General or Idiosyncratic Item Effects: What Is the Good Target for Models?

Pierre Courrieu, Arnaud Rey

General or idiosyncratic item effects: what is the good target for models?

Pierre COURRIEU and Arnaud REY

CNRS, Aix Marseille Université, LPC UMR 7290, 13331, Marseille, France

Final publication of this manuscript:

Courrieu, P., & Rey, A. (2015). General or idiosyncratic item effects: what is the good target for models? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5),* 1597-1601. DOI: 10.1037/xlm0000062

Running head: Idiosyncratic item effects

Corresponding author:

Pierre Courrieu,

Laboratoire de Psychologie Cognitive (LPC),

UMR 7290, CNRS-Université d'Aix-Marseille,

Centre Saint-Charles, 3 place Victor Hugo,

13331 Marseille Cedex 3, France

E-mail: pierre.courrieu@univ-amu.fr

Abstract. Recently, Adelman et al. (2013) formulated severe criticisms about approaches based on averaging item RTs over participants, and associated methods for estimating the amount of item variance that models should try to account for. Their main argument is that item effects include stable idiosyncratic effects. In this commentary, we provide supplementary empirical evidences that this assertion is indeed valid. However, the actual implications of this result are not those defended in Adelman et al. (2013), where there seems to be a confusion between the precision of measures and the nature of target effects. Indeed, basic statistical considerations show that any arbitrary data precision level can be achieved in all cases using an appropriate number of observations per item, while general and idiosyncratic item effects are both targets of interest for modelling, but in different questionings.

Key words. Idiosyncratic item effects; systematic item variance; reading model testing; large-scale behavioural databases; intraclass correlation coefficient

Introduction

Recently, Adelman, Marquis, Sabatos-DeVito, and Estes (2013) proposed to model individual participants performance in the word-naming task. Their paper includes severe criticisms about approaches based on averaging item RTs over participants, while using an intraclass correlation coefficient (ICC) for estimating the amount of item-related variance that models should try to account for (Courrieu, Brand-D'Abrescia, Peereman, Spieler, and Rey, 2011; Courrieu and Rey, 2011; Rey, Courrieu, Schmidt-Weigand, and Jacobs, 2009). The Adelman et al. (2013) criticisms are based on the idea that there is no general item effect in the RTs, but that item effects depend on each possible participant in a way that is not random, and that does not reduce to a simple linear transformation of a general item effect. As a consequence, "analysis techniques that treat individual differences as noise will necessarily overestimate the amount of noise contributing to the mean RT for each word. This overestimation of noise results in an underestimation of the variability that a model should explain, leading to an overestimation of the success of models" (Adelman et al., 2013, p. 1038).

The authors present empirical evidences supporting their idea, based in particular on Kristof's (1973) method. However, the idea that there is no general item effect must not be considered in a radical sense since, averaging RTs over participants in word naming tasks, one commonly observes that the percentage of available systematic item variance is greater than 80% (Courrieu et al., 2011; Rey and Courrieu, 2010; Rey, Courrieu, Madec, and Grainger, 2013). This would not be possible if there was no general item effect common to all participants. Thus, the correct modelling of an

individual participant item effect probably involves a general item effect plus an independent idiosyncratic item effect. The question is now: what is the relative contribution of these two item effects, on the average? If the general item effect contributes for 80%, it is clear that the idiosyncratic item effect contributes for less than 20% of the observable item variance, since in addition, there is always a non-zero contribution of the random noise.

Adelman et al. (2013) consider, as an example, the case where the idiosyncratic effect accounts for 10%, and the noise accounts for the remaining 10% of the observable item variance. They argue that in this case "an analysis that treats individual differences as noise would only set a target of 80% variance explained for a model, when, in fact, 90% could in fact be explained" (p. 1038). However, by accounting for idiosyncrasies, we would generate an idiosyncratic model accounting for the behaviour of an undetermined number of participants having similar idiosyncrasies, but the generalization power of this model for other randomly chosen participants would probably be quite poor. So, the first question is in fact: what do we plan to do with the model? The answer obviously depends on the context and the goal of the investigation, and there is no universal better choice. If we plan to model general mechanisms governing the reading process, then we must randomly sample a number of participants in the general population and take their average item effect as the target, which ensures the best generalization power of the model for this population. Now, if one plans to model individual performance in the perspective of clinical or educational applications then an idiosyncratic modelling is certainly preferable. However, even in this case, one will probably need references concerning the general population, and references

concerning particular subpopulations (e.g. dyslexics), which requires a general targeting in the considered populations.

Moreover, as Adelman et al. (2013) noted, the state of the art in modelling the reading performance is far from satisfactory. So, one can understand that many researchers prefer trying to identify general mechanisms governing the reading process, while not complicating the picture with a profusion of idiosyncrasies in a first time. It is in this perspective that a number of very large-scale behavioural databases have recently been developed, freely providing to researchers behavioural item level data for thousands words in various languages (ELP: Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007; FLP: Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, Augustinova, & Pallier, 2010; DLP: Keuleers, Diependaele, & Brysbaert, 2010; BLP: Keuleers, Lacey, Rastle, & Brysbaert, 2012). In these databases, the general modelling perspective is clearly favoured, and the item effects are averaged over randomly sampled participants. Since disqualifying these databases, and associated methodologies, could have detrimental consequences on a number of research activities and on the selection of publications, we think that it is necessary to carefully examine the critical arguments of Adelman et al. (2013), and their actual implications.

The remaining of the paper is organized as follows. First, we describe the statistical problem in a simple formal way, in order to clarify the critical argument of Adelman et al. (2013). After this, we test the validity of this argument on two independent data sets, and we estimate the contribution of idiosyncratic item effects. Finally, we discuss the actual implications of the results, showing that Adelman et al. (2013) draw an abusive conclusion from a valid argument.

Formalising the critical hypothesis of Adelman et al. (2013)

In order to be sure of what we are speaking about, let us rapidly describe the problem in a simple formal way.

Let x be an experimental measure, such as a word naming time, for instance. The usual approach assumes that x can be decomposed as follows, for the item i and the participant j :

$$x_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij} , \qquad (1)$$

where μ is the grand mean of x in the data population, α is the participant effect, β is the item effect, and ε is a random effect. The three effects are assumed to be independent, with zero means, and variances var(α), var(β), and var(ε), respectively. In these conditions, one can easily show (see Courrieu et al., 2011) that if one averages x over n independent participants, then the proportion of systematic item variance in the resulting variable has the expected value:

$$\rho = \mathrm{var}(\beta)/(\mathrm{var}(\beta) + \mathrm{var}(\varepsilon)/n). \qquad (2)$$

Setting $q=\mathrm{var}(\beta)/\mathrm{var}(\varepsilon)$, the expression (2) can also conveniently be written as:

$$\rho = nq/(nq+1). \qquad (2')$$

One can note that (2) is the expected value of a well-known intraclass correlation coefficient (ICC), namely the "ICC(2,k)" in the nomenclature of McGraw and Wong (1996). The simplest way of estimating this coefficient, from an x-data table of m items by n participants, consists in replacing the variances in (2) by their estimates (mean squares) as provided by a standard ANOVA (Courrieu et al., 2011). There are other ways of estimating the ICC, such as a permutation resampling Monte Carlo method (Rey et al., 2009), and the various methods for estimating Cronbach's alpha coefficient (Cronbach, 1951), which is theoretically equal to (2).

Assume that a theoretical model provided predictions having a squared correlation r^2, or a squared determination coefficient R^2, with the observed item means. Then the ratio r^2/ICC, or R^2/ICC, is the proportion of systematic item variance accounted for. Moreover, one can show that if the model predictions are not correlated with the data noise, then the ratio r^2/ICC, or R^2/ICC, is the true squared correlation of the model predictions with the underlying item effect β, and it does not depend on the data precision (Courrieu and Rey, 2011).

Now, Adelman et al. (2013) pointed out that the assumption that there is a general item effect β is wrong. Their idea is that the item effect depends on the considered participant in a way that is not random, and that does not reduce to a simple linear transformation of β. Accordingly, we must modify (1) as:

$$x_{ij} = \mu + \alpha_j + \beta_{ij} + \lambda_{ij} , \qquad (3)$$

where the item effect depends now on the participant j, and λ is a noise variable of zero mean and variance var(λ), that accounts for the random fluctuations of the performance as they can be observed in repeated measures.

We can think of β_i in (1) as the expected value of β_{ij}'s of all possible participants. Setting $\delta_{ij} = \beta_{ij} - \beta_i$, one can write (3) as:

$$x_{ij} = \mu + \alpha_j + \beta_i + \delta_{ij} + \lambda_{ij} . \qquad (3')$$

Identifying the terms of (3') with those of (1), we can conclude that:

$$\varepsilon_{ij} = \delta_{ij} + \lambda_{ij} . \qquad (4)$$

An important consequence of this is that $var(\varepsilon) = var(\delta) + var(\lambda) > var(\lambda)$. Moreover, denoting the item effect variable of participant j as $\beta_{.j}$, we have also:

$$\beta_{.j} = \beta + \delta_{.j} , \qquad (4')$$

with the consequence that $var(\beta_{.j}) = var(\beta) + var(\delta_{.j}) > var(\beta)$, if $\delta_{.j}$ is independent of β.

Assume that one performed two experiments, the first one using n1 randomly selected participants, and each participant providing one measure per item, while the second experiment used only one participant providing n2 repeated measures per item. Assuming also that n1=n2=n, we obtain:

$$\rho 1 = var(\beta)/(var(\beta) + var(\varepsilon)/n) < \rho 2 = var(\beta_{.j})/(var(\beta_{.j}) + var(\lambda)/n). \qquad (5)$$

This formally expresses the idea that "analysis techniques that treat individual differences as noise will necessarily overestimate the amount of noise contributing to the mean RT for each word. This overestimation of noise results in an underestimation of the variability that a model should explain, leading to an overestimation of the success of models" (Adelman et al., 2013, p. 1038). Fortunately, the prediction (5) can easily be tested on available data, what we do hereafter.

Tests on empirical data

The data sets

In this section, we test the prediction (5) on two independent data sets. The first data set is the one from Adelman et al. (2013)[1], which consists of word naming times collected for 2820 English words, 4 participants, and 50 repeated measures for each word and each participant. The second data set consists of word naming times collected for 200 French words, 48 participants, and 4 repeated measures for each word and each participant. This last set of RTs was collected during an experiment primarily devoted to the study of ERPs, which was presented in Rey, Madec, Grainger, and Courrieu (2013).

Analysis method

In the first data set (Adelman et al., 2013), we have n1=4 and n2=50, while in the second data set (Rey et al., 2013), we have n1=48 and n2=4. Thus, the condition n1=n2=n is met in none of these data sets, however, given an ICC for n' measures per

item, we can easily compute the corresponding q ratio, and then compute the corresponding ICC for n measures per item using (2'). The q ratio for a given ICC r with n' measures per item is simply given by:

$$q = r / n'(1-r). \quad (6)$$

Using this q value in the formula (2'), we can extrapolate the corresponding ICC for any arbitrary n. For the first data set, we choose n=n2=50, directly obtaining 4 ICCs (one for each participant), and extrapolating 50 ICCs with n'=4 (one for each repetition). This gives a set of 4 ICCs with a random repetition factor, and a comparable set of 50 ICCs with a random participant factor. For the second data set, we choose n=n1=48, directly obtaining 4 ICCs (one for each repetition), and extrapolating 48 ICCs with n'=4 (one for each participant). This gives a set of 4 ICCs with a random participant factor, and a comparable set of 48 ICCs with a random repetition factor.

Before computing the ICCs of the various data tables, we transformed the raw RTs into their z-scores (Faust, Balota, Spieler, and Ferraro, 1999), for each participant and each repetition separately. This transformation allows correcting possible heteroscedasticities and frequently improves the ICC values.

Finally, for each data set, we compare the ICCs obtained with a random participant factor to those obtained with a random repetition factor, using a distribution-free Wilcoxon-Mann-Whitney test. This test being based on ranks, its result is independent of the particular choice of n for the ICCs, and it would be exactly the same using the q ratios as well.

Results

Adelman et al. (2013) data set

The ICCs obtained for the 4 participants (D, A, M, and U) with a random repetition factor, and the ICC 99% confidence intervals are: ICC= 0.8744 [0.8655, 0.8829] for D, ICC= 0.9150 [0.9090, 0.9208] for A, ICC= 0.6927 [0.6710, 0.7136] for M, and ICC= 0.8743 [0.8654, 0.8829] for U, with an average ICC of 0.8391 (sd= 0.0995). It is interesting to compare these values with the "target from hypothetical correct model" estimated using a very different method in Adelman et al. (2013). These estimates appear in their Table 5 (p. 1045), second row, and they are: 87.96% for D, 91.63% for A, 69.81% for M, and 87.17% for U. Clearly, these estimates are almost equal to the corresponding ICCs.

The 50 extrapolated ICCs with a random participant factor ranged between 0.6803-0.8613, with an average ICC of 0.7940 (sd= 0.0373). This is 0.0451 lower than the average ICC with a random repetition factor, and the Wilcoxon-Mann-Whitney test gave $U(4, 50)= 161$, $p < 0.0482$. Thus this result clearly supports the hypothesis (5).

Rey et al. (2013) data set

The ICCs obtained for the 4 presentations of the words, with a random participant factor, and the ICC 99% confidence intervals are: ICC= 0.8372 [0.7914, 0.8765] for the first presentation, ICC= 0.7897 [0.7306, 0.8405] for the second presentation, ICC= 0.8112 [0.7581, 0.8567] for the third presentation, and ICC= 0.7972

[0.7402, 0.8462] for the fourth presentation, with an average ICC of 0.8088 (sd= 0.0209).

The 48 extrapolated ICCs with a random repetition factor ranged between 0.6560-0.9602, with an average ICC of 0.8645 (sd= 0.0777). This is 0.0557 greater than the average ICC with a random participant factor, and the Wilcoxon-Mann-Whitney test gave $U(4, 48)= 51$, $p < 0.0298$. Thus, at new, the result clearly supports the hypothesis (5).

In summary, we observed in the two tested data sets significant contributions of idiosyncratic item effects, corresponding to about 4.51%-5.57% of the item variance, on the average.

What can we conclude from these results?

At this point, there is a great temptation of agreeing with the opinion of Adelman et al. (2013), since their main argument is visibly valid, and this is per se an important result. What have we missed? May be just some trivial consideration such as the fact that the arithmetic mean is an unbiased, consistent estimator of its parent parameter.

In fact, in their reasoning, Adelman et al. (2013) implicitly set n1=n2, which is a prerequisite for validating the inequality (5). However, relaxing this implicit hypothesis, one obtains a quite different picture. Consider the data set from Adelman et al. (2013), let q1 be the q ratio with a random participant factor, and let q2 be the q ratio with a

random repetition factor. For simplicity, we use the average ICCs (with n=50) in the following estimation. Using (6), we obtain:

$q1 = 0.7940 / (50\times(1-0.7940)) = 0.0771$, and $q2 = 0.8391 / (50\times(1-0.8391)) = 0.1043$.

After (2'), one can obtain equal ICC values if $n1\times q1 = n2\times q2$, that is if $n1 = n2\times(q2/q1)$. In the present case, we have $(q2/q1)=1.353$. Thus we can achieve the same precision in item means by collecting 35.3% more measures per item when these measures are provided by distinct participants than when these are repeated measures provided by a unique participant.

Similarly, considering the data set from Rey et al. (2013), we have $q1=0.0881$, $q2=0.1329$, and $(q2/q1)= 1.5082$. Thus we can achieve the same precision in item means by collecting 50.82% more measures per item when these measures are provided by distinct participants than when these are repeated measures provided by a unique participant.

Collecting about 35%-51% more measures per item in the multi-participant case than in the one-participant case is not an insuperable task. For instance, one can achieve the same precision in item means by collecting 50 repeated measures per item from a unique participant, or by collecting one measure per item from each of about 68-76 participants, which is probably easier to do when the number of items is very large.

The above observation contrasts with the definitive assertion of Adelman et al. (2013): "analysis techniques that treat individual differences as noise will necessarily

overestimate the amount of noise contributing to the mean RT for each word". This is simply not the case if one collects an appropriate number of observations per item.

In fact, as a consequence of the consistency of the average estimator, one can always achieve any arbitrary precision level in item means by choosing an appropriate n, provided that q>0. Assume that we have an estimate of q, and we want an ICC equal to r. Then it suffices to choose n using the simple formula:

$$n = r / q(1-r). \qquad (7)$$

Now, assume that in an experiment using n1 randomly sampled participants as the random factor, and in another experiment using n2 repetitions as the random factor (with one participant j), we choose n1 and n2 in order to obtain ICCs very close to 1 in both cases. Then the item means converge towards noise free values in both cases, however, these values are those of the general population $\mu + \beta$ in the first case, while the target $(\mu + \alpha_j) + (\beta + \delta_j)$ includes idiosyncrasies in the second case. What is the most appropriate target for a theoretical model? An idiosyncratic model will provide more detailed predictions for the participant j, but it will probably provide quite poor generalization for other participants. At the contrary, a general model will miss a number of idiosyncratic details, however, its generalization power will be better for new randomly sampled participants. So, the target to be preferred completely depends on the goal of the investigation, and a priori there is no goal better than other ones.

Returning to the criticism of Adelman et al. (2013), we observe that it is based on a confusion between the item mean precision, which determines the proportion of

systematic item variance, and the nature of the target effects (general or idiosyncratic), which just depends on the object of the research. Accounting for a maximum part of the item variance in the general population is not the same thing as accounting for a maximum part of the item variance for an individual. However, in both cases, one can always make the available proportion of systematic item variance arbitrarily large, simply by using an appropriate number of observations per item.

Conclusion

At the end of this commentary, we can confirm the finding of Adelman et al. (2013) that there are stable idiosyncratic item effects in word naming times. This was verified on two independent data sets, in two different languages (English an French). When one averages RTs over participants, the variance corresponding to the idiosyncrasies is transferred into the random variance, which contributes to lower the proportion of systematic item variance in comparison to the case where one averages the same number of repeated measures from the same participant.

However, contrarily to the assertion of Adelman et al. (2013) that the above transfer of variance necessarily leads to an underestimation of the proportion of systematic item variance, simple statistical considerations show that the same precision can be achieved in both collective or individual item means provided that one uses appropriate numbers of observations per item. Experimental results indicate that one must use about 35-51% more observations per item in the collective case than in the individual case, in order to compensate the increase of the random variance in the collective case.

Finally, one must remember that the target item effect is not the same in the collective case than in the individual case, precisely because there are stable idiosyncratic effects. In the collective case, one obtains a general item effect that can suitably generalize to other random samples of participants from the same population. However, in the individual case, one obtains an idiosyncratic item effect that can possibly generalize to an undetermined number of potential participants having similar idiosyncrasies, but that probably poorly generalizes to the general population. It is clear that the better choice completely depends on the goal of the study, but there is in no way matter to disqualify general psychology approaches that average observations over randomly sampled participants, in order to account for general mechanisms of reading.

References

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39(4),* 1037–1053. doi: 10.1037/a0031829

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods,* 39, 445–459.

Courrieu, P., Brand-D'Abrescia, M., Peereman, R., Spieler, D., & Rey, A. (2011). Validated intraclass correlation statistics to test item performance models. *Behavior Research Methods, 43*, 37-55. doi: 10.3758/s13428-010-0020-5

Courrieu, P., & Rey, A. (2011). Missing data imputation and corrected statistics for large-scale behavioral databases. *Behavior Research Methods, 43*, 310-330. doi: 10.3758/s13428-011-0071-2

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychological bulletin*, *125*(6), 777-799.

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.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1, 174*. doi:10.3389/fpsyg. 2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44(1)*, 287-304.

Kristof, W. (1973). Testing a linear relation between true scores of two measures. *Psychometrika, 38*, 101–111. doi:10.1007/BF02291178

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1(1)*, 30-46.

Rey, A., & Courrieu, P. (2010). Accounting for item variance in large-scale databases. *Frontiers in Psychology 1:200*. doi:10.3389/fpsyg.2010.00200

Rey, A., Courrieu, P., Madec, S., Grainger, J. (2013). The unbearable articulatory nature of naming: on the reliability of word naming responses at the item level. *Psychonomic Bulletin & Review, 20(1),* 87-94. doi:10.3758/s13423-012-0336-5

Rey, A., Courrieu, P., Schmidt-Weigand, F., & Jacobs, A.M. (2009). Item performance in visual word recognition. *Psychonomic Bulletin & Review, 16(3)*, 600-608

Rey, A., Madec, S., Grainger, J., & Courrieu, P. (2013). Accounting for variance in single-word ERPs. Paper presented at the *54th Annual Meeting of the Psychonomic Society*, Toronto, Canada, November 14-17.

Note

1. The authors wish to thank Dr. James Adelman who kindly provided his raw data.