Psychologica Belgica
2011, 51-1, 5-13

COMBINING SPEED AND ACCURACY IN COGNITIVE PSYCHOLOGY: IS THE INVERSE EFFICIENCY SCORE

OF ERRORS (PE)?

Raymond BRUYER(1) & Marc BRYSBAERT(2)[1]
(1) University of Louvain-la-Neuve & (2) Ghent University

Experiments in cognitive psychology usually return two dependent variables: the percentage of errors and the reaction time of the correct responses. Townsend and Ashby (1978, 1983) proposed the inverse efficiency score (*IES*) as a way to combine both measures and, hence, to provide a better summary of the findings. In this article we examine the usefulness of *IES* by applying it to existing datasets. Although *IES* does give a better summary of the findings in some cases, mostly the variance of the measure is increased to such an extent that it becomes less interesting. Against our initial hopes, we have to conclude that it is not a good idea to limit the statistical analyses to *IES* without further checking the data.

Speed and accuracy as dependent variables

Most studies in experimental cognitive psychology involve participants performing some task. These studies typically return two dependent variables (*DV*): the proportion of errors (*PE*) and the latency of the correct responses (i.e., the time elapsed between the onset of the stimulus and the onset of the response), expressed as the *Reaction Time* or *RT*. Most of the time the variables are analysed separately, which tends to complicate the interpretation.

First, authors check whether the conclusions based on *PE* and *RT* go in the same direction, or whether there is evidence for a speed-accuracy trade-off. In the latter case, the conditions with faster responses have higher error rates. In such a situation it usually is impossible to reach a convincing conclusion. When *PE* and *RT* point in the same direction, authors tend to focus on the *RT* analysis, unless the percentage of errors is high (e.g., more than 15%) or the *PE* analysis returns a significant effect in the predicted direction whereas the *RT* analysis does not.

---

1. Raymond Bruyer, Institute of Research in Psychological Science; Systems and Cognition Neuroscience, University of Louvain-la-Neuve, Belgium; Marc Brysbaert, Department of Experimental Psychology, Ghent University, Belgium.
The authors wish to thank Salvatore Campanella, Frédéric Joassin and Mandy Rossignol.
Correspondence concerning this article should be addressed to Raymond Bruyer, Institut de recherche en sciences psychologiques, Place du cardinal Mercier, 10, 1348 Louvain-la-Neuve. E-mail: Raymond.bruyer@uclouvain.be

The situation would be simplified if *PE* and *RT* could be integrated into a single *DV*, which appropriately weighs the impact of speed and accuracy. Such a measure has been proposed by Townsend and Ashby (1978, 1983).

The inverse efficiency score

To deal with the issue of how to combine speed and error, Townsend and Ashby (1978) proposed the "inverse efficiency score" (IES; see also Townsend & Ashby, 1983). *IES* can be thought of as an observable measure that gauges the average energy consumed by the system over time (or the power of the system; Townsend & Ashby, 1983, p. 204). It consists of *RT* divided by $1 - PE$ (or by *PC*, the proportion of correct responses). So, for a given participant the mean (or median) *RT* of the correct responses in a particular condition is calculated and divided by $(1 - PE)$ or by *PC*:

$$IES = \frac{RT}{1 - PE} = \frac{RT}{PC}$$

Since *RT*s are expressed in ms and divided by proportions, *IES* is expressed in ms as well. For instance, if a participant in a particular condition responds with an average *RT* of 652 ms and makes 5% errors, then *IES* = 652/(1-.05) = 652/.95 = 686 ms.

At first sight, *IES* seems to have all the properties which we would want the combined measure to show. When two conditions have the same average *RT* but differ in *PE*, then the *IES* of the condition with the higher *PE* will increase more than the *IES* of the condition with the lower *PE*. So, a condition A with *RT* = 650 ms and *PE* = .05 will have *IES* = 684 ms, whereas a condition B with *RT* = 650 ms and *PE* = .07 will have *IES* = 699 ms. Similarly, when there is a trade-off between speed and accuracy, the *IES* effect will compensate for the differences in *PE*. Take, for instance, condition A with *RT* = 650 ms and *PE* = .05, and condition B with *RT* = 640 ms and *PE* = .06. Then condition A has *IES* = 684 ms, against *IES* = 681 ms for condition B (see, however, below for a serious limitation of this use of *IES*).

For the above reasons, a number of researchers in various fields of experimental cognitive psychology have started to use the *IES* measure (e.g., Akhtar & Enns, 1989; Christie & Klein, 1995; Goffaux, Hault, Michel, Vuong, & Rossion, 2005; Jacques & Rossion, 2007; Kennett, Eimer, Spence, & Driver, 2001; Kuefner, Viola, Vescovo, & Picozzi, 2010; Minnebusch, Suchan, & Daum, 2009; Murphy & Klein, 1998).

An example where *IES* works well

A study by Rossignol, Bruyer, Philippot, and Campanella (2009) nicely illustrates the potential of *IES versus* separate analyses of *PE* and *RT*. Participants had to identify the emotional expression displayed by morphed faces, which were blends of an emotional expression and a neutral expression. Six different emotional expressions, each blended with the neutral expression, were used and participants had to say which emotion they thought was shown. The relative contribution of the emotional expression varied from 10% to 90% in steps of 10%. As expected, participants were faster when the emotional expression was stronger: *RT* decreased with increasing impact of emotional expression (Figure 1). However, the longest *RT*s were not observed for the condition with 10% emotional expression, but for the condition with 30% emotional expression. This is the type of pattern typically observed when two different emotions are morphed (e.g., anger and sadness). When one emotion strongly dominates (e.g., 90% anger and 10% sadness, or 10% anger and 90% sadness), *RT*s are fast. Somewhere in the middle of the continuum, there is a category boundary (from anger to sadness) around which *RT*s are long. Did the results of Rossignol et al. (2009) imply that neutral faces were a category as well? Evidence against this interpretation was found in *PE*. Accuracy dropped for morphs with low contributions of the emotional expression (particularly 10% and 20%), because participants could no longer identify the emotion that was displayed.

When the latencies of the correct responses were weighted by the proportion of correct responses, that is to say when *IES* was calculated, another picture emerged. The *IES*-curve kept rising up to the 10% condition (Figure 1), in line with a Signal Detection view saying that the neutral expression consisted of noise against which the emotional signal was perceived. Given that the latter is a better description of what was going on in the study of Rossignol et al. (2009), the *IES* curve provides us with a more accurate picture of the data than the *RT* curve.

IES is not always better than RT

Unfortunately, further analysis indicated that the situation is not always as clear-cut as in the example above. Take, for example, a study published by Goffaux et al. (2005). In this study, participants were shown triads of faces: a target face at the top of a triangle, and two comparison faces at the bottom angles. The comparison faces consisted of the target and a distractor. All faces were filtered so that only low spatial frequencies (*LFS*), only high spatial frequencies (HSF), or the full spectrum of spatial frequencies (Full) were displayed. In addition, the distractor face could differ locally (i.e., had differ-

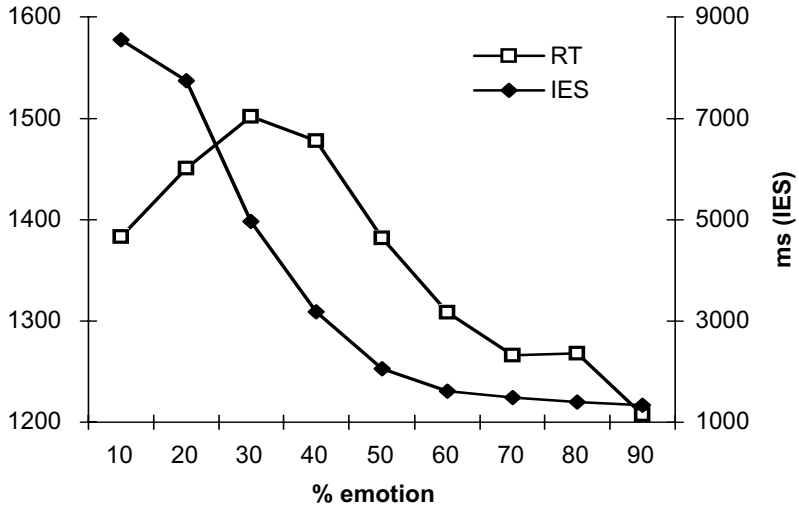

Figure 1

The effect of the morphing manipulation of facial expression on correct latencies and IES. The abscissa indicates the % of emotional expression in the morphs, and the ordinates indicate correct RT (left ordinate) and IES (right ordinate), both in ms. Data from Rossignol et al., 2009

ent features), globally (had a different overall shape), or both locally and globally. Participants had to indicate which comparison face matched the target face. A 3x3 (filter * type of distractor) ANOVA was computed. The main effect of distractor was significant for both *RT* and *IES*. However, for *RT*, this was due to an advantage of the conditions local and local + global, which did not differ from each other, over the condition global, whereas for *IES* all pairwise comparisons were significant. Thus, a new effect emerged after the *IES* transformation. The interaction between spatial filter and type of distractor was significant both for *RT* and *IES*, but had a different shape. When global distractors were used, the analysis of *RT*s revealed an advantage of *HFS* and Full (which did not differ from each other) over LSF. However, with the *IES* measure there was no significant difference between the three conditions. Thus, an effect observed with *RT* disappeared after the *IES* transformation.

We ran similar analyses of other studies we had access to. Four patterns emerged: (a) no change after transformation (Campanella, Bruyer, Froidbise, Rossignol, Joassin, Kornreich et al., 2010; Christie & Klein, 1995; Jacques & Rossion, 2007; Kennett et al., 2001); (b) the disappearance of significant effects with *IES* (Kuefner et al., 2010); (c) the apparition of significant effects with *IES* (Bruyer, Mejias, & Doublet, 2007); and (d) the disappearance and

apparition of significant effects in the same study (Bruyer, Leclère, & Quinet, 2004; Constant, Lancereau, Gillain, Delatte, Ferauge, & Bruyer, in press; Goffaux et al., 2005; Joassin, Maurage, Campanella, & Bruyer, 2006; Rossignol et al., 2009). So, it is not the case that *IES* always clarifies matters. It looks pretty much like every type of change is possible with the introduction of *IES*.

The reason why *IES* has no straightforward relationship with *RT* is that it combines two variables subject to sampling error. This increases the variability of the measure. In addition, it is not clear whether the division of *RT* by *PC* is always a good reflection of the relative weights of speed and accuracy. To examine these weaknesses, we made use of an effect that is well documented and for which there are many data available. One of the best established effects in cognitive psychology is the word frequency effect in lexical decision: participants decide much better that a letter string is a word rather than a nonword when the letters make a high-frequency word than when they make a low-frequency word. In addition, there are now lexical decision times for thousands of words, making it possible to get a clear image of the word frequency effect across the entire range. For instance, in the so-called French Lexicon Project, Ferrand, New, Brysbaert, Keuleers, Bonin, Méot et al. (2010) assessed lexical decisions for 38,335 French words, which were seen by 25 participants each (the entire study involved 975 participants).

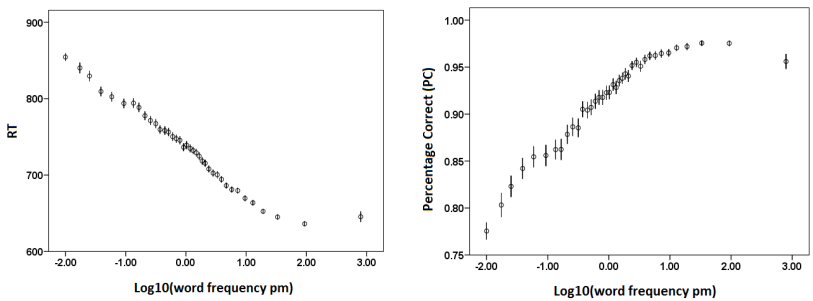

Figure 2

The word frequency effect in the French Lexicon Project (Ferrand et al., 2010). Left panel: Mean RT of the correct responses; right panel: PC. Frequency = log10 frequency per million. Each dot is the average of 1000 words. Bars indicate the 95% confidence interval

Figure 2 shows the *RT*s and *PC*s as a function of log10 word frequency per million (pm; so 0 = a frequency of 1 pm, 1 = a frequency of 10 pm, 2 = a frequency of 100 pm, and -1 = a frequency of .1 pm). Both *RT* and *PC* show a clear relationship with word frequency: *RT*s are about 200 ms slower for words with a frequency below .1 pm than for words with a frequency above

100 pm. At the same time, many more errors are made for the low-frequency words (*PC* = .77) than for the high-frequency words (*PC* = .96).

Figure 3 shows the effect for *IES*. Although the effect looks stronger, the confidence intervals indicate that there is much more noise in the data. Indeed, whereas word *RT*s never exceeded 1,500 ms, individual word *IES*-values went up to more than 15,000 (large *RT* and *PC* < .10). As a result, the percentage of variance explained by frequency is much less for *IES* (R^2 = .12 for a 3 degree polynomial) than for *RT* (R^2 = .33 for a 3 degree polynomial). Only when the analysis was limited to those words with a *PC* of .90 and more did we find equivalent percentages of variance explained for *RT* and *IES* (R^2 = .28).

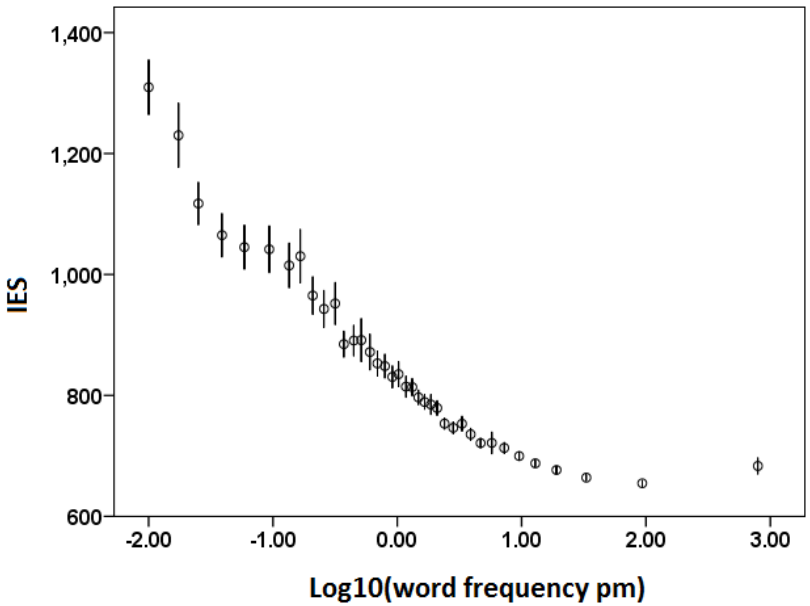

Figure 3
The word frequency effect in the French Lexicon Project for IES. Bars indicate the 95% confidence interval

Situations in which *IES* may be problematic

The preceding analyses suggest that *IES* is better not used when *PC* < .90. When high errors rates are observed, three problems arise (Akhtar & Enns, 1989). The first problem is that high numbers of errors imply that the number of correct responses is low, so that estimates of correct latencies become

rather unstable. The second problem is that some of these "correct" responses are probably the result of guesses or even "mistakes" (e.g., the participant wanted to indicate that the low-frequency word was not a word, but pressed the wrong button). Finally, the multiplication of *RT* due to *PC* is not linear, but accelerates the lower *PC* becomes (e.g., *RT* is multiplied by 1 when *PC* = 1.0, by 1.1 when *PC* = .90, but by 1.4 when *PC* = .70 and by 1.7 when *PC* = .60). In all likelihood, the multiplication weight of accuracy becomes too high for low *PC*-values. This is particularly a problem in situations where the lower limit of correct guessing is low, for instance in the experiment of Rossignol et al. (2009) where six different expressions were used. Although in a lexical word/non-word decision experiment, it can be defended that no *PC*s under .60 are meaningful (otherwise the participant is simply guessing or does not know the word), in an experiment with 6 response alternatives meaningful *PC*s can go as low as 1/6 = .17, meaning that for these accuracy levels *RT*s can be multiplied by up to 6.

Furthermore, Townsend and Ashby (1978, 1983) warned that the *IES* only works when there is a positive correlation between *RT* and *PE*: the *IES* transformation should be used only if *high* and *linear* correlations are evidenced. So, Townsend and Ashby advise ***not*** to use *IES* in cases of speed-accuracy trade-off. A preliminary check to make sure that *RT* and *PE* are positively correlated was done by Akhtar and Enns (1989), but seems to be absent in many other studies. Indeed, Christie and Klein (1995), Goffaux et al. (2005), Jacques and Rossion (2007), Kennett et al. (2001), Kuefner et al. (2010), and Murphy and Klein (1998) simply reported the usual separate analyses of accuracy, correct latencies, and *IES* without any further information. Minnebusch et al. (2009) did not analyse accuracy and correct latencies but only *IES*. Kennett et al. (2001) and Murphy and Klein (1998) mentioned the problem of the speed-accuracy trade-off, saying that some participants might show the effects of interest for speed and others for accuracy, but did not address the issue directly in their statistical analyses.

Furthermore, as our example of the French Lexicon Project shows, a positive correlation between *RT* and *PE* does not guarantee that more variance will be explained in the *IES* measures than in the *RT* measure (the correlation between *RT* and *PE* in Ferrand et al. (2010) is $r[38333] = 0.56$, $p < 0.001$).

Conclusions

Although we set out with high hopes that the *IES* measure could be a better and a more concise variable than *RT* and *PE* to convey the findings of a cognitive psychology experiment, our analyses have shown that the "blind" use of *IES* as *DV* is likely to lead to interpretation problems. The main issue is that *IES* increases the variability of the data. This became very clear in the statis-

tical analysis of the data from the French Lexicon Project: although the frequency effect on average seemed clearer for *IES* than for *RT* (Figures 2 and 3), the percentage of variance explained in *IES* by word frequency was less than half of the variance explained in *RT*. This remained the case even when the words were limited to those with *PC* > .66. Only for the words with *PC* > .90 was the amount of variance explained in *IES* and *RT* the same. Further analyses will have to indicate whether some other combination of *RT* and *PC* [e.g., log (IES)] is a better measure.

The increase in variability will be particularly problematic in studies with small numbers of observations per condition (as is mostly the case in cognitive psychology). If a condition only has 20 stimuli, then 2 errors already make a difference of .10 in *PC*. Depending on where in the range this 10% falls, *RT* multiplications will range from 1.1 (from 1.00 to .90) to 2.0 (from .20 to .10). For *RT*s around 600 ms, this means increases from 60 to 600 ms, which are considerably larger than the *RT* effects generally investigated in cognitive psychology (which tend to be in the order of 20-60 ms). In general, an increase in the variance of the *DV* will diminish the power of the experiment. Occasionally, however, it may result in a spurious effect, when by chance a few more mistakes are made in one condition than in the other. This will lead to Type I errors (the illusion of having found a significant difference, whereas in reality there is none), which is particularly a problem if journals are more likely to publish "statistically significant" effects than null-effects. In our experience, spurious effects are most likely in the interaction terms of multivariable experiments.

All in all, *IES* only seems to have value when the number of errors is small and when there is a high correlation between *RT* and *PE*, indicating that both variables are in unison. Even then, it is presumably safer to calculate *RT* and *PE* as well, to make sure that *IES* is in line with them. On the basis of our analyses, we have to conclude that it is not a good idea to limit the analyses to *IES* without any further checking of the data.

References

Akhtar, N., & Enns, J.T. (1989). Relations between covert orienting and filtering in the development of visual attention. *Journal of Experimental Child Psychology, 48*, 315-334.

Bruyer, R., Leclère, S., & Quinet, P. (2004). Ethnic categorisation of faces is not independent of face identity. *Perception, 33*, 169-179.

Bruyer, R., Mejias, S., & Doublet, S. (2007). Effect of face familiarity on age decision. *Acta Psychologica, 124*, 159-176.

Campanella, S., Bruyer, R., Froidbise, S., Rossignol, M., Joassin, F., Kornreich, C., Noël, X., & Verbanck, P. (2010). Is two better than one? A cross-modal oddball

paradigm reveals greater sensitivity of the P300 to emotional face-voice associations. *Clinical Neurophysiology,121*, 1855-1862.

Christie, J., & Klein, R. (1995). Familiarity and attention: does what we know affect what we notice? *Memory and Cognition, 23*, 547-550.

Constant, E.L., Lancereau, J., Gillain, B., Delatte B., Ferauge, M., & Bruyer, R. (in press). The deficit in negative emotional information processing in schizophrenia: does it occur in all patients? *Psychiatry Research.*

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488-496.

Goffaux, V., Hault, B., Michel, C., Vuong, Q.C., & Rossion, B. (2005). The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception, 34*, 77-86.

Jacques, C., & Rossion, B. (2007). Early electrophysiological responses to multiple face orientations correlate with individual discrimination performance in humans. *NeuroImage, 36*, 863-876.

Joassin, F., Maurage, P., Campanella, S., & Bruyer, R. (2006). Is associative priming a valid method to differentiate the serial and parallel models of face identification? *Visual Cognition, 14*, 199-216.

Kennett, S., Eimer, M., Spence, C., & Driver, J. (2001). Tactile-visual links in exogenous spatial attention under different postures: convergent evidence from psychophysics and ERPs. *Journal of Cognitive Neuroscience, 13*, 462-478.

Kuefner, D., Viola, M.C., Vescovo, E., & Picozzi, M. (2010). Natural experience acquired in adulthood enhances holistic processing of other-age faces. *Visual Cognition, 18*, 11-25.

Minnebusch, D.A., Suchan, B., & Daum, I. (2009). Losing your head: behavioral and electrophysiological effects of body inversion. *Journal of Cognitive Neuroscience, 21*, 865-874.

Murphy, F.C., & Klein, R.M. (1998). The effects of nicotine on spatial and non-spatial expectancies in a covert orienting task. *Neuropsychologia, 36*, 1103-1114.

Rossignol, M., Bruyer, R., Philippot, P., & Campanella, S. (2009). Categorical perception of emotional faces is not affected by aging. *Neuropsychological Trends, 6*, 29-49.

Townsend, J.T., & Ashby, F.G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory. Vol. 3.* (pp. 200-239). Hillsdale, N.J.: Erlbaum.

Townsend, J.T., & Ashby, F.G. (1983). *Stochastic modeling of elementary psychological processe*s. Cambridge: Cambridge University Press.

Received June 23, 2010
Revision received September 30, 2010
Accepted October 1, 2010