Why to treat subjects as fixed effects

James S. Adelman

University of Warwick

Zachary Estes

Bocconi University

Corresponding Author: James S. Adelman

Department of Psychology,

University of Warwick,

COVENTRY,

CV4 7AL,

UK.

Telephone: +44 (0) 24 7615 0233 Electronic mail: J.S.Adelman@warwick.ac.uk

Z. Estes gratefully acknowledges the support of the Center for Research on Marketing and Services (CERMES) at Bocconi University.

Abstract

J. S. Adelman, S. J. Marquis, M. G. Sabatos-DeVito and Z. Estes (2013) collected word naming latencies from 4 participants who read 2820 words 50 times each. Their recommendation and practice was that R^2 targets set for models should take into account subject idiosyncrasies as replicable patterns, equivalent to a subjects-as-fixed-effects assumption. In light of an interaction involving subjects, they broke down the interaction into individual subject data. P. Courrieu and A. Rey's (this issue) commentary argues that (1) single-subject data need not be more reliable than subject-average data, and (2) anyway, treating groups of subjects as random samples leads to valid conclusions about general mechanisms of reading. Point (1) was not part of Adelman et al.'s claim. In this reply, we examine the consequences of using the fixed-effect assumption. It (1) produces the correct target to check if by-items regression models contain all necessary variables; (2) more accurately constrains cognitive models; (3) more accurately reveals general mechanisms; and (4) can offer more powerful tests of effects. Even when individual differences are not the primary focus of a study, the fixed-effect analysis is often preferable to the random-effects analysis.

We (Adelman, Marquis, Sabatos-DeVito, & Estes, 2013) evaluated the processes involved in word naming using a new data set that was somewhat unusual for the field: It had only four subjects, each reading 2820 words 50 times. This design's purpose was to permit us to treat systematic individual differences as replicable statistical patterns requiring explanation. Since there was a significant subject $\times$ item interaction, we broke down the interaction on a subject-by-subject basis. We concluded (1) that current knowledge of the factors affecting reading falls well short of accounting for all the systematic variance in the data; and (2) that when using mega-study data, modelers have compared their models to variance-explained targets that are too lenient. Courrieu and Rey (this issue) argued that (1) single-subject data are not necessarily more reliable than subject-average data (though on their empirical tests, they were in this case) and (2) anyway, averaging over "randomly sampled" participants is more informative about "general mechanisms of reading" (p. XX).

Our arguments did not rely on a comparison of the reliability of the single-subject data to the subject-average data — as assumed in their point (1) — but rather on different approaches to estimating the reliability of subject-average data. The difference between the two approaches results in analyses that correspond to the traditional distinction between a fixed effect (ours) and a random effect (theirs) of subjects, or rather that of the subject $\times$ item interaction. We, of course, do not believe that subjects are fixed effects in the traditional sense (being exhaustive of the population). However, we will demonstrate the counter-intuitive point that although the fixed-effect analysis has narrower scope, it is more useful for assessing general mechanisms, both for the comparison we performed and more generally. Whilst different analyses can be useful for different research questions, the questions that are scientifically or practically interesting in the context of experiments on word naming are better answered by the fixed effect approach.

Decomposition of item variance

The fixed approach and the random approach differ in how they treat subject idiosyncrasies, and different estimates result. Consider an experiment like ours with items (I), subjects (S) and replicates of each subject-item (S $\times$ I) combination. The variance to be explained is equivalent to the mean-squares associated with items, MS_I, which is an overestimate of the variance due to the actual item effect because this estimate is contaminated by variance from subjects and noise. An item-R^2 target says what proportion of the MS_I is due to the actual item effect. The random approach assumes that the target of interest is the *average of the population* from which the subjects are randomly sampled, whereas the fixed approach assumes that the *observed subjects* are the target of interest. The random target is therefore $T_R = 1 - MS_{S\times I}/MS_I$ and the fixed target is therefore $T_F = 1 - MS_W/MS_I$, where MS_W is the variance within each subject-item combination. T_R is designed to treat subject idiosyncrasies as to-be-left-unexplained and T_F is designed to treat them as to-be-explained. If one uses (as we did) a simple signal-noise dichotomy, i.e. $R^2 = 1 -$ proportion noise, then things to-be-left-unexplained are noise. The random approach is therefore equivalent to treating subject idiosyncrasies as (structured) noise. $E(MS_{S\times I}) \geq E(MS_W)$, with equality only when there are no stable individual differences in item effects. If there are such stable differences, $E(MS_{S\times I}) > E(MS_W)$, and $E(MS_{S\times I}/MS_I) > E(MS_W/MS_I)$. That is, the noise estimate for T_R is greater than that for T_F. If — as we will show — MS_W/MS_I is the more useful definition of noise for the problems at hand, then $MS_{S\times I}/MS_I$ is an overestimate of the noise. This is what we meant when we wrote "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). This statement compares two analysis techniques (T_R and T_F) that both generate targets relating to MS_I, that is, for the subject-average item means. Courrieu and Rey have instead generated and tested a hypothesis (Inequality 5) that compares the

magnitude of T_R for the subject-average item means with that of a single-subject target (not T_F), infelicitously referring to this as a claim of ours to justify single-subject experiments (the "second experiment" on p. 5). Their comparison is not informative as to whether to use this single-subject target, because the total variance of this measure is different to the measure with which it is compared (T_R).

We believe their confusion results because (a) this statement was part of our design's justification, but the part being justified was the use of replicates (required to calculate MS_W); and (b) we did ultimately analyze the data on a per-subject basis.

Breaking down the subject $\times$ item interaction constrains models

Our analyses ultimately focused on analyzing each subject separately, because we found a significant subject $\times$ item interaction, and an interaction limits the conclusions that can be drawn from the main effects. Therefore, we omitted to report the overall T_F of 93.22%, which is substantially greater than the T_R that Courrieu and Rey (this issue) report of 79.40%. It may not have been obvious that we were testing the subject $\times$ item interaction because we used Kristof's (1973) method instead of the standard $F = MS_{S\times I}/MS_W$. This F test analysis is of course also significant, but it does not indicate if the interaction was driven by overall speed, so Kristof's test was more informative (others might use a z-scoring approach to remove "general speed," but see Adelman, Sabatos-DeVito, Marquis, & Estes, 2014, for criticism).

The formal argument that it is advisable to break down fixed-effect interactions corresponds to an important substantive point: Individual differences add information that is necessary to properly identify the constraints on general cognitive mechanisms. Consider the length effect, which we considered separately for regular words and for exception words. Such effects are sometimes believed to be absent for naming short words on the basis of Weekes's (1997) study, but mega-study data typically show a small inhibitory effect. In our Figure 1, for exceptions, all showed the inhibitory effect of length, except M, who showed no effect; for regulars, D and A showed the inhibitory effect, M showed a facilitatory effect, and U showed no effect. By taking the fixed approach, M's idiosyncratic pattern adds information that is not

available in the average data. It suggests that the length effect could be associated with a process whose influence is not central to reading. It also refutes any model that necessarily predicts an inhibitory length effect. In contrast, the random-effects logic of "general psychology" as described by Courrieu and Rey would infer from a significant average inhibitory length effect that a correct model would always predict such an inhibitory length effect. Accounting for M's pattern is not a simple matter of adding idiosyncracies to such a model, but instead would require a different set of mechanisms.

The fixed-effect model is the correct comparison for regression models

Most analyses of mega-studies use regression models on item means. Adelman et al. (2013) calculated (a) such regression models and (b) R^2 targets to compare them with. The mega-study regressions are by-items, so they treat subjects as fixed effects. A regression model of this sort will adapt its regression coefficients to match the idiosyncracies of its subject sample as well as possible, and its R^2 will reflect how well it has done so. The correct target for a regression model that takes subject idiosyncracies into acccount (like those normally used on mega-studies) is a target that takes subject idiosyncracies into account, namely our T_F target. T_R is lower and therefore too lenient. We illustrate this (i) with a thought experiment for an ideal case, and (ii) with simulations of a more realistic set of data.

Illustration 1: Thought experiment

Consider some task for which there are two possible strategies. Strategy 1 is affected (facilitated) only by variable A, and strategy 2 is affected (facilitated) only by variable B. Stimuli are selected for a study in which A and B are not correlated, as in an ideal experiment. Replicates are collected (each participant responds more than once to each stimulus), so that T_F can be calculated. Each subject reliably uses a single pure strategy, and there is no trial-to-trial variation in response times. That is, some subjects' RTs are perfectly correlated with A, and the other subjects' RTs are perfectly correlated with B. For a regression model with A and B as predictors of these average data, there are always some coefficient on A and B that can fit the average data

perfectly, so its R^2 will always be 100%. The T_F target is based on the idea that how well each subject replicates *himself* is informative as to how well a model can do; there is no trial-to-trial variation, each subject replicates himself perfectly, so T_F is 100%. This correctly reflects how well the regression model does. The T_R target is based on the idea that how well subjects replicate *each other* is informative as to how well a model can do. When the sample contains some subjects using strategy 1 and some using strategy 2, whose RTs correlate 0, the average correlation will be less than 1. Then $T_R < 100\% = R^2$. T_R is an underestimate that incorrectly indicates that the regression is doing better than possible. Note that the average data offer no way to discern the correct cognitive model. To address any concern that there may be something peculiar to this ideal case that will not generalize, the next illustration is more similar to the word naming situation we have been considering.

Illustration 2: Simulated data

We simulated artifical data from a model that contained the subject-idiosyncratic effects in the form of subject-specific coefficients in a linear model from lexical predictors to response times. For simplicity, we included only three lexical predictors, log. frequency (SUBTLEX + 1), length and a hypothetical to-be-discovered variable with a standard normal distribution of scores over items, which we will call *foo*. The subject idiosyncracies were simulated as follows: Frequency: a $-3 \times \chi^2(5)$ distribution. Length: a $N(6, 30)$ distribution. Foo: a $9 \times \chi^2(2)$ distribution.

For each simulation (representing an experiment), we sampled new subject idiosyncracies and normally distributed trial-to-trial noise. Given the known noise distribution, we used analytic results to produce item-mean R^2 targets under fixed-effect and random-effect models, and compared these to the correct regression model with all three variables, and the "currently-known" regression model with only log. frequency and length. T_R was considered both with and without z-scoring of subjects.

We ran 1000 simulations, in which 4 subjects read each of 2712 items 50 times each, using a noise standard deviation of 250. Across all simulated experiments, the mean subject-average item R^2 of the correct 3-variable model was 83.82%. The

average T_F was 83.77%. The mean absolute deviation between the correct model's R^2 and T_F was 0.35%. T_F thus accuractly indicated what we should expect a model to do.

In contrast, the average T_R was 66.84%, or 67.59% (with z-scoring), which are underestimates of what the correct model achieves. The incomplete two-variable model (without foo) explained 66.73% of the variance on average. T_R would therefore incorrectly label an incomplete set of predictors as adequate to explain the data.

The T_R target leads to the wrong conclusion. It does so because — whether or not one is presently interested in the subject variation — the subject variation in the effects is systematically associated with cognitive processes that are of the same form but different magnitude across subjects.

The fixed-effect model is the correct comparison for cognitive models

Just as multiple regression models have coefficients, cognitive models explaining multiple lexical variables also have numerical parameters. These parameters can be adapated to the idiosyncracies of a data set or indeed individual participants to improve the fit. This is the same property of models that makes T_F rather than T_R correct for regression models (see above). Although modelers sometimes argue that their model parameters are not free, the qualitative mechanisms are the theoretical content of the model, and the parameters could be altered to save the theoretical content if the data required it. Indeed, given there are no truly random samples, parameters *should* be optimized wherever possible (for further discussion, see Adelman & Brown, 2008). So, T_F is also the correct target for cognitive models.

The fixed-effect model is appropriate for some designs of experiment

The fixed-effect model for subjects can also be of use elsewhere. For cognitive modeling, we are sometimes interested in whether a variable affects cognition, and this is not logically identical to the effect being non-zero on average over the population: a variable can facilitate for some people, inhibit for others and we would want to account for it in a cognitive model, even if it averages out to zero. Clark (1973) considered a within-subjects between-items design. As well as rejecting F_1, he rejected F_2, which uses a subject-fixed (item-random) model. He recommended less

powerful tests based on subject-random (item-random) models because he argued F_2 produces significant results for a case (his (4b)) which it should not: When the effect has zero mean but subjects vary around zero. But declaring such an effect not significant says — incorrectly — that the variable should not be accounted for by a cognitive model. F_2 might give a significant result in this case, and this would — correctly — indicate that a model should account for it.

General mechanisms are obscured — not clarified — by averaging over subjects

Statistical generalization of the average across experiments with new subjects is not the same as scientific generalization. The statistical generalizations that Courrieu and Rey (this issue) consider apply only to the mean of the population, but the regularities that govern cognition need not hold for the mean. If, for instance, practice improves response latencies with a power law, but at different rates for each subject, the average data will not show a power law, which is the regularity a cognitive model should predict in this case (Estes, 1956). This problem holds for any cognitive mechanism that is not mathematically equivalent to a linear regression model. As such, individual differences should not be an afterthought in cognitive modeling.

The utility of mega-studies

Finally, Courrieu and Rey (this issue) suggest that our intent is to "disqualify" (p. 5) a wide range of mega-studies from publication. For the purposes of setting model R^2 targets and therefore directly testing cognitive models, we prescribe fixed-effect analyses that require replicates. There is no disqualification of mega-studies with replicates, if they are presented with the prescribed T_F analysis, nor without replicates, if they are used for any of the other purposes of mega-studies (see Balota, Yap, Hutchison, & Cortese, 2012, for numerous examples). That is: mega-studies without replicates should no longer be used for setting model targets, and mega-studies with replicates should use T_F.

Conclusion

Courrieu and Rey (this issue) argued that single-subject data are not always more reliable than subject-average data, but this was not our claim. They do

accurately identify our article as stressing that investigating the structure of individual differences is critically important for assessing models. However, they argue to the contrary that average data can model "general mechanisms." The random effects model this implies provides only limited information about the cognitive mechanisms that are generally true. Treating subjects as fixed effects results in a more accurate picture of what a model can and should account for.

References

Adelman, J. S., & Brown, G. D. A. (2008). Methods of testing and diagnosing models: Single and dual route cascaded models of word naming. *Journal of Memory and Language*, *59*, 524–544.

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*, 1037–53.

Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., & Estes, Z. (2014). Individual differences in word naming: A mega-study, item effects, and some models. *Cognitive Psychology*, *68*, 113–160.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition, Vol. 1: Models and methods, orthography and phonology* (pp. 90–115). Hove, England: Psychology Press.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

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

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *Quarterly Journal of Experimental Psychology*, *50A*, 439–456.