-8

Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62. doi:https://doi.org/10.1016/j.cognition.2014.09.005

Rinkenauer, G., Osman, A., Ulrich, R., Müller-Gethmann, H., & Mattes, S. (2004). On the locus of speed–accuracy trade-off in reaction time: inferences from the lateralized readiness potential. *Journal of Experimental Psychology: General*, *133*, 261–282. doi:https://doi.org/10.1037/0096-3445.133.2.261

Röder, B., Kusmierek, A., Spence, C., & Schicke, T. (2007). Developmental vision determines the reference frame for the multisensory control of action. *Proceedings of the National Academy of Sciences*, *104*, 4753–4758. doi:https://doi.org/10.1073/pnas.0607158104

Sanders, A. F. (1998). Elements of human performance: Reaction processes and attention in human skill. Mahwah, NJ: Erlbaum.

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 222–250. doi:https://doi.org/10.1037/a0026003

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence*, *51*, 28–46. doi:https://doi.org/10.1016/j.intell.2015.05.002

Spence, C., Kingstone, A., Shore, D. I., & Gazzaniga, M. S. (2001a). Representation of visuotactile space in the split brain. *Psychological Science*, *12*, 90–93. doi:https://doi.org/10.1111/1467-9280.00316

Spence, C., Shore, D. I., Gazzaniga, M. S., Soto-Faraco, S., & Kingstone, A. (2001b). Failure to remap visuotactile space across the midline in the split-brain. *Canadian Journal of Experimental Psychology*, *55*, 133–140. doi:https://doi.org/10.1037/h0087360

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*, 1–34. doi:https://doi.org/10.1016/j.cogpsych.2011.10.002

Thura, D., Guberman, G., & Cisek, P. (2017). Trial-to-trial adjustments of speed–accuracy trade-offs in premotor and primary motor cortex. *Journal of Neurophysiology*, *117*, 665–683. doi:https://doi.org/10.1152/jn.00726.2016

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

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148–174. doi:https://doi.org/10.1016/j.cogpsych.2015.02.005

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*, 635–654. doi:https://doi.org/10.1080/09658210902998047

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550–592. doi:https://doi.org/10.1037/0033-295X.108.3.550

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*, 653–673. doi:https://doi.org/10.3758/s13428-016-0721-5

Vandierendonck, A. (2018). Further tests of the utility of integrated speed–accuracy measures in task switching. *Journal of Cognition*, *1*, 8. doi:https://doi.org/10.5334/joc.6

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011–1026. doi:https://doi.org/10.3758/BF03193087

Van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*, 463–473. doi:https://doi.org/10.1016/j.jmp.2009.09.004

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology*, *44*, 1048–1056. doi:https://doi.org/10.1016/j.jesp.2007.10.009

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220. doi:https://doi.org/10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*, 767–775. doi:https://doi.org/10.3758/BF03192967

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology*, *6*, 336. doi:https://doi.org/10.3389/fpsyg.2015.00336

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671. doi:https://doi.org/10.1080/09541440802205067

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*, 1229–1235. doi:https://doi.org/10.3758/PBR.15.6.1229

Wagenmakers, E.-J., van der Maas, H. J., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22. doi:https://doi.org/10.3758/BF03194023

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85. doi:https://doi.org/10.1016/0001-6918(77)90012-9

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition*, *34*, 668–684. doi:https://doi.org/10.3758/BF03193587