Behav Res (2017) 49:653–673
DOI 10.3758/s13428-016-0721-5

A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure

André Vandierendonck[1]

Published online: 4 March 2016

Abstract In cognitive research, speed and accuracy are two important aspects of performance. When analyzed separately, these performance variables sometimes lead to contradictory conclusions about the effect of a manipulation. To avoid such conflicts, several measures that integrate speed and accuracy have been proposed, but the added value of using such measures remains unclear. The present paper compares the relative utility of seven integrated performance measures, namely four variations on a binning procedure that weights response times of correct and incorrect trials differently, and three measures that combine averaged speed and accuracy scores. The properties of these integrated measures were explored in three simulation studies. The first study compared three binning measures and showed that one measure failed to grasp the performance difference between two conditions. The second study showed that the sampling distributions of the measures were symmetric, except for a strong skewness on the rate correct score. The third study varied the trade-off and the effect sizes of speed and accuracy in four different combinations of size and direction of speed and accuracy effects. These studies highlighted some further shortcomings of the binning measures. The combination measures performed well, but linear integration of speed and accuracy and rate correct score were most efficient in detecting effects and accounting for a larger proportion of the variance. The paper concludes that these combination measures are useful provided that the speed and accuracy data are also inspected.

✉ André Vandierendonck
Andre.Vandierendonck@UGent.be

[1] Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium

Keywords Speed · Accuracy · Integrated speed-accuracy scoring

Introduction

Research on human performance requires paradigms that induce changes in performance as expressed in response time, accuracy or both. For example, tasks involving incompatibility are typically performed slower and are often more error-prone than tasks with compatible stimuli and/or responses (Kornblum, Hasbroucq, & Osman, 1990; MacLeod, 1991; Stroop, 1935); similarly, situations requiring task switching lead to slower responding and/or increased error rates (e.g., Kiesel et al., 2010; Vandierendonck, Liefooghe, & Verbruggen, 2010). In such paradigms, the imposed variations may have different effects on response speed and accuracy, possibly as a result of differences in the speed-accuracy balance.

Contradictory findings in these two important aspects of performance might be avoided when measurements of response time (RT) and accuracy (proportion of errors; PE) are integrated into a single measure. Integrative measures that have been proposed include the inverse efficiency score (IES; Townsend & Ashby, 1978), the rate correct score (RCS; Woltz & Was, 2006), and a Bin score, i.e., a score based on partitioning the RTs in the data set into bins (Hughes, Linck, Bowles, Koeth, & Bunting, 2014). Although such measures have occasionally been used, there is no general agreement about the utility of such integrated measures and their efficiency in detecting performance differences. Bruyer and Brysbaert (2011) applied the oldest of these measures, IES, to published data sets and obtained mixed results. Only when the RTs and PEs are correlated or when the proportions of errors are rather low, did IES seem to offer some help. These authors advised against using IES without also inspecting RT and PE. Similar conclusions regarding IES followed from the study of Hughes et al. (2014). These authors applied IES, RCS and a Bin score to some data sets and concluded that RCS and the Bin score are more reliable than IES and the single measures of RT and PE.

The latter findings might be taken to indicate that by using RCS or the Bin score all our problems are solved. Unfortunately, nothing is further from the truth. In fact, notwithstanding the good news reported by Hughes et al., some important problems remain to be solved. First, it is surprising that RCS was better than IES in the Hughes et al. study, because they are both a ratio of correct responses and correct RTs (IES) or all RTs (RCS). Second, the Bin measure proposed by Hughes et al. (2014) has two important disadvantages: (1) it measures the performance difference between two conditions (e.g., the switch cost or the congruency cost) so that its use is limited to situations involving a single contrast between two conditions; (2) also, and more importantly, when there is no difference between the two conditions (e.g., neither RT nor PE switch cost), the Bin score yields a positive number that is substantially larger than zero. The present article, therefore, rejoins the debate about integrated measures of latency and accuracy, in order to achieve a clearer picture regarding the utility of integrated performance measures. More specifically, improvements on the Bin score will be considered, and as an alternative to IES and RCS, a linear combination of RT and PE will be proposed. These measures will be tested on the basis of a series of Monte Carlo simulations.

Integrated measures of response time (RT) and proportion of errors (PE)

Before considering integrated measures of speed and accuracy in more detail, it is important to delineate the focus of such measures. Indeed, in some tasks, typically well-learned tasks, it may be hypothesized that speed and accuracy of performance are driven by common or overlapping processes. Consequently, when the response process is speeded, for example by instructions or by the presence of a response deadline, responding will become more error-prone, with the occurrence of choking under pressure as an extreme case (Beilock, Kulp, Holt, & Carr, 2004). However, in other tasks, speed and accuracy relate to different underlying mechanisms. Categorization and concept learning tasks using well-defined categories provide an example in point. In such a task (e.g., Trabasso & Bower, 1966), errors are indicative of a state of not yet knowing the categorization rule. The occurrence of an error signals the need to change the currently tested rule, with the result that more intricate processing will occur after an error than after a correct response (e.g., White, 1972). Speeding up responding will not result in a dramatic increase in the number of errors, but it may interfere with the processes involved in selecting a new rule after an error. In other words, in tasks of the latter type, integration of speed and accuracy into a single measure would not help at all to achieve a more stable and informative measurement of performance. For that reason, the present scrutiny of measures that integrate speed and accuracy scores is restricted to tasks for which it can reasonably be hypothesized that they—at least in part—result from shared processes.

The oldest and most frequently used measure that integrates RT and PE is the inverse efficiency score, or IES (Townsend & Ashby, 1978). Its definition is quite simple, namely

$$\mathrm{IES} = \frac{\mathrm{RT}}{1 - \mathrm{PE}} \qquad (1)$$

where RT is the subject's average (correct) RT of the condition, and PE is the subject's proportion of errors in the condition. As an example, if average correct RT is 500 ms, and the proportion of errors is .10, IES will be 500/(1−.10) = 556. IES can be considered as the RT corrected for the amount of errors committed.

The rate correct score or RCS (Woltz & Was, 2006) is defined as

$$\mathrm{RCS} = \frac{c}{\sum \mathrm{RT}} \qquad (2)$$

where c is the number of correct responses in the condition, and the denominator refers to the sum of all RTs in the set of trials under consideration. If there are 100 trials with 90 correct responses and the average RT is 0.500 s, RCS = 90/50 = 1.8.[1] This score can be interpreted as the number of correct responses per second of activity.

The Bin score cannot be expressed in a simple mathematical formula. Its calculation, as defined in Hughes et al. (2014), assumes that there are two conditions, basically a control and an experimental condition (e.g., task repetition and task switch trials to estimate the task switch cost), and is performed by executing the following steps:

1. Over all participants and all trials belonging to the control condition, calculate the average RT (RT_c).
2. Over all participants and all correct trials belonging to the experimental condition, calculate RT-RT_c; sort all these values from small to large and calculate the deciles. Each decile constitutes a bin; these bins are numbered 1 to 10.

[1] This example implicitly changes the definition of Equation 2 by taking the proportion correct and the average RT instead of the sum of all RTs. This also has the advantage of making the measure comparable across conditions.

3. For each participant, count the number of correct difference scores (RT-RT$_c$) of the experimental condition in each of the 10 bins (n_i); also count the number of error trials (n_e) and assign them to the "bad" bin. The score can then be calculated as follows:

$$\left(\sum_{i=1}^{10} n_i \times i\right) + n_e \times 20,$$

where i is the number of the bin (1–10); note that the bad bin is assigned a weight (or penalty) of 20.
4. The obtained score expresses the size of the difference in performance (accuracy and latency) between the experimental and the control conditions. In the remainder of this article, the score calculated following these steps will be referred to as Bin-o (bin-original).

At this point, a few comments about this calculation procedure are in order. First, the errors committed in the control condition seem to be ignored, so that only the errors committed in the experimental condition are part of the difference score. Hence, the final score does not properly reflect the accuracy difference between the two conditions. Second, the final score does not only express the difference between the two conditions but also the variability within the experimental condition (not the variability within the control condition). Third, assume that there is absolutely no difference in performance between the two conditions (repetition and switch): same mean RT, same PE, and same RT distribution. If the score expresses the performance difference between the two conditions, it should be zero (or very close to zero). In fact, the score will still be substantial. This can be seen if at step 2, instead of the difference between experimental correct RTs and average control RT, the differences between the correct control RTs and the average control RT are calculated. Because there is variability within the control condition, the final score will not be zero, but will express this variability. Fourth, consider two experiments, one based on 100 trials and another based on 200 trials per condition. The Bin scores obtained in the former experiment will be much smaller than the Bin scores in the latter experiment, because the score is not based on proportions of responses but on absolute numbers of responses.

This inspection of the calculation procedure suggests that the Bin-o score does not yield a fair estimate of the performance difference between the two conditions. However, based on these observations, the calculation procedure can be adapted to provide a score with potentially better measurement properties. Here, two adaptations are proposed. The first adaptation involves the following changes:

1. At step 1, instead of taking all RTs of the control condition, *include only the correct RTs* to calculate the average of the control condition.
2. At step 2, calculate the difference between all correct RTs over *all trials* (both control and experimental conditions) and sort them in ten bins.
3. At step 3, first take the participant's RT differences of the control condition and count the number per bin; count the number of error trials in the control condition and assign them to be the bad bin. Calculate the overall score as defined in step 3 of the original procedure, and take this as the integrated performance measure of the participant in the control condition. Next, do the same for the experimental condition. Instead of a difference score, now two scores are obtained, one for each condition. A difference score can be obtained by subtracting the control score from the experimental score, but as also argued by Hughes et al., this has a number of disadvantages (amongst others, larger variance) and is better avoided.

This adapted procedure, which will be referred to as Bin-a (bin adapted) addresses the first three potential drawbacks mentioned in response to the procedure to calculate Bin-o: the errors in the control condition are no longer ignored, the RT variability in the control condition is accounted for, and if the control condition and the experimental condition yield the same or very similar scores, it is possible to infer the absence of a difference between the two conditions. One issue remains: if there is a difference in the number of trials between conditions, the scores will not be comparable.

The latter remark can be addressed in a further adaptation of the calculation procedure, namely by adding a step in which for each subject the response count per bin is converted to a proportion per bin. No other changes are needed. This score will be referred to as Bin-p (bin proportional).[2]

Thus far, the measures considered are IES, RCS, Bin-o, Bin-a, and Bin-p. Whereas the binning scores are linear combinations of RT and PE measures, IES and RCS are in fact non-linear measures; these scores are not the result of a linear combination of RT and PE. Considering that in cognitive psychology mostly statistical procedures are used that are based on the (general) linear model, a new integrated measure is proposed here that is based on a linear combination of RT and PE. This linear integrated speed-accuracy score (LISAS) is defined as:

$$\text{LISAS} = \text{RT}_j + \frac{S_{\text{RT}}}{S_{\text{PE}}} \times \text{PE}_j \qquad (3)$$

[2] In these adapted calculation procedures, the weight or penalty assigned to the error bin remains the same as the penalty used by Hughes et al., namely 20. In fact, the choice of the penalty is arbitrary. Exploration with different values of the penalty suggests that there is no optimum and that the impact of the penalty largely depends on specific characteristics of the sample being considered, e.g., the variability of error responses, the covariance between RT and PE measures, the presence of speed-accuracy trade-offs, etc. For these reasons, the penalty was kept at 20, and this value was used in all calculations of Bin scores in this article.

where RT_j is the participant's mean RT in condition j, PE_j is the participant's proportion of errors in condition j, S_{RT} is the participant's overall RT standard deviation, and S_{PE} is the participant's overall PE standard deviation. Weighting of the PE with the ratio of the RT and PE standard deviations is done to achieve a similar weight of the two components, RT and PE. Like IES, this measure yields an estimate of RT corrected for the number of errors. As an example, consider an average correct RT of 500 with a standard deviation of 100, and an error rate of .10 with a standard deviation of .05, LISAS will be 500 + 100/.05 × .1 = 700.

The properties of these six integrated measures will be evaluated in three studies based on Monte Carlo simulations. The first study focuses on the alleged shortcoming of Bin-o and the potential advantages of Bin-a and Bin-p. In a preview, this study will confirm that Bin-o is not a useful measure. Because selection of appropriate measures at least in part depends on their distributional properties and how well these fit the assumptions of normality often required in statistical procedures, the next study covers these distributional properties, and was based on a large sample of artificial data. As it is well known that the balance between RT and PE can be deliberately modified by speed-accuracy trade-off strategies, Study 3 focuses on the role of speed-accuracy trade-offs in a variety of cases with similar or different effects for RT and PE.

Study 1

In view of the alleged drawbacks and the potential to improve measurement with the binning method, the first study was designed to evaluate these issues. On the basis of a set of artificial data, this study tested whether it is indeed the case that the Bin-o measure fails to adequately and fully represent the size of the difference between the control and experimental conditions. As the two adapted measures are proposed to overcome these shortcomings, it may be expected that these measures efficiently capture the performance difference between the experimental and the control condition. However, as the Bin-o and Bin-a measures are based on absolute numbers of observations, it is also expected that both measures would yield different results solely due to the number of observations or trials within the conditions. In contrast, Bin-p is based on proportions and should not be affected by the number of observations per condition. All these different aspects were implemented in a single study, by means of a design encompassing a within-subject factor representing a comparison between a control condition (which could be task repetition, congruence, easier task, etc.) and an experimental condition (task switch, incongruence, difficult task, …), and two between-subject factors. One of the latter factors represented the presence or absence of a performance cost, and the other factor involved a variation of the number of trials in each of the conditions (high versus low number of trials in both conditions).

Method

Artificial data were generated for a 2 (Cost absent or present) × 2 (Number of trials: high or low) × 2 (Trial type: control vs. experimental) factorial design with repeated measures on the last factor. For the within-subject part of the design, the following structural model was defined:

$$X_{ij} = \mu + \alpha_i + \pi_j + \varepsilon_{ij} \tag{4}$$

where μ is the overall performance mean, α_i refers to trial type (control vs. experimental), π_j refers to the preferred performance level of statistical subject j, and ε_{ij} refers to the error term. For the generation of the RT data, $\mu = 500$, the value of π_j was sampled from a Gaussian distribution with zero-mean and standard deviation 100, and the value of α_i was zero in the cost-absent condition, and −10 (control trials) or +10 (experimental trials) when a cost was present. As RT data typically have a positive skew, the error term was generated from an exponentially modified Gaussian distribution, also known as an ex-Gaussian distribution, which is obtained as the convolution of a Gaussian and an exponential distribution (Heathcote, Popiel, & Mewhort, 1991; Ratcliff, 1979; Ratcliff & Murdock, 1976), which is defined as

$$f(t|\mu;\sigma;\tau) = \frac{1}{\tau(2\mathit{\Pi})^{1/2}} e^{-\left(\frac{\sigma^2}{2\tau^2}+\frac{t-\mu}{\tau}\right)} \times \int_{-\infty}^{(t-\mu)/\sigma-(\sigma/\tau)} e^{\left(-y^2/2\right)} dy \tag{5}$$

where t is the time, μ and σ are parameters of the gaussian distribution, and τ represents the mean and standard deviation of the exponential distribution. The mean of the ex-Gaussian distribution is $\mu + \tau$, its variance is $\sigma^2 + \tau^2$, and its skewness is $\frac{2\tau^3}{(\sigma^2+\tau^2)^{3/2}}$; μ represents the modus of the distribution. Ex-Gaussian distributed random numbers were generated by taking the sum of a Gaussian distributed random value N(0,1) and an exponentially distributed random value. The latter was obtained from a uniformly distributed value u that was transformed according to the following formula:

$$-\ln(u) \times \tau \tag{6}$$

with $\tau = 1.5$.

For the generation of the PE data, $\mu = .10$, π_j was sampled from a Gaussian distribution with standard deviation .04, and the value of α_i was zero in the cost-absent condition and ±.05 when a cost was present. A uniformly distributed random value between 0 and 1 was then sampled and compared to the sum $\mu + \alpha_i + \pi_j$ to decide whether the current response was correct (0) or incorrect (1). In the conditions with many trials, 260 trials were registered (130 per trial type). Only 130 trials (65 per trial type) were registered in the conditions with few trials. Each of the cells of the 2 × 2 between-subjects part of the design contained 30 statistical subjects.

Results and discussion

The descriptive statistics of the obtained sample of artificial data are displayed in Table 1. This table contains the means and the standard deviations of RT, PE, Bin-o, Bin-a, and Bin-p per cell of the design. As is shown in Table 1, the RT and PE means did not differ much between the Control and Experimental conditions in the Cost Absent conditions, and this was also the case for the Bin-a and Bin-p means. In contrast, these four measures yielded performance differences between the two trial types in the Cost Present conditions. In the Cost Absent condition, the Bin-o scores were large and only slightly smaller than in the Cost Present conditions. An effect of the number of trials per condition was only present in the Bin-o and Bin-a measures. These observed trends were tested by means of a 2 (Cost Presence) × 2 (Number of Trials) × 2 (Trial Type) ANOVA applied to each measure separately (except for Bin-o where the factor Trial Type was not available, as this is a difference score). The results of these analyses are shown in Table 2.

Three measures showed exactly the same pattern of results, namely RT, PE, and Bin-p, with a significant main effect of Trial Type and interaction of this factor with Cost Presence, while none of the other effects attained significance. In other words, for these three measures, the difference between control and experimental trials was significant, but only in the conditions where a trial type cost was present. These measures did not vary with the number of registered trials. The pattern was different for Bin-a: in addition to these same two effects, Number of Trials was significant and interacted with these two significant effects, thus producing a significant interaction of Number and Trial Type and a significant triple interaction. In other words, like RT, PE, and Bin-p, the Bin-a measure was only sensitive to the contrast between experimental and control trials when a trial type cost was present. But in contrast to these three measures, Bin-a was sensitive to the number of trials resulting in higher scores in conditions with more trials and a larger trial type cost when more trials were included.

For Bin-o, only a 2 × 2 design was applicable, and this analysis revealed significant main effects of Cost Presence and Number of Trials, but their interaction was not significant. This shows that Bin-o scores were larger when a cost was present (M = 1507) than no cost was present (M = 1374). Similarly, Bin-o scores were also higher when the number of trials was larger (M = 1951) than when it was smaller (M = 929).

To further assess whether the combined score obtained with the binning procedure captures the differences present in both RT and PE measures, linear multiple regression analyses were conducted with RT and PE measures as predictors and each of the binning measures in turn as dependent variable. As the binning scores are based on a sum of weighted RT differences additively combined with weighted PE, linear regression is a suitable technique for testing the relationship between these measures and the RT and PE measures. The RT and PE averages per trial type of all statistical subjects in the sample were the predictors of the difference between the experimental and control trials as measured by the binning measures. Table 3 displays the results of these regression analyses: for each bin measure (dependent variable), the *t*-values associated with each predictor are displayed together with the multiple correlation coefficient and the coefficient of determination.

In line with the critical comments formulated in the introduction regarding Bin-o, it appears that the PE of the control

Table 1 Means (standard deviations between brackets) of the measures response time (RT), proportion of errors (PE), Bin-o, Bin-a, and Bin-p as a function of the cells of the design of the simulated 2 (Cost Presence) × 2 (Number of trials) × 2 (Trial type) factorial design in Study 1

	Cost absent				Cost present			
	Many		Few		Many		Few	
	C[a]	E	C	E	C	E	C	E
RT	599 (101)	596 (100)	572 (102)	568 (101)	578 (107)	600 (108)	562 (129)	586 (130)
PE	0.105 (.035)	0.113 (.045)	0.104 (.036)	0.106 (.046)	0.053 (.046)	0.149 (.049)	0.050 (.042)	0.154 (.046)
Bin-o[b]	1,879 (412)		869 (198)		2,024 (394)		989 (235)	
Bin-a	1,860 (452)	1,878 (412)	871 (208)	869 (198)	1,569 (431)	2,024 (393)	741 (277)	989 (235)
Bin-p[c]	7.15 (1.74)	7.22 (1.58)	6.70 (1.60)	6.69 (1.52)	6.03 (1.66)	7.79 (1.51)	5.70 (2.13)	7.61 (1.81)

[a] The letters C and E refer to the control and experimental trial types

[b] The Bin-o measure expresses a difference between the control and the experimental condition and has therefore only one value for the factor of Trial type

[c] Note that the Bin-p averages can be derived from the Bin-a averages by dividing the latter by the number of trials (260 in the conditions with many trials and 130 in the conditions with few trials)

Table 2 Results of the analyses of variance applied to the measures response time (RT), proportion of errors (PE), Bin-o, Bin-a, and Bin-p in Study 1 on the basis of a 2 × 2 × 2 factorial design. The table displays the value of the *F*-test, its probability level, and the effect size expressed in partial eta-squared

Effects	RT			PE			Bin-o			Bin-a			Bin-p		
	F	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Cost Presence (A)	<1	>.9	0.0	<1	>.4	0.0	4.89	0.05	0.04	<1	>.5	0.0	<1	>.6	0.0
Number (B)	1.07	>.3	0.01	<1	>.8	0.0	288.9	0.001	0.71	237.8	0.001	0.67	1.45	>.2	0.01
Trial Type (C)	90.5	0.001	0.44	329.4	0.001	0.74				363.1	0.001	0.76	330.9	0.001	0.74
A x B	<1	>.75	0.0	<1	>.7	0.0	<1	>.8	0.0	<1	>.5	0.0	<1	>.6	0.0
A x C	170.5	0.001	0.60	271.8	0.001	0.79				330.2	0.001	0.74	310.5	0.001	0.73
B x C	<1	>.7	0.0	<1	<.9	0.0				36.1	0.001	0.24	<1	>.7	0.0
A x B x C	<1	>.5	0.0	1.2	>.25	0.01				24.6	0.001	0.18	1.37	>.2	0.01

Note All *F*-values have 1 and 116 degrees of freedom. As the Bin-o measure yields a single score for the effect of Trial Type, the design was reduced to a 2 × 2 between-subject design

condition is indeed not related to the Bin-o score at all. Furthermore, neither the experimental RT value nor the control RT value significantly contributes to the Bin-o difference score in this sample. In fact, the Bin-o score depends to a large extent on the proportion of errors committed in the experimental or difficult condition as these are heavily penalized. One could have expected that the RT scores of the experimental trials contributed more as their difference to the overall mean control RT mean is taken into account in the calculation, but this expectation was not confirmed in this analysis. To further explore this observation, another regression analysis was performed in which the RT and PE difference score between the experimental and the control condition were used as predictors of the binning measures. For Bin-o neither of these differences contributed to the prediction; the multiple regression coefficient was only .08. In Bin-a (R = .90) and Bin-p (R = .99) both cost predictors strongly contributed to the final score.

This first study fully corroborates the alleged shortcomings of Bin-o as an integrated measure of performance. The Bin-o measure does not validly combine RT and PE information and therefore it should never be used. The adaptations implemented in Bin-a and Bin-p, on the contrary, seem to work. Both measures validly combine RT and PE scores, but Bin-a suffers from the drawback that its value varies with the number of observations. By taking proportions of observations instead of absolute numbers, Bin-p seems to capture well the two performance components without any of the drawbacks observed with the two other binning variants.

Table 3 Results of the regression analyses with response time (RT) and proportion of errors (PE) scores on control and experimental trials as predictors of the cost scored by the binning measures: *t*-values, and coefficient of determination (R^2)

	RT-C	RT-E	PE-C	PE-E	R^2
Bin-o	0.51	0.38	0.01	2.80	0.35
Bin-a	−3.64	3.50	−11.52	12.89	0.82
Bin-p	−13.69	12.92	−41.58	46.37	0.98

Note The columns RT-C, RT-E, PE-C, PE-E contain the *t*-values (df = 115) of, respectively, RT in control, RT in experimental, PE in control, and PE in experimental trials as predictors of the cost observed on the measures in the rows (for Bin-o, the binning score; for Bin-a and Bin-p the difference of these scores in the control and experimental trials). The probability of all reported *t*-values was < .001, except for the *t*-values with respect to the Bin-o score, where only the PE-E predictor reached significance ($p < .01$)

On the basis of the findings in this first study, the Bin-o measure was excluded from any further evaluations. It should be noted further that if the number of trials in the conditions is kept constant, no distinction between Bin-a and Bin-p is possible because it can be shown that in every condition Bin-p equals Bin-a divided by the number of trials in the condition. For that reason, as the following studies in the paper did not vary the number of observations per condition, Bin-a and Bin-p were treated as the same measure. Its utility was further examined together with the three other integrated measures. Because most statistical methods that are used in data analysis are based on particular assumptions about the distribution of the measurements being analyzed, the next study examined the major characteristics of the probability distribution of the four remaining integrated measures.

Study 2

It is well known that both RT and PE measures show deviations from the normal distribution; within samples obtained from one or more subjects (i.e., the sample distribution) RTs are usually positively skewed and the shape of the PE distribution tends to vary with the average percentage of errors. When these measures are combined to form an integrated measure, it is possible that the integration results in even stronger deviations from the normal distribution. In order to examine the properties of the four remaining measures, two artificial data sets were generated on the basis of the model described in Equation 4. Because it is possible that the results differ with the direction of the effects, two subsets were produced, one with larger RT and larger PE in the experimental than in the control condition (effects in same direction) and one with RT and PE effects in opposing directions, namely larger RT and smaller PE in the experimental than in the control condition.

Method

Two data sets of 1,000 samples based on 20 statistical subjects were generated on the basis of Equation 4 with 250 observations in the control condition and 250 observations in the experimental condition. In one data set the RT and PE effects were in the same direction with longer RTs and more errors in the experimental condition; in the other data set, the RTs were also longer in the experimental condition, while the PEs were larger in the control condition. For generating the RT data, α_i was ±10, μ was 500 and π_j was sampled from a normal distribution with zero mean and a standard deviation of 100. Random error (σ_ε) was sampled from an ex-Gaussian distribution based on the sum of a Gaussian distributed value N(0, 1) and an exponentially distributed value with $\tau = 1.5$; the obtained value was then multiplied by the standard deviation of the RT error distribution (100). Thus an RT value was produced for each trial.

For the PE data, α_i was ±.0145, π_j was sampled from a normal distribution with zero mean and a standard deviation of .04. Because at the trial level errors are absent (0) or present (1), the generated value was compared to a uniformly distributed random number to decide whether the trial was correct or not. The RT and PE values for α_i were selected in such a way that the obtained effect sizes for the RT and PE variables were on average about equal.

Results and discussion

Mean, standard deviation, and skewness were calculated for each measure separately per condition within each of the samples of both data sets. Within both data sets, the values of these statistics varied over the samples. Averages of these statistics over the 1,000 samples in each data set yield an estimate of these statistics in the sample of samples (i.e., the sampling distribution). These averaged results are shown in Table 4. However, the question addressed in this study concerns the sample distributions and, in particular, the skewness of each measure in the samples. Therefore, Table 4 also displays the 95 % confidence interval (CI) of the skewness over all the samples.

Table 4 shows that the skewness of the sample RT distributions varied from strongly negative to strongly positive, with an average near to zero, in both data sets. Due to averaging the RTs per subject, the RT distribution becomes more symmetric at the level of the sample, conforming to the central limit theorem. Instead of simply adding all the RTs, the Vincent adding procedure (for more details see Heathcote et al., 1991; Ratcliff, 1979) could have been used. However, as this is not a common practice, the standard methodology for aggregating data was used in these simulations.

Within the data set with effects in the same direction, in each sample and for each measure the means differed between the two conditions. Within the data set with opposing effects, clear differences were present for correct RT and all RTs, PE, and the Bin measures, but not for IES, RCS, and LISAS. In the latter three measures, the opposing effects seemed to balance each other out.

A first noteworthy observation concerns the *absence of clear differences* between the standard deviations of the integrated measures across the two samples. In other words, whether the composing effects go in the same direction or in opposite directions does not matter much for the standard deviations of the integrated measures. The other noteworthy observation relates to the variations in skewness of the distribution of the different measures. As already indicated, for RT, skewness varied widely over the samples, but was on average rather close to zero (little or no skewness). The PE distribution was positively skewed to similar extents in both data sets, with more samples showing positive skew than samples showing negative skew; overall, the deviation from zero was rather small. Within the integrated measures, skewness was close to zero for Bin-p, IES, and LISAS. Finally, RCS was more strongly positively skewed in both samples, with a small proportion of the samples showing some degree of negative skewness and the majority of samples showing large to very large degrees of positive skewness. Closer inspection of the data revealed that the cases with small or even negative skewness in RCS were obtained from samples with very strong positive RT skewness.[3] Figure 1 shows the shape of the distributions for each of the measures and confirms the presence of some asymmetry in RCS and PE.

It may seem surprising that even with small deviations from symmetry in the RT and PE measures, RCS is the only integrated measure showing an important degree of skewness. In fact, this feature is inherent in the way RCS is calculated, namely as the ratio of the proportion of correct responses to the response time. A simple example can clarify this. Consider three cases with the same proportion correct responses, namely 0.9 and with RTs of 0.6, 0.7, and 0.8 s; the respective RCS values are 1.5, 1.29, and 1.13. Although the RT difference

[3] Explorations with larger values of τ and with inclusion of an ex-Gaussian distribution for sampling the value of π_j have shown that samples with negatively skewed RT distributions are likely to occur. As far as RCS is concerned, a positive skewness only occurred in conjunction with negative skewness of the RT distribution.

steps have the same size (0.1), the steps from the smaller to the larger RCS values increase (0.16 for the step from the smallest to the central value and 0.21 for the next step). This example shows that with increasing RCS values the spread between the values increases even though no such difference in spread is present in RT and PE (or 1-PE). That the positive skew in RCS is related to the calculation is also confirmed by the fact that the distribution of 1/IES is positively skewed while the distribution of IES is not.

Irrespective of whether the integrated effects are in the same or in opposing directions, the overall means (as can easily be inferred from Table 4) and the standard deviations of the integrated measures are not affected much. Only with respect to skewness did some variations occur across the different measures. Of the two basic measures, only PE showed some small positive skewness, and like the RT measure, all integrated measures, except RCS, yielded a skewness close to zero. The RCS measure, in contrast, showed a stronger positive skewness.

An additional consideration concerns the binning measures: due to the way these measures are calculated, their properties are dependent on the complete sample because all the observations in the sample are used to define the bins, from which the scores per subject are derived. The latter property constitutes an important potential drawback for these measures, because the score of a particular subject in the context of one sample may be strongly different from the score based on the same performance but calculated in the context of another sample. Evidently, this should never be the case, because statistical applications assume that subjects are sampled and scored independently from one another.

Fig. 1 Histograms of the distributions of all the measures in the control condition of the data set with effects in the same direction based on all the observations in the 1,000 samples of the data set. Each bar in the histogram shows the frequency (scaled per 1,000) in intervals of 0.5 standard deviations; the central bar (0) displays observations in the interval 0.5 standard deviation below and above the mean. The distributions in the other experimental condition and in both conditions of the other data are almost identical, and are therefore not displayed ▶

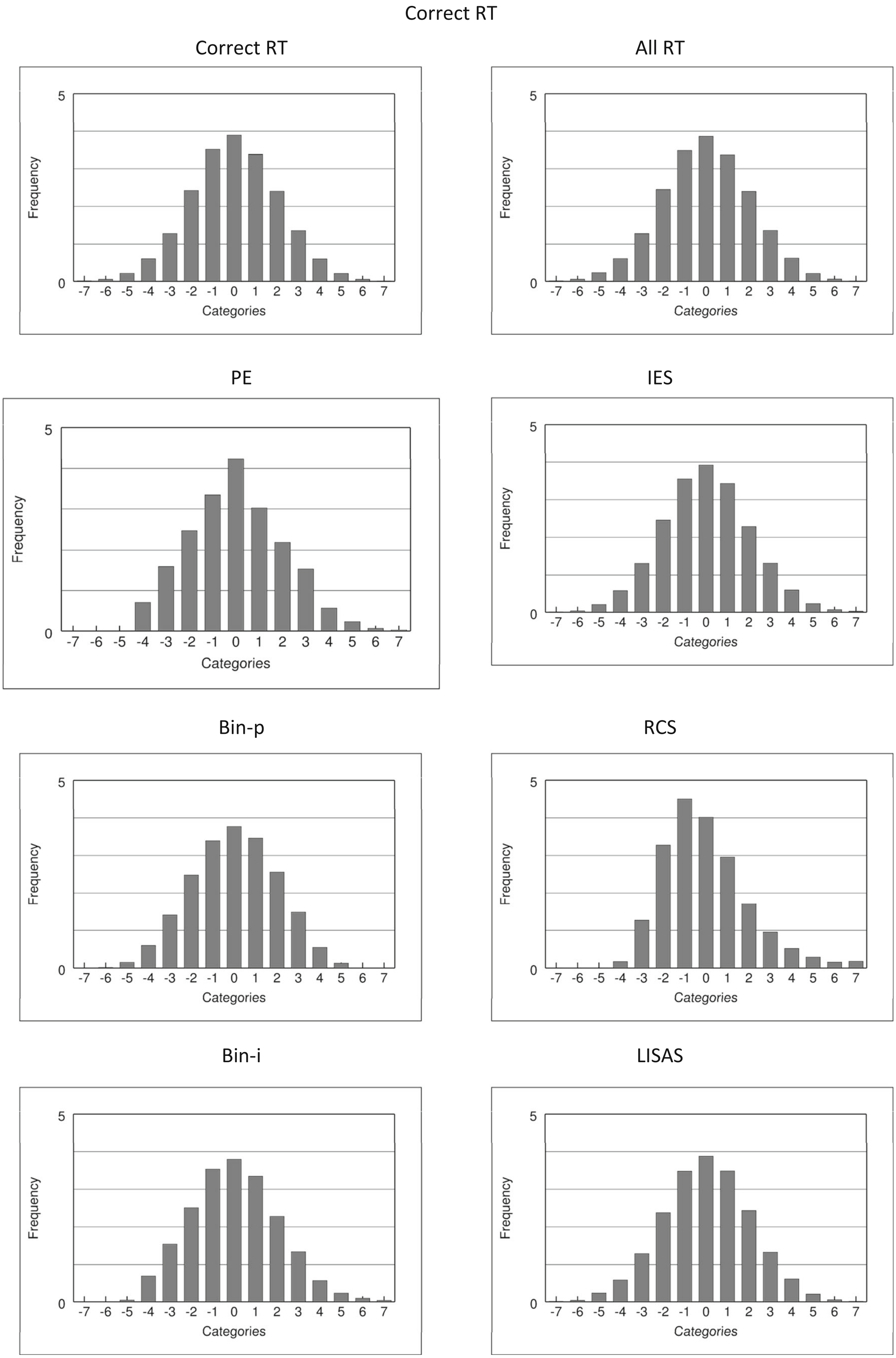
In order to test whether it would be practical to calculate bin-scores using only the RT and PE scores of the individual, the Bin-p scores for the two samples were calculated using only the data available per statistical subject. The findings are shown in Table 4 on the row labeled Bin-i (Bin-p, with individual-based scoring) and in Fig. 1 in the panel labeled Bin-i. Clearly, the averages were quite similar to those based on the complete sample and the standard deviations were much smaller. However, the individual-based scores yielded a small positive skewness (also visible in the slightly longer rightward tail in Fig. 1). To further explore the similarity between the individual-based and sample-based measures, product–moment correlations were calculated between the two sets of

Table 4 Averages of the means, standard deviations, and skewness values and 95 % confidence interval of the skewness values of response time (RT), proportion of errors (PE) Bin-p, IES, RCS, and LISAS in the control and the experimental condition in the two data sets of Study 2

		Same direction effects					Opposing direction effects				
		M	SD	Skewness			M	SD	Skewness		
				0.025	M	0.975			0.025	M	0.975
Correct RT	C[a]	640	100	−0.967	−0.001	1.076	640	100	−0.961	0.024	1.076
	E	660	100	−1.021	−0.004	1.101	660	100	−0.973	0.020	1.074
All RT	C	640	99	−0.974	0.001	1.052	640	100	−0.971	0.023	1.084
	E	660	99	−1.011	−0.005	1.134	660	100	−0.981	0.017	1.073
PE	C	0.086	0.043	−0.618	0.218	1.180	0.116	0.045	−0.827	0.165	1.144
	E	0.116	0.044	−0.776	0.141	1.097	0.087	0.043	−0.707	0.194	1.190
Bin-p	C	6.61	1.49	−0.870	−0.036	0.832	7.05	1.45	−0.880	−0.013	0.876
	E	7.30	1.45	−0.981	−0.111	0.794	6.89	1.48	−0.914	−0.068	0.821
Bin-i[b]	C	6.57	0.64	−0.647	0.204	1.157	7.00	0.67	−0.735	0.153	1.145
	E	7.31	0.65	−0.827	0.125	1.087	6.90	0.63	−0.731	0.180	1.109
IES	C	702	115	−0.926	0.062	1.127	726	119	−0.912	0.108	1.193
	E	748	120	−0.918	0.063	1.148	724	115	−0.975	0.095	1.116
RCS	C	1.47	0.26	−0.385	0.754	2.246	1.42	0.25	−0.394	0.709	2.373
	E	1.37	0.24	−0.416	0.738	2.229	1.42	0.24	−0.443	0.693	2.291
LISAS	C	689	102	−0.979	−0.008	1.063	709	101	−0.977	0.027	1.035
	E	728	102	−1.038	−0.012	1.111	710	102	−1.090	0.016	1.067

[a] The letters C and E refer to control and experimental conditions, respectively

[b] Bin-i refers to the Bin-p measure calculated on the data of one single subject instead of the data available in the complete sample



Correct RT

Correct RT

All RT

PE

IES

Bin-p

RCS

Bin-i

LISAS

scores over the entire sample. The correlations were very similar: .431 for effects in the same direction, .430 for effects in the opposing direction ($p < .001$). This suggests that there may be some promise in the usage of individual-based Bin scores, although it is not clear to which extent the individual-based scores are more noisy than the sample-based scores. In order to further explore these characteristics, the Bin-i measure was also included in the following studies.

In conclusion, Study 2 shows that the integrated measures closely approach the normal distribution. In the present study, RCS showed a rather strong positive skewness over the range of samples available. That may be a drawback in particular applications, but this is no reason to exclude this measure from the evaluation performed in the following two studies. As this skewness which is situated at the level of the sampling distribution was restricted to samples with a rather strong negative RT skewness, there is no impediment to the usage of the RCS measure as long as the distributions of RT, PE, and RCS are checked for symmetry. In fact, it is safe to perform these checks for any measure that is being used.

Study 3

After the assessment of the distributional properties of the integrated measures, the next study addresses the central question regarding the potential advantages of using integrated measures. This concerns two questions. First, is each of the integrated measures capable of detecting significant effects when RT and/or PE effects are present but vary in strength, and if so how good is this detection performance? Second, do the integrated measures account for a larger proportion of the variance than each of the composing measures do, and if so, to which extent? Importantly, some integrated measures of performance may be more sensitive to variations in the balance of the two components, RT and PE. For that reason speed-accuracy trade-off (SAT), effect size of the experimental-control contrast in RT and PE, and the direction of the RT and PE effects were varied over four Monte Carlo simulations: one with RT and PE effects in the same direction, one with no PE effects, one without RT effects, and one with RT and PE effects in opposing directions. The inclusion of the variation of the speed-accuracy trade-off strategy (trading speed for accuracy, no trade-off, or trading accuracy for speed) also has the additional advantage that the degree of covariance between speed and accuracy is varied.

Method

Artificial data were generated for a within-subject factor contrasting a control (or easy) and an experimental (or difficult) condition. The data were generated on the basis of the same structural model as in Studies 1 and 2 (see Equation 4). This within-subject factor was combined with a between-subject factor with three levels representing variations in the speed-accuracy trade-off strategy. The three trade-off strategies were: trade speed for accuracy, balanced speed and accuracy, and trade accuracy for speed. In the balanced condition, the structural model was applied independently to both measures resulting in near-zero correlation between RT and PE. Trading speed for accuracy was achieved by increasing the RT and decreasing PE after an error, while gradually reversing this effect after a correct answer. As short RTs occur less often with errors, this augments the correlation between RT and PE. Trading accuracy for speed was achieved in a similar way by increasing PE and decreasing RT after a correct answer, while gradually reversing this effect after an error. As incorrect answers are less often associated with longer RTs, this also leads to a positive correlation between RT and PE.

Thus, the complete design involved a 3 (SAT: speed-accuracy trade-off) × 2 (Trial Type) factorial combination with repeated measures on the last factor. The between-subject factor representing the three trade-off strategies was not expected to interact with the within-subject factor, because the strategies were applied irrespective of the trial type and are thus expected to have the same effects with the two trial types.

This 3 × 2 design was used in each of the four simulations that varied the RT and PE effects. These simulations involved 1,000 samples each. Every sample contained ten statistical subjects per SAT condition; each condition involved ten blocks of 65 trials (the first trial of each block was not included in the data analyses). In each statistical subject, the effect size implemented was sampled from a pre-specified range for the within-subject factor as specified in the structural model of Equation 4 (see Table 5 for details). In one simulation (Case A) the RT and PE effects occurred in the same direction; the second simulation (Case B) had variable RT effect sizes but no PE effects; the third simulation (Case C) contained variable PE effect sizes but no RT effects; the fourth and final

Table 5 Effect size parameters used in the simulations of Study 3

Parameter	Label	RT	PE
μ		500	0.10
α_i	Trial type	40	0.06
π_j	Subject	50	0.15
σ_ε^2	Error variance	19,600	0.05

Note Except for μ and σ_ε^2, the values specify the maximum absolute value the parameter could take. In every statistical subject, a value between 0 and the value given in the table was randomly selected from a Gaussian distribution for π_j, and from a uniform distribution for α_i. To satisfy the assumptions of a fixed effects model that $\sum\alpha_i = 0$, the sampled (positive) value was used for one level and the negative of the value was used for the other level. RT data were sampled from an ex-Gaussian distribution as in Studies 1 and 2

Table 6 Confidence interval and median of the absolute value of the effect sizes sampled for response time (RT) and proportion of errors (PE) in the four simulation cases in Study 3

	RT effect size			PE effect size		
	0.025	0.50	0.975	0.025	0.50	0.975
Case A	0.99	20.03	38.56	0.002	0.032	0.058
Case B	0.78	20.75	39.04	0	0	0
Case C	0	0	0	0.001	0.032	0.059
Case D	0.998	21.24	39.08	0.002	0.032	0.059

Note The table displays the 95 % percent confidence interval of the effect sizes sampled for the data generation in the four simulation cases. The values given are the points in the distribution with a probability of .025, .5 (median), and .975

simulation (Case D) included RT and PE effects in opposing directions. The distribution of the effect sizes used in each simulation case are shown in Table 6.

Results and discussion

Descriptive statistics

Before looking into the results relevant to the present research question, the data are summarized. The means and standard deviations of the sampling distributions collected in each of the four simulation cases are displayed in Tables 7, 8, 9, and 10 for the four variations in RT/PE effects (Cases A–D). These tables show the (correct) RT data and PE data, as well as the means and standard deviations of the integrated measures (Bin-p, Bin-i, IES, RCS, and LISAS). These tables show that the SAT conditions had the intended effects on both RT and PE with longer RTs and lower PEs than the neutral or balanced condition when speed was traded for accuracy. Likewise, RT was shorter and PE was larger than in the neutral condition when accuracy was traded for speed. The effect of Trial type was also clearly present but variable over the four simulation cases. For all five integrated measures, the tables show that in the four simulation cases, the averages were consistent with an effect of Trial Type, but the size of the difference between control and experimental condition varied over the four cases. Correlations between RT and PE varied between .56 and .67 ($p < .001$) in the trade-speed conditions, between .16 and .25 ($p < .001$) in the trade-accuracy conditions, and between −.03 and .02 ($p > .33$) in the neutral conditions.

In each of the 1,000 samples of the four simulation cases, the 3 (SAT) × 2 (Trial Type) design was subjected to analyses of variance. All these analyses were performed separately for each measure because the measures use different metrics. Table 11 displays the average effect size (partial eta-squared) related to Trial type for each of the integrated measures in all four simulation cases. The table shows that the effect sizes for RT and PE were quite similar in Cases A and D, while the PE effect size was near zero in Case B, and the effect size of RT was near zero in Case C. This confirms that the scheme used for the data generation worked fine. Interestingly, the average effect sizes of all the integrated measures were larger than those of both RT and PE when these effects were in the same direction (Case A), while they were lower than the largest effect size of either RT or PE in the other three cases. More specifically, the effect sizes of the integrated measures were lower than the effect size of the only effective component measure in the cases where only one component varied (RT in Case B and PE in Case C). It is also worthwhile to note that in Case D the integrated measures still succeeded in achieving some explanatory power, although it could be expected that in the case with opposite RT and PE effects, these effects would balance each other out.

Efficiency of RT and PE integration

By comparing significant effects in RT and PE on the one hand, and an integrated measure on the other hand, it is possible to assess how well the integrated measure is capable of detecting an effect, given that the effect is present with some

Table 7 Means (standard deviations between brackets) of the sampling distributions based on 1,000 samples for response time (RT), proportion of errors (PE) Bin-p, Bin-i, inverse efficiency score (IES), rate correct score (RCS), and linear integrated speed-accuracy score (LISAS) in the simulation case with RT and PE effects in the same direction (case A)

	Trade speed		Neutral		Trade accuracy	
	Control	Experimental	Control	Experimental	Control	Experimental
RT	751 (27)	795 (27)	691 (20)	731 (21)	421 (21)	460 (21)
PE	.076 (.031)	.121 (.037)	.104 (.038)	.149 (.042)	.178 (.044)	.234 (.047)
Bin-p	7.63 (0.55)	8.53 (0.60)	7.35 (0.57)	8.34 (0.61)	6.22 (0.73)	7.40 (0.77)
Bin-i	6.39 (0.48)	7.42 (0.54)	6.77 (0.57)	7.85 (0.62)	7.87 (0.65)	9.05 (0.67)
IES	828 (58)	929 (72)	787 (46)	882 (58)	530 (48)	629 (62)
RCS	1.22 (.08)	1.11 (.08)	1.30 (.07)	1.17 (.07)	1.95 (.15)	1.67 (.13)
LISAS	825 (44)	903 (46)	757 (29)	833 (31)	538 (34)	613 (34)

Table 8 Means (standard deviations between brackets) of the sampling distributions based on 1,000 samples for response time (RT), proportion of errors (PE) Bin-p, Bin-i, inverse efficiency score (IES), rate correct score (RCS), and linear integrated speed-accuracy score (LISAS) in the simulation case with RT but no PE effects (case B)

	Trade speed		Neutral		Trade accuracy	
	Control	Experimental	Control	Experimental	Control	Experimental
RT	754 (26)	794 (27)	689 (21)	730 (21)	419 (20)	459 (21)
PE	.098 (.032)	.099 (.032)	.124 (.038)	.124 (.038)	.203 (.042)	.203 (.042)
Bin-p	7.90 (0.56)	8.27 (0.53)	7.61 (0.57)	8.00 (0.55)	6.63 (0.71)	6.90 (0.70)
Bin-i	6.69 (0.49)	7.12 (0.46)	7.05 (0.57)	7.50 (0.55)	8.23 (0.64)	8.62 (0.61)
IES	856 (61)	901 (62)	805 (49)	852 (50)	548 (48)	599 (52)
RCS	1.20 (.08)	1.14 (.07)	1.28 (.07)	1.21 (.06)	1.91 (.15)	1.74 (.13)
LISAS	842 (44)	883 (44)	771 (30)	811 (29)	552 (32)	592 (33)

strength in the RT and PE measures. As the size of each effect in RT and PE varies over the samples, it may be expected that as the effect size of RT and PE in the sample considered increases, the probability for an integrated measure to detect the effect also increases. This capability to detect effects as significant given effects of a particular size in the RT and PE data, will be referred to as the detection efficiency, which can be defined as

$$E = \frac{N_i}{N_e} \tag{7}$$

where E is the efficiency of detection, N_e is the number of samples in which the RT and PE measures find an effect of some size e, and N_i is the number of samples in which the integrated measure detects a significant effect. The degree of efficiency is expected to increase with the strength of the relationship, i.e., the effect size present in RT and PE. Five levels of effect size for RT and PE were defined using partial eta-squared (η_p^2) by partitioning the samples into five subsets. The first level contains the samples in which both effect sizes were smaller than .2 (this is the only level where nonsignificant effects could occur). The second level contains the samples in which both effect sizes were smaller than .4 after excluding the samples of level 1. Similarly, levels 3, 4, and 5 contain the samples with both effect sizes smaller than, respectively, .6, .8, and 1.0, after excluding the samples already assigned to the previous levels. As an example, when in a particular sample the effect size for RT is .42 and the effect size for PE .24, this sample will be assigned to level 3 (both are too large for level 1, only the PE effect satisfies the criterion for level 2, but both satisfy the criterion for level 3). The number of samples at each level and the average effect size over both measures in each sample are displayed in Table 12 for each of the four simulation cases. This table shows that the average effect size increases as the level increases. As can be expected from the example given, the averages tend to be lower than the nominal limits used to define the partitioning because the effect sizes of PE and RT are randomly sampled.

The detection efficiency of the integrated measures was calculated for each of these effect size levels by taking the ratio of the number of samples in which the integrated measure detected a significant effect ($p < .05$)[4] and the number of samples included at that level in line with Equation 7. Figure 2 displays the detection efficiency of the five integrated measures in the study as a function of the effect size levels of the *between-subject effect* for each of the four simulation cases.

The pattern of findings was very similar across the four cases (the four panels of Fig. 2). Most striking is the observation that the trend of the Bin-i measures deviated from the trend displayed by the other measures: in contrast to the other integrated measures the trend for Bin-i detection efficiency was to decrease with the level of the RT and PE effect size. As can be seen in Tables 7, 8, 9, and 10, the Bin-i averages tended to be larger in Trade-accuracy than in the Trade-speed condition of the SAT manipulation, suggesting that the gain in accuracy compensated the loss of speed, but the gain in speed did not seem to compensate the loss in accuracy in this integrated measure. As this measure is calculated on the basis of bins derived from the individual's data, the end result may be less reliable than the end result obtained in the Bin-p measure, which is derived from sample-based bins. Interestingly, Bin-p, IES, and RCS showed an increased probability of detecting a significant SAT effect as its effect size in RT and PE increased whereas LISAS showed 100 % detection of the effect at each effect size level. It is also noteworthy that LISAS, RCS, and IES were able to detect the effect with an overall probability of 1.00, 0.99, and 0.96, respectively, but Bin-p was far less efficient (overall probability less than .30). It thus seems that the Bin measures were not very efficient at detecting the between-subject effect. Overall, the Bin-p detection

[4] In fact, only effect sizes smaller than .14 (for F values with 1 degree of freedom in the numerator, namely Trial type) or .20 (for F values with 2 degrees of freedom in the numerator, namely SAT and its interaction with Trial type) were nonsignificant.

Table 9 Means (standard deviations between brackets) of the sampling distributions based on 1,000 samples for response time (RT), proportion of errors (PE) Bin-p, Bin-i, inverse efficiency score (IES), rate correct score (RCS), and linear integrated speed-accuracy score (LISAS) in the simulation case with PE but no RT effects (case C)

	Trade speed		Neutral		Trade accuracy	
	Control	Experimental	Control	Experimental	Control	Experimental
RT	772 (23)	776 (23)	710 (17)	710 (17)	439 (18)	440 (18)
PE	.077 (.029)	.123 (.037)	.102 (.037)	.149 (.041)	.173 (.044)	.231 (.048)
Bin-p	7.84 (0.49)	8.39 (0.59)	7.51 (0.55)	8.15 (0.60)	6.28 (0.74)	7.20 (0.79)
Bin-i	6.62 (0.42)	7.23 (0.53)	6.96 (0.54)	7.63 (0.60)	8.01 (0.64)	8.81 (0.69)
IES	852 (53)	909 (68)	807 (45)	858 (54)	551 (47)	599 (58)
RCS	1.19 (.07)	1.14 (.07)	1.27 (.06)	1.21 (.07)	1.88 (.14)	1.76 (.15)
LISAS	846 (40)	884 (44)	774 (27)	813 (29)	553 (33)	591 (33)

probabilities were quite low, and the Bin-i detection rates dropped as the RT and PE effect size increased. IES, RCS, and LISAS were very efficient as they detected almost all the effects at all levels of effect size. Moreover, note that the pattern as well as the level of performance was the same in the four simulation cases for all the measures.

The detection efficiency of the five integrated measures with respect to the *within-subject effect* is shown in Fig. 3. In panel A (simulation Case A: RT and PE effects in the same direction), the detection efficiency of all five measures increased with the effect size level of RT and PE, and almost perfect detection performance was reached (.98–.99) as levels 1 and 2 contained only few cases. Panel B (simulation Case B: only RT effects) shows that the detection capability for each measure also increased with the effect size of RT and PE. However, large differences were observed among the measures. The best performance was achieved by LISAS (overall .83), RCS (.82), and IES (.78), while Bin-p (.71) and Bin-i (.75) attained lower levels but were still very efficient. For all measures, performance increased monotonically with effect size level. In Panel C (simulation Case C: only PE effects) all measures attained a high level of performance from level 3 on (at least .90). The spread between the curves was large at level 2. In this simulation case, overall best performance was achieved by Bin-p (.87) and Bin-i (.86), followed by LISAS (.83), IES (.80), and RCS (.78). All five measures attained a detection efficiency above .50 from level 2 on, except RCS (.48 at level 2). In Panel D (simulation Case D: opposing RT and PE effects), the pattern was quite different in shape from the three other panels with some crossing-over of the lines in the graph and a non-monotonic trend for all the measures, except RCS and LISAS. Averaged over all measures, detection rate was not higher at level 5 than at level 4. At these two levels, all five measures performed quite well (.60 detection or more), and achieved medium detection performance at level 3. Average detection rates were very close together with the highest score for Bin-p (.57 overall) followed by Bin-i (.54), and LISAS (.51); RCS (.47) and IES (.46) attained the poorest average. Given that in this particular case the RT and PE effects were in opposing directions, it may be considered that the performance of all five measures was at a remarkably good level.

Table 10 Means (standard deviations between brackets) of the sampling distributions based on 1,000 samples for response time (RT), proportion of errors (PE) Bin-p, Bin-i, inverse efficiency score (IES), rate correct score (RCS), and linear integrated speed-accuracy score (LISAS) in the simulation case with RT and PE effects in opposing directions (case D)

	Trade speed		Neutral		Trade accuracy	
	Control	Experimental	Control	Experimental	Control	Experimental
RT	755 (25)	791 (25)	690 (21)	731 (20)	419 (21)	459 (20)
PE	.121 (.035)	.075 (.030)	.150 (.041)	.103 (.035)	.235 (.047)	.176 (.042)
Bin-p	8.17 (0.58)	7.98 (0.50)	7.98 (0.61)	7.70 (0.53)	7.14 (0.79)	6.46 (0.70)
Bin-i	6.99 (0.52)	6.80 (0.44)	7.43 (0.62)	7.19 (0.52)	8.67 (0.70)	8.24 (0.61)
IES	882 (65)	870 (57)	834 (55)	830 (46)	574 (57)	577 (48)
RCS	1.18 (.08)	1.16 (.07)	1.24 (.07)	1.23 (.06)	1.84 (.16)	1.79 (.13)
LISAS	863 (44)	865 (41)	794 (31)	795 (29)	573 (34)	575 (33)

Table 11 Average (standard deviations within brackets) effect size (η_p^2) of Trial type obtained with response time (RT), proportion of errors (PE) Bin-p, Bin-i, inverse efficiency score (IES), rate correct score (RCS), and linear integrated speed-accuracy score (LISAS) in each of the four simulation cases (A-D) of Study 3

	Case A	Case B	Case C	Case D
RT	.67 (.28)	.65 (.30)	.04 (.05)	.64 (.30)
PE	.58 (.25)	.03 (.05)	.59 (.26)	.60 (.24)
Bin-p	.75 (.18)	.39 (.25)	.57 (.26)	.39 (.26)
Bin-i	.77 (.17)	.46 (.27)	.55 (.25)	.36 (.25)
IES	.69 (.17)	.50 (.28)	.43 (.22)	.30 (.24)
RCS	.76 (.18)	.54 (.27)	.45 (.25)	.32 (.25)
LISAS	.79 (.16)	.59 (.30)	.50 (.25)	.36 (.27)

Note Case A has RT and PE effects in the same direction, B has no PE effects, C has no RT effects, and D has RT and PE effects in opposing directions

Proportion of explained variance

Apart from knowing that the integrated measures are efficient in detecting an effect when an effect is present in one or both composing measures, it is also important to know whether an integrated measure accounts for a larger amount of the variance than each of the composing measures do. This was investigated by assessing the proportion of samples in which the effect size associated with Trial type was larger than the largest effect size of RT or PE. Table 13 displays these proportions for the five integrated measures in each simulation case. The table shows that, except for IES, the integrated measures achieved this in .48 (RCS) to .62 (LISAS) of the samples when the RT and PE effects were in the same direction. In the other simulation cases, the integrated measures rarely attained a higher effect size than RT and PE (up to .16 in Case B for LISAS; up to .29 in Case C for Bin-p; and not higher than .025 in Case D again for Bin-p). In other words, when only one of the two components had an effect (either RT or PE, Cases B and C) the integrated measures did not frequently achieve a much higher effect size than that single component, but when the effects were in opposite directions, the integrated measures very rarely achieved a higher effect size than the components did. It is nevertheless remarkable that the binning measures performed extremely well in Case C as well with respect to the proportion of samples in which a higher effect size than RT and PE was obtained as in the detection efficiency scores (Fig. 3). In combination with a rather poor performance in Case B, this may be attributed to the high penalty which is applied to errors in the binning measures, thus creating a bias towards detection of PE-related effects.

Summary

The present study varied the size and the direction of the effects of RT and PE while keeping the average effect sizes of these measures in balance. The efficiency of the integrated measures at detecting the RT and PE effects present was tested both at the level of the SAT manipulation (between subjects) and the difficulty of the task (within subjects). With respect to the SAT manipulation, the binning measures by and large failed to detect the very strong RT and PE effects, while the other integrated measures performed excellently (see Fig. 2). In contrast, all measures performed excellently in the detection of the within-subject effects in simulation Case A. They all detected an effect in more than 98 % of the samples and the obtained effect sizes were larger than the RT and PE effects in about half of the cases, except for IES (Table 13). Efficiency dropped dramatically but was still very good when only one of

Table 12 Average effect size (η_p^2) of response time (RT) and proportion of errors (PE) effects at each of the five effect size levels, and number of samples (between brackets) at each level in Study 3

		1	2	3	4	5
Case A	Between	.0 (0)	.0 (0)	.0 (0)	.49 (48)	.53 (952)
	Within	.07 (15)	.21 (28)	.38 (71)	.57 (346)	.73 (540)
	Interaction	.07 (829)	.17 (161)	.25 (10)	.0 (0)	.0 (0)
Case B	Between	.0 (0)	.0 (0)	.0 (0)	.49 (47)	.52 (953)
	Within	.05 (143)	.17 (86)	.28 (101)	.37 (201)	.46 (469)
	Interaction	.06 (905)	.16 (93)	.21 (2)	.0 (0)	.0 (0)
Case C	Between	.0 (0)	.0 (0)	.0 (0)	.48 (58)	.52 (942)
	Within	.05 (128)	.17 (87)	.28 (140)	.38 (498)	.45 (147)
	Interaction	.06 (827)	.17 (168)	.30 (5)	.0 (0)	.0 (0)
Case D	Between	.0 (0)	.0 (0)	.0 (0)	.49 (43)	.53 (957)
	Within	.07 (20)	.22 (31)	.37 (66)	.57 (341)	.72 (542)
	Interaction	.07 (815)	.17 (179)	.28 (6)	.0 (0)	.0 (0)

Note The numbers between brackets add up to 1,000 for all the effects

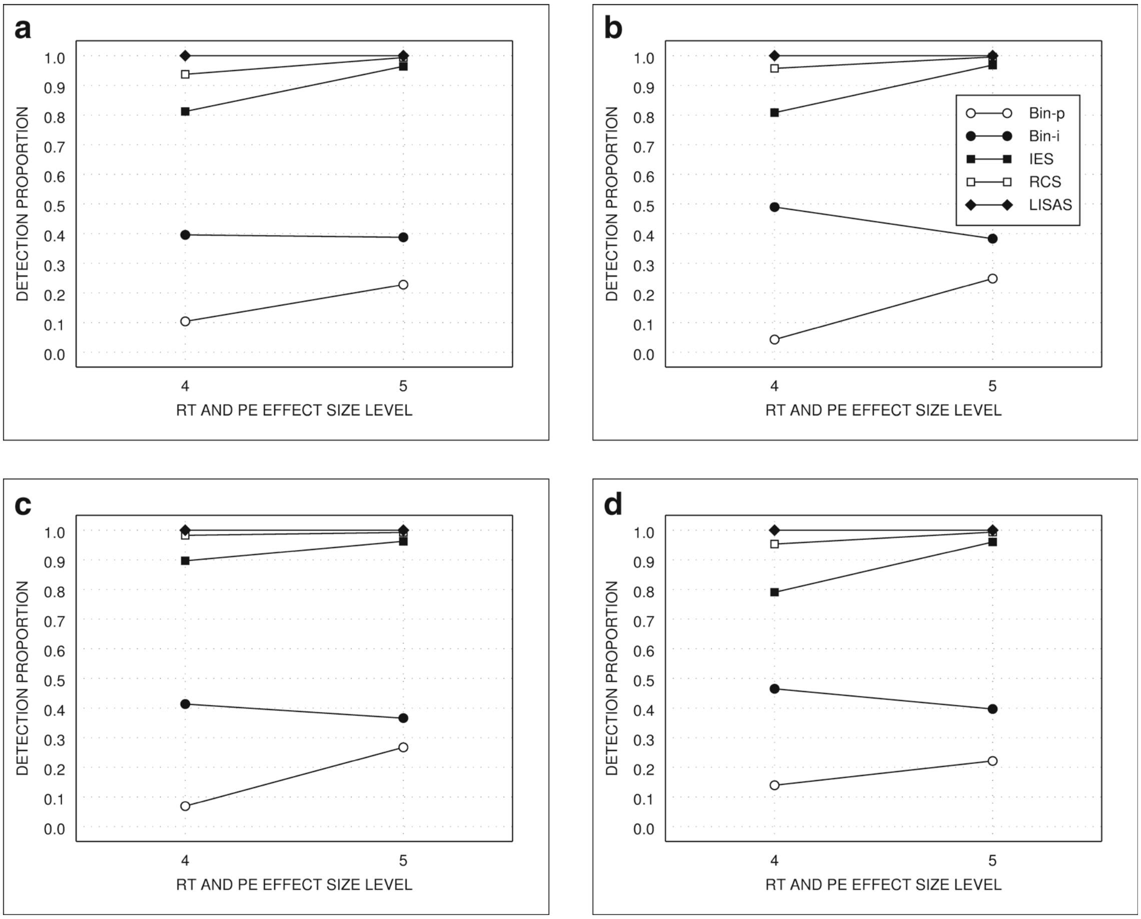


Fig. 2 Probability of detecting the between-subject effect of speed-accuracy trade-off (SAT) by each of the integrated measures as a function of the effect size of the response time (RT) and proportion of errors (PE) effect in the four simulated cases (**A**: RT and PE effects in the same direction; **B**: no PE effects; **C**: no RT effects, and **D**: effects in opposing directions). Five levels of effect size (η_p^2) were distinguished as explained in the text. As all the RT and PE effect sizes were at levels 4 and 5, only these levels are shown in the graph

the basic effects RT or PE was present; the frequency of larger effect sizes than those of either RT or PE were far less frequent but still substantial for RCS and LISAS in Case B (only RT) and for Bin-p, Bin-I, and LISAS in Case C (only PE). Finally, in Case D (opposing effects), detection efficiency was more variable, especially at higher RT and PE effect levels, and rarely ever an integrated measure achieved a larger effect size than RT and PE. Nevertheless, even in the latter case, most of the measures obtained quite good detection efficiency at the higher RT/PE effect sizes. It is possible, though, that this rather good performance is due to the fact that the effect sizes varied from sample to sample in such a way that a weak RT effect may be combined with a strong PE effect in the other direction or vice versa.

General discussion

Many experimental paradigms in cognitive psychology rely on both speed and accuracy of performance. To the extent that speed and accuracy are the result of common or overlapping processes, these paradigms could benefit from the availability of valid performance measures that integrate speed and accuracy aspects of performance. Following up on recent research about the advantages of such integrated measures, and in particular the binning measure proposed by Hughes et al. (2014), the present paper investigated the usefulness of some measures that integrate speed and accuracy of performance into one single score. Several integrated measures were compared in three studies, namely four measures based on the binning procedure, Bin-o (the original binning procedure proposed by Hughes et al.), Bin-a (an adapted version avoiding some of the potential shortcomings of Bin-o), Bin-p (a further adaptation based on proportions rather than absolute numbers), and Bin-i (the same as Bin-p but based on a binning of the subject's own data only), and three measures that combine RT and PE scores in a more direct way, namely IES (Equation 1; Townsend & Ashby, 1978), RCS (Equation 2; Woltz & Was, 2006), and a new linear combination labeled LISAS (Equation 3). All studies used artificially generated data to evaluate the usefulness of these measures under strictly controlled conditions. All the data were generated on the basis of a simple model based on a

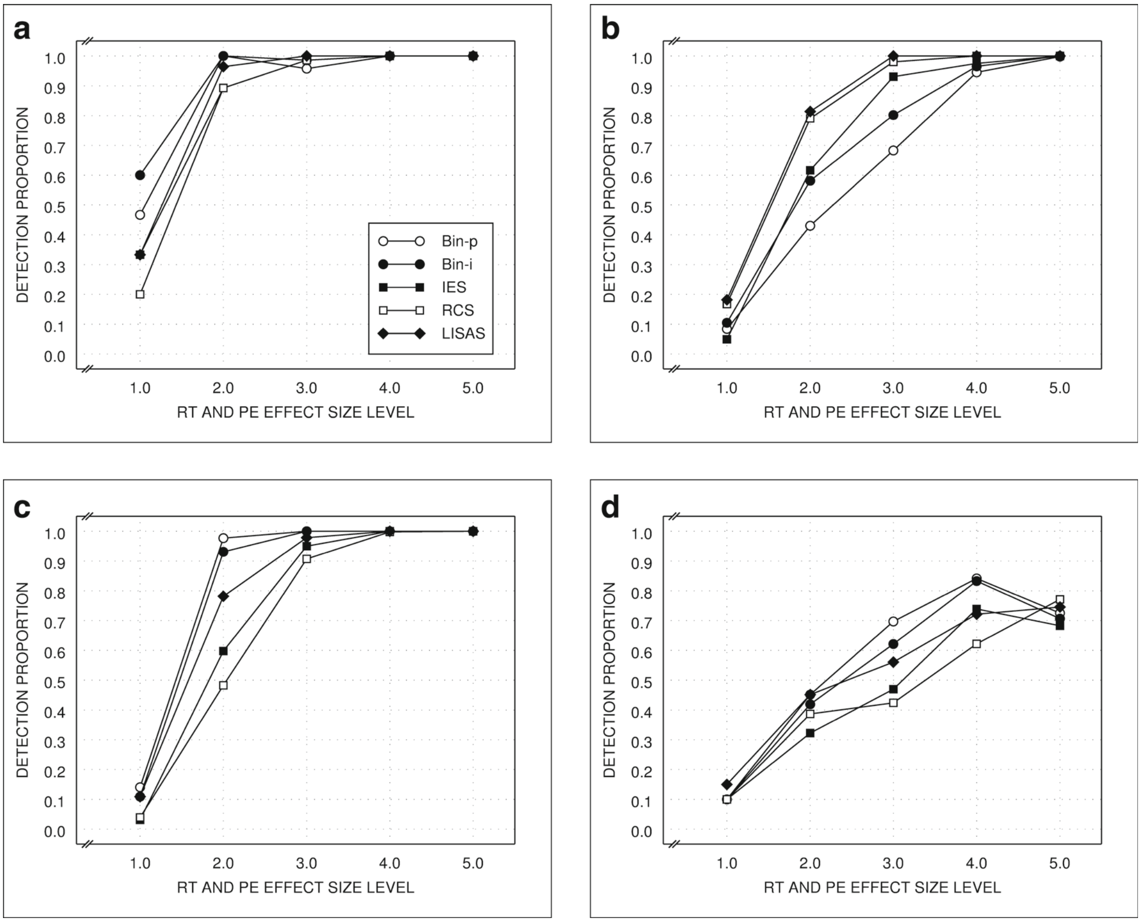


Fig. 3 Probability of detecting the within-subject effects by each of the integrated measures as a function of the effect size of the response time (RT) and proportion of errors (PE) effect in each of the simulated cases (**A**: same direction effects for RT and PE; **B**: no PE effects; **C**: no RT effects; and **D**: opposing effects). Five levels of effect size (η_p^2) were distinguished, as explained in the text

within-subject factor representing the comparison of a control and an experimental condition.

Properties of integrated measures: General summary

Table 14 summarizes the main findings. The table contains the answers to a series of simple yes/no questions about each integrated measure. The first question concerns the validity of the measures: do the measures provide a score based on information from both components, RT and PE? Although IES, RCS, and LISAS quite transparently combine RT and PE information, the question was raised because of a number of considerations that were provoked by the rather complex calculation procedure used for the binning measures. Study 1, which was designed to clarify this issue, showed that the Bin-o measure indeed lacks validity as only RT and PE information from the experimental condition is used to calculate a difference score between the control and the experimental condition. As a consequence, Bin-o is not capable of detecting the absence of a difference between the two conditions and if a difference exists the measure expresses only the occurrence of errors and the variability in speed within the experimental condition.[5] Hence, Bin-o is not a valid measure. It is remarkable that Hughes et al. (2014) report several studies without testing the construct validity of the measure. Study 1 also showed that the other binning measures did not suffer from these shortcomings and can basically be considered as

Table 13 Proportion of samples in which the integrated measures obtained a larger effect size for the factor control versus experimental condition than the maximum shown by the composing response time (RT) and proportion of errors (PE) measures

	Case A	Case B	Case C	Case D
Bin-p	0.487	0.039	0.287	0.024
Bin-i	0.553	0.058	0.195	0.014
IES	0.154	0.047	0.032	0.004
RCS	0.480	0.134	0.076	0.008
LISAS	0.619	0.160	0.126	0.015

[5] This may have implications for a correct interpretation of the results obtained in studies that use this particular integrated measure (Draheim, Hicks, & Engle, 2016).

Table 14 Summary of the findings of the seven integrated measures studied in the present paper: response time (RT), proportion of errors (PE) Bin-p, Bin-i, inverse efficiency score (IES), rate correct score (RCS), and linear integrated speed-accuracy score (LISAS)

Measure	Valid?	Symmetric?	Contrast efficient?	Other efficient?	Added value?	Balance?
Bin-o	−	NA	NA	NA	NA	NA
Bin-a	0	NA	NA	NA	NA	NA
Bin-p	+	+	+	−	+	−
Bin-i	+	+	+	−	+	−
IES	+	+	+	+	−	+
RCS	+	−	+	+	+	+
LISAS	+	+	+	+	+	+

The body of the table summarizes the properties regarding validity of the measure (Valid?), whether the sampling distribution is symmetric rather than skewed (Symmetric?), whether the measure efficiently detects a contrast when it is present in RT and PE (Contrast efficient?), whether the measure efficiently detects an effect in a variable outside the contrast (Other efficient?), whether the measure accounts for a larger part of the variance than the components (Added value?), and whether the two components are integrated in a balanced way (Balance?). Each cell contains either a minus sign (property not present), a plus sign (property present), a zero (property only partly present), or the indication "NA" (not applicable)

measures that combine RT and PE into a score that validly represents both components. However, Study 1 also showed that like Bin-o, the Bin-a measure varies with the number of trials in the experiment, which impedes fair comparisons across experiments or conditions with unequal numbers of trials. Fortunately, the usage of Bin-a can be avoided: in conditions with equal numbers the Bin-a scores vary linearly with the Bin-p scores which are based on the same data, except that the calculation is based on proportions rather than absolute numbers. For that reason, Bin-a was not further included in the present studies. Moreover, given that Bin-p captures the same information without limiting the comparability of the obtained scoring, Bin-a should be avoided as well.

The second question mentioned in Table 14 relates to the statistical properties of the (sampling) distributions of the remaining measures, i.e., after exclusion of Bin-o and Bin-a. Study 2 showed that overall mean, standard deviation, and skewness of the sampling distributions of RT, PE, and the integrated measures did not vary much between a variation with RT and PE effects in the same direction and a variation with RT and PE effects in opposing directions. This study also showed that the degree of skewness of all measures widely varied over the samples, with a bias toward positive skewness for RCS especially in samples with low positive and negative skew in the RT distribution. Although the asymmetry in the RCS distribution may be a matter of concern when it does occur, it is not a sufficient basis to completely reject this measure. Depending on the sample at hand, deviations from symmetry may occur in all measures.

In relation to the occurrence of asymmetry in the distributions of RCS and occasionally in some of the other measures, one could consider the possibility of always checking the skewness in the sample at hand. Unfortunately, the skewness measure is quite sensitive to outliers and is in general only stable when it is calculated on a very large number of observations. Hence, estimates of the degree of skewness in samples of the size usually taken in cognitive research may be expected to be rather unreliable, and consequently there is not much need for concern, except for samples that yield clearly asymmetric distributions.

The three next questions displayed in Table 14 were tested in Study 3 and they concern two important properties of integrated measures, namely: (1) to what extent do they recover the effects present in the component measures, and (2) do they account for more of the variance than each of the components and to what extent? These two properties correspond to the detection efficiency of these measures (Equation 7) and to the degree to which they account for more of the variance than the component measures do. Two of the questions in Table 14 address detection efficiency. The first of these questions concerns the efficiency with which the control-experimental contrast is detected conditional on the presence of a contrast effect in RT and/or PE. The table shows that all five integrated measures are quite successful in this respect. Indeed Fig. 3 shows that irrespective of whether the RT and PE effects are in the same or opposing directions, or whether one of these effects is absent, all the integrated measures report an integrated effect and the likelihood of detection generally increases with the size of the effect in the components, except for large PE and RT effect sizes in opposing directions. Although all the measures pass this test, some of the measures are more efficient than others, but which ones are more efficient depends on a number of factors. I will return to this point later in this General discussion.

The second efficiency question concerns how well the integrated measures detect contrasts in variables orthogonal to the main contrast. In Study 3, this concerned the variations in trade-off strategy. This factor had really big effects in RT and PE, and only IES, RCS, and LISAS were capable of extracting this information. The two binning measures (Bin-p and Bin-i) dramatically failed to do so (Fig. 2).

The second desirable property of integrated measures concerns its added value, namely the extent to which more of the variance is accounted for than by the component measures. Table 14 shows the answers in the column labeled "Added Value?" When the RT and PE effects were in the same direction, the integrated measures, except IES, were very efficient in accounting for an even bigger amount of the variance than the maximum achieved by the components. In all the other cases, the integrated measures were less successful. In cases with opposing RT and PE effects, the integrated measures rarely ever accounted for a larger proportion of the variance than the components. In the cases where only one of the two components had an effect, the integrated measures accounted for a larger part of the variance than the effective component in a rather small part of the samples, except for the binning measures in the case with only a PE effect. This exception can probably be accounted for by assuming that these measures assign a larger weight to the PE component, which is directly related to the next question.

The final issue addressed in Table 14 concerns the question whether the two components, RT and PE, are represented in a balanced way in the integrated measures. In particular in the cases with unbalanced and opposing effects, Study 3 provided some indications that the binning measures are more sensitive to the PE than to the RT effect, probably because of the high penalty applied to errors.

Potential limitations of the present studies

Before drawing any conclusions from this comparative overview of the properties of the diverse integrated measures, it is important to first discuss the potential limitations of the present simulations. Two issues may be of importance: the adequacy of using Monte Carlo simulations, and the adequacy of the present methodology.

The main issue at stake here is whether it is appropriate to use Monte Carlo simulations rather than reanalyzing relevant existing data. Monte Carlo simulation has several advantages over the usage of existing data. First, it is possible to obtain large numbers of samples without too much effort. Second, the samples can all be replications of the same basic design. Third, the simplest possible design can be used so that the data are not obscured by other factors that could result in strategic adaptations in real subjects. For all these reasons, the present paper completely relies on simulations.

A related concern is whether the most adequate choices have been made for the present studies. The main question addressed in the present paper concerns the utility of integrated measures of speed and accuracy. The simplest design in which this can be implemented is by using a contrast between a control and an experimental condition which differ from each in speed and/or accuracy. The model in Equation 4 defines this simple situation. In order to achieve as much realism as possible it was further assumed that subjects differ from each other in their personal speed and accuracy (π_j) in Equation 4. In order to check on strategic effects of speed-accuracy trade-off, Study 3 also included trade-off strategy as a factor.

One design choice which may raise some concern relates to the decision to ensure that the PE and RT effect sizes were of a comparable size. One could indeed object that in practice it will never occur that opposing RT and PE effects will be equally strong, and that exactly in these imbalanced cases an integrated measure may be useful to obtain a clear picture of the basic findings. Nevertheless, if an integrated measure is biased towards one of the components, how would it be possible to detect this bias if the same component has also a stronger effect size in the sample. Hence, usage of a design in which RT and PE effects were as much possible in balance is the best way to find out about biased weighting of the components.

Utility of the integrated measures

Keeping in mind the summary of the findings thus far and the considerations regarding potential limitations of the present series of simulations, it is now time to turn to the main issue that motivated the present study, namely the question concerning the utility of using integrated speed-accuracy measures. It should already be clear by now that there is no simple yes/no answer to the question whether the usage of such measures has advantages. Several factors may play a moderating role. The factors that will be considered concern the implicit bias of integrated measures, the utility of integrating speed and accuracy when they have opposing effects, the utility when only one of the components yields clear effects, and the utility when the effects reinforce each other. In the elaboration of each of these conditions it is not only useful to try to come to a general answer, in most cases it will also be useful to check whether some measures are more suitable for the specific condition under discussion.

Does it matter whether an integrated measure is biased towards one of the components? Assume for the sake of the argument that it does not matter whether an integrated measure is biased towards one of the components, and that there exists a measure M that *validly* integrates RT and PE performance into a single measure. Further assume that this measure is biased towards PE. If the PE effect is rather weak in the sample, the chance that M shows a significant effect would be rather small and the study would be bound to conclude that in this case, there is no integrated effect of speed and accuracy. On the contrary, if the PE effect in the study is rather strong, the study would be likely to conclude the opposite. Next, consider the possibility that M is biased towards RT: if the RT effect is weak, the measure is likely not to detect an integrated effect, whereas if the RT effect in the study is strong, this measure would be likely to detect an integrated speed-accuracy effect. In other words, if one does not know with some degree of certainty what the implicit weights of the measure are, it is difficult to trust any of the effects detected by the measure when the opposing effects are not in balance.

What could the researcher do then? In fact, there are only two options. The first option is that the researcher of such a study decides on the weights given to the two components RT and PE. As long as this is not an arbitrary choice but a well motivated choice based on theoretical considerations, this could be acceptable to the scientific community. The second option, and in my view the best one, is to give equal weight to the two components, RT and PE.

Is such a control over the weights of the components possible? In the three direct measures it is. For example in IES, it is possible to multiply the numerator of Equation 1 by a factor k (k > 0) so as to give a smaller (k < 1) or a larger weight (k > 1) to the RT component. Similar operations are possible for RCS and LISAS. Can this also be done for the binning measures? As the Bin-p and Bin-i scores are obtained as the sum of the proportion of RT differences in each of the ten bins on the one hand and the proportion of errors weighted by 20 on the other hand, a change in the weight assigned to the proportion of errors should do the trick. However, in order to do this properly, it is necessary to find the "neutral" point, i.e., the value at which the RT and the PE component are given equal weight. If such a neutral point exists, it is possible to find it in the context of a simulation study covering a large set of samples. Unfortunately, this value would probably depend on the combination(s) of RT and PE effect size used in the simulation. The result would therefore not be generalizable to other situations. Hence, each researcher who wishes to use a Bin measure and prefers to control the weight balance, would first have to run a simulation to find out about the neutral point. It can be done, but it is not very practical. Consequently, it seems that only the integrated measures based on a mathematically transparent combination of the two components leave room for the researcher to change the relative weights assigned to the two components.

Thus far, it can be concluded that it is important to know about the bias implicitly present in the integrated measure if it is to be useful. Yet, the question remains whether integrated measures can be useful when the effects of speed and accuracy are contradictory. It can be argued that in such a situation, the most important information is available in the RT and PE effects themselves, and that an integration of these two sources of information can only be meaningful if there is some theoretical or empirical basis to calculate a weighted average of the two components or if that is not possible to make an integration that gives equal weight to their effects (not necessarily to the components themselves). This opposing effect situation is probably the most important one that can be encountered. Clearly, a neat conclusion will only be possible if, on the one hand, there is some difference in effect sizes in the sample and this difference can be trusted, and on the other hand, a fair integrated measure is applied.

In situations where only one of the two components has a clear and reliable effect, an integration of the two components may be useful, especially if it helps to account for a larger proportion of the variance. In the simulations of Study 3, the design ensured that the RT or PE effects were effectively zero in the population from which the artificial data were sampled. In practice, this translates to situations where there is a significant RT effect joined with a nonsignificant PE effect, or vice versa, a significant PE effect joined with a nonsignificant RT effect. These are two interesting, but quite different cases, which are therefore discussed separately.

First consider the situation with nonsignificant PE effects. Is the usage of an integrated speed-accuracy measure useful in such a situation? In research with paradigms were the focus is usually on RT measures (because they have better statistical properties than PE measures), most researchers will not care much about such outcomes. In particular, researchers will not bother about nonsignificant PE effects if the means are in the expected direction but are not significantly different. The fact that the researchers have no knowledge of a suitable measure to integrate both effects so as to achieve a more solid statistical conclusion may be at the basis of the choices being made. Indeed, until recently only one integrated measure was more widely known, namely IES. As the present article shows, today, there are at least five measures from which a choice can be made. Could or should this make a difference? The present paper shows that some of these measures are useful in this particular context. Study 3 showed that in such situations, detection efficiency was high for LISAS (83 % overall) and RCS (82 %), and that these measures also accounted for more variance than the components in 16 % and 13 %, respectively, of the samples. These two measures are obviously useful in situations with weak PE effects, bearing in mind that the RCS distribution may be skewed.

The situation with nonsignificant RT effects is quite different. Many researchers will prefer not to trust the absence of any robust RT effects and will be reluctant to conclude anything from significant PE effects. Again usage of an integrated measure may be considered here. However, in view of the possibility that some measures give more weight to PE than to RT information, it is possible that an integrated measure will not be able to extend the information beyond the significant PE effect. This suspicion is strengthened by the finding (see Panel C of Fig. 3) that the more biased measures (Bin-p and Bin-i) achieve the best detection rates. However, also in these situations, LISAS and RCS, and to a lesser extent IES, hold some promise. All three measures attained a sufficiently high detection efficiency (83 %, 78 %, and 80 %, respectively) while accounting for more of the variance in some of the samples (13 % for LISAS and RCS; 8 % for IES). Hence, to the extent that the RT effects are not really zero as in the present simulations, but have some weak significant effects, it might be worthwhile to consider the usage of one of these three integrated measures which do not show strong biases towards one of the components.

The results of Study 3 show that the integrated measures in general perform best in situations where the RT and PE effects are clearly present and point in the same direction. This is also the kind of situation in which researchers, in general, do not feel the need to use integrated measures. After all, this kind of situation raises few concerns because both RT and PE have effects that support the same conclusion. Yet, this is exactly the situation in which the integrated measures attained the highest levels of detection efficiency and accounted for more of the variance in a high proportion of all samples. It is also the kind of situation in which the potential biases towards one or the other component are least obvious and are least likely to lead to incorrect conclusions. In fact, all five integrated measures detected the effects in almost 100 % of the samples, and they accounted for more of the variance than the component measures in about 50 % of the samples, except for IES which was successful in only 15 % of the samples, while LISAS did so in more than 60 % of the samples. Once more, it seems that some of the integrated measures may be helpful to reach unambiguous conclusions more often than is possible with separate RT and PE measures. It should be stressed, though, that in this kind of situation too, some of the measures are more useful.

In consideration of all the issues discussed thus far and of the summary of the findings schematized in Table 14, it seems fair to conclude that there are several reasons to avoid the usage of the binning measures. First, they require a quite elaborate calculation procedure. Second, the binning measures result in an imbalanced combination of PE and RT variation, which can only be rectified by extensive additional calculations. Third, when the design in which they are used also contains other factors in addition to the main contrast, it is quite likely that they will fail to detect even large effects associated with these other factors. Fourth, as mentioned at the occasion of Study 2, the Bin-p measure requires the data from the complete sample which raises some doubts about the independence of the scores per subject. However, this can be avoided by using Bin-i instead, but the stability of this measure will require large numbers of trials per condition.

Thus only IES, RCS, and LISAS are left for further consideration. These three measures are all valid, and allow the researcher to vary the weights assigned to the speed and accuracy components in calculating the measure. Nevertheless, these three measures differ from each other in the efficiency with which they can recover effects present in the two components and the extent to which they can account for more of the systematic variance than the component RT and PE measures can. Table 14 indicates that IES would best be avoided because it rather infrequently succeeds in accounting for a larger proportion of the variance than the RT and PE measures account for. The findings of Study 3 further show that LISAS better recovers the experimental-control contrast as well as the effect of a factor orthogonal to this contrast than RCS does. In turn, RCS has a higher detection efficiency than IES. In sum, if one wants to maximize the chances of finding a reliable integrated speed-accuracy effect, there are better choices than IES, even though IES is a valid measure and the results obtained by using IES are expected to be trustworthy. To the extent that skewness of the distribution is considered to be a contraindication, the RCS measure should be avoided.

The present simulations do not only support suggestions about which integrated measures are more efficient than the others, by varying the direction of the RT and PE effects, information is also obtained about the situations in which integrated measures may or may not be useful. The introduction already explained that not all experiments that involve speed and accuracy can take advantage from integrated speed and accuracy performance scoring. When there is no reason to assume that speed and accuracy are driven by the same or by overlapping processes, it would seem counterproductive to use integrated speed-accuracy scores. For example, when testing the hypothesis that in dual-task situations the speed of performance is slowed due to the coordination of the two tasks, while errors are produced when the capacity limit for processing information is exceeded, integrated measures are better not considered. In contrast, when it is warranted to assume that speed and accuracy have a common basis, integrated scoring may be considered. However, the results of Study 3 show that the decision to use such measurements requires an inspection of the speed and accuracy data. More than an advice (as formulated by Bruyer & Brysbaert, 2011), it is in fact a necessity. When RT and PE effects are observed to be in the same direction, much can be gained by using one of the better integrated measures. Even when only one of the two effects attains significance and the differences point in the same direction, usage of integrated measures may still be advantageous. However, when the PE and RT effects are in opposing directions and they are of equal strength, not much is to be gained from the usage of integrated measures. Therefore, it is a necessity to always test the component effects, RT and PE, before calling on integrated scoring.

Conclusion

When speed and accuracy rely on common processes and when the effects of speed and error rate are showing differences in the same direction, the measures that directly integrate RT and PE can be useful. However, it remains necessary to always first test the direction of the RT and PE effects. Under these conditions, two measures, namely LISAS and RCS, are likely to be advantageous by yielding an integrated effect size that recovers the information present in both component measures and that accounts for a larger proportion of the variance than these component measures do. Usage of IES in these circumstances will not lead to invalid results, but the likelihood of being advantageous is much smaller. For many reasons explained in this article, the binning measures are better avoided.

References

Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General, 133*(4), 584–600. doi:10.1037/0096-3445.133.4.584

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*(1), 5–13.

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and ccuracy: The relationship between working memory capacity and task switching as a case example. *Perspectives on Psychological Science, 11*(1), 133–155. doi:10.1177/1745691615596990

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response-time distribution: An example using the Stroop task. *Psychological Bulletin, 109*(2), 340–347. doi:10.1037/0033-2909.109.2.340

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*(3), 702–721. doi:10.3758/s13428-013-0411-5

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin, 136*(5), 849–874. doi:10.1037/a0019842

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus–response compatibility: A model and taxonomy. *Psychological Review, 97*(2), 253–270. doi:10.1037/0033-295x.97.2.253

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163–203. doi:10.1037/0033-2909.109.2.163

Ratcliff, R. (1979). Group reaction-time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86*(3), 446–461. doi:10.1037/0033-2909.86.3.446

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83*(3), 190–214. doi:10.1037//0033-295x.83.3.190

Stroop, J. R. (1935). Studies of inteference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. N. J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 199–239). New York: Lawrence Erlbaum Associates.

Trabasso, T., & Bower, G. H. (1966). Presolution dimensional shifts in concept identification: A test of the sampling with replacement axiom in all-or-none models. *Journal of Mathematical Psychology, 3,* 163–173. doi:10.1016/0022-2496(66)90009-5

Vandierendonck, A., Liefooghe, B., & Verbruggen, G. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin, 136*(4), 601–626. doi:10.1037/a0019791

White, R. M. (1972). Relationship of performance in concept identification problems to type of pretraining problem and response-contingent feedback intervals. *Journal of Experimental Psychology, 94,* 132–140. doi:10.1037/h0032804

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition, 34*(3), 668–684. doi:10.3758/bf03193587