# BEYOND F2 Issue #2, April 18, 2011

While my expert panel gets its act together, I thought it would be useful to summarize the responses I have already received from the “users” list. I will try to organize this around the original topics.

**1.1  What do we gain by using mixed-effects modeling?**

**Jeroen Raaijmakers** has this to say:

“Current statistical software will allow this for both traditional ANOVA models as well as for ME models (see e.g. SPSS). Both approaches are better than the traditional minF' approach since the test statistic is (a variant of) F' rather than minF'. The ME model is preferred since the model that it is based on more closely resembles the assumed linear model (i.e., in the linear model the effect parameter for a random factor is a variance and this is what the program estimates). In practice, it turns out that the two approaches give similar results (in SPSS with complete data exactly the same results). In the literature, there are statements that the ME approach works better when there is a substantial amount of missing data (there will probably be demonstrations of this but I haven't seen them).“

**KIF**: I have heard it said that the code SPSS uses contains errors, but I don’t know whether this makes much difference compared to the output from the *lmer* routine in R. It would be good to know the answer. Also, it would be very useful to run simulations with varying proportions of missing data.

**Hugo Quené** referred us to his paper (Quené & Van den Bergh (2004) Speech Communication), in which the following advantages are discussed:

\* no sphericity required

\* no balanced data sets required (but see your question below)

\* no homoskedasticity required, see below. For example, if the between-item variance is larger in Cond.A than in Cond.B, then this may be included explicitly in the MEM. The gain is that the between-item variances are more realistic. With non-homoskedasticity, the reported p values are inflated (too low), which creates a serious risk of capitalization on chance.

**KIF:** So it would help the novice to know what exactly is meant by “*this may be included explicitly in the MEM”.* I take this to mean that you include (1+Factor|item) as a random effect.

**Roger Levy:** It's much better than minF': minF' is by all reasonable assessments incredibly conservative. Of course, one has to make some assumptions about the underlying generative process giving rise to the data in order to assess nominality of p-values. One of the nice things about MEMs is that they can faithfully mirror something close to the generative structure that our ANOVA analyses implicitly assume to underlie our data. If we use these generative structures to simulate data and analyze them with ANOVAs it immediately becomes clear that minF' is horribly conservative. (Hal Tily and I have some simulations which demonstrate this.) The MEMs (unsurprisingly, given that they use exactly the generative process underlying the simulated data) give nominal p-values on the simulated data.

**1.2  Are observations independent?**

**Jeroen Raaijmakers** comments “Independence is not a required assumption in repeated measures designs. The F-test is still valid provided that the covariance matrix of the repeated measurements has a specific form (compound symmetry). The various corrections basically adjust for deviations from this requirement.”

In response to my comment about Greenhouse-Geisser corrections, **Roger Levy** comments:

**Levy**: “I'm not familiar with the Greenhouse-Geisser correction, but a quick Googling suggests that it is a correction for violations of sphericity. (What I'm more familiar with is using Hotelling's T^2 in this situation.) Actually we're better off with MEMs than ANOVAs in this respect because an MEM with appropriately specified random-effects structure can fit arbitrary covariance matrices and is thus applicable to cases where sphericity is violated.”

“I personally find it conceptually useful to think of \*conditional\* independence as the fundamental relation between stochastic variables (e.g., as I cover in my textbook draft, available at http://idiom.ucsd.edu/~rlevy/textbook/text.html). That is, all assessments of independence are conditional on some given state of knowledge. MEMs model \*hierarchically\* structured data. Let us discard item effects for the moment and consider only subject-level effects. In such a situation, the underlying assumption of the mixed-effects model is that two observations arising from the same subject are conditionally independent \*if and only if\* you know the proclivities of the individual subject (e.g., slow in condition A, slower in condition B) in each condition. Since we don't have veridical knowledge of these proclivities, the observations are not conditionally independent; but MEMs take this lack of independence into account by modeling these subject-specific proclivities and taking them into account in inferences about the "fixed effects". And this ability of MEMs to explicitly model extends to the case where we have multiple crossing hierarchies (e.g., subjects and objects).

**KIF:** Just to make sure that we are all on the same page, **sphericity** is obtained when “the variances among all possible pairwise differences of the repeated measures factor are equal” (don’t know where that definition came from). So this has to do with homogeneity of covariances between each level of a repeated measures factor and all other levels. So what if the variances are not homogeneous? Is this also a violation of sphericity, or is this heteroskedacity? I’m confused because in my notes it says the following:

“A model lmer(RT ~ A + (1|Subj) + (1|Item)) assumes homoskedasticity of the variance across the levels of A (so sphericity).” This implies that sphericity and homoskedacity are equivalent terms.

**Alan Garnham** comments**:**

“There are several aspects of mixed-effect modelling that make me feel very queasy. One is the independence issue, and how it affects interpretation, though I don't think repeated measures is a huge problem, because MEM gives you much more freedom in modelling the variance-covariance matrix - you can estimate every term separately if you want to (losing DFs in the process).

**KIF**: When you say "every term", what are you referring to (remember that I'm trying to discuss these issues in terms that everyone will understand).

**Garnham**: “I mean every element in the variance-covariance matrix. It's hard to know what people do and don't know, and the obvious way to say what it means is to use technical terms. I guess pretty much everyone knows that traditional ANOVA assumes homogeneity of variance (for both repeated measures factors and between-S/between-I factors). Those variances are the (lead) diagonal elements in the matrix. In MEM you don't assume they are the same. I don't know how this works in the Bates (lmer) procedures (ie whether its the default - I'd guess it is), as I've only looked at them very briefly. But stepping back from the software, the way to estimate the different variances is from the sample variances.”

I also hope people know about the sphericity assumption for repeated measures. This is approximately a homogeneity of covariance assumption (among the levels of the repeated measure). Covariance isn't an issue for levels of a between factor. In MEM again you can estimate each covariance from sample data. In SPSS the default covariance structure for repeated measures is called DIAG and has heterogeneous variance but all covariances zero. SPSS allows a relatively small number of possible covariance structures (see manual or online help) including UNSTRUCTURED - all elements potentially different.”

**KIF**: I hope someone will explain what it means that SPSS assumes that the covariances are zero. Does lmer (MEM) cope with unequal covariances? (More on this later). As for the sphericity issue, those of us who applied the Greenhouse-Geisser correction when a repeated measures factor had more than two levels will recognize this problem.

In answer to the question “are observations independent?”, **Quené** answers:

Yes. This needs to be verified by inspection of residuals. If there is additional correlation between factors, due to subjects, then these additional random effects could also be included in the model, e.g. as  
 resp~1+Condition+(1|Subject)+(0+Condition|Subject)

etc.

**KIF:** So here it is important that we understand MEM syntax. So (1+Condition|Subject) means that we allow for the possibility that subjects differ in the effect of the factor Condition. My notes tell me that (0+Condition|Subject) is used if Condition is a covariate, not a fixed effect. Is this correct?

**Quené:** It is indeed assumed that the same subject yields the same individual component for every response s/he produces: Subject i is +200 ms relative to the population mean, for all his/her responses. In a simple repeated-measures design there are therefore two random components in MEM, 1. the individual subject effect, and 2. the residual or within-subject effect capturing noise within subject. In a less simple psycholinguistic experiment, there are not two but three random components in MEM, 1. subject, 2. item, 3. residual. This is difficult to estimate: data from multiple subjects are used to estimate item components, and vice versa.

**Sarah Brown-Schmidt poses a further question:**

My question for you and your panel relates to your question #1.2 regarding the independence of observations. There is certainly a non- independence of samples taken from the same subject; it seems like MEM has a way of handling this issue, and I look forward to hearing more about it. What, however, can be done about the non-independence of one response to the next-- that is, response n+1 should be more highly correlated with response n+2 than response n+15. This might be a minor problem in say, object naming, but scales up to being a massive one in studies of fixations over time, where sampling rates are now high enough (2000hz) that fixation position from one sample to the next is very highly correlated.

**KIF:** Given that we have been routinely including the effect of PreviousRT in all our analyses (and it’s always highly significant), I certainly hope that MEM does cope with this problem. Is there a limit on how high the correlation can go?

**1.3  How well does lmer() cope with unbalanced designs?**

**Raaijmakers**: Technically, the term for such a procedure is "cheating". More generally, you would then let the data affect the design used instead of choosing the design before collecting the data.

**Levy:** Well, if in your analysis you allowed for the possibility that different conditions may have different effects on different lists (either as random effects or fixed effects), then yes, lmer would effectively weight the observations in this situation and tell you that your treatment effect fails to generalize across lists.

**KIF:** Good point. I’ve been wondering whether Lists needs to be included in the model when we have counterbalanced designs. In traditional F1.F2 analyses, it’s absolutely critical, since the treatment effect can actually reverse in direction if the items were not carefully matched. Initially, I thought it was necessary to include Lists as a variable in MEM, but it made no difference at all. But I can see that in this case, it would make a difference. In a counterbalanced design, I assume Lists would have to be a fixed effect. So would the model look like this? :

RT ~ FactorA \* Lists + (1|subject) + (1|item)

**1.4  Getting a significant treatment effect by including extra variables as predictors.**

**Raaijmakers**: In order to avoid making decisions about the model to be used dependent on the data observed, I generally prefer to use the most elaborate model (i.e. the model with all of the factors included that were in the experimental design used) unless there are a priori reasons to prefer a less complete model (such as deleting a nuisance factor from the design in order to increase the degrees of freedom for the test that you are really interested in). Including such factors indeed removes part of the error variance but that does not necessarily mean that it will affect the results. That will depend on the ratio of the error variance removed to the number of degrees of freedom lost. But generally speaking I would not advocate a procedure to do both analyses and then using the most convincing one. That would introduce too much bias in the procedure.

**Levy:** On average, adding control predictors that actually have explanatory power for your dependent variable should help you find effects. For any particular dataset, however, you can never know for certain how much of the effect of adding a control predictor on inferences about a predictor you really care about arises from idiosyncracies of your particular data (i.e. noise). If a given control increases your significance in half your studies and decreases your significance by equal amounts in the other half, I would tend to be skeptical!

I think that you're indirectly alluding to another issue, though, which is that the flexibility afforded by MEMs allows a researcher to try many different analyses. There is an issue of trust that a researcher is presenting the model that best reflects the data rather than the model that best supports the story one wants to tell. Each additional control variable one considers introduces more possibilities, and a bunch of "controls" some of which happen to help your story and some of which don't may be tempting fruit indeed.

**KIF:**  This makes it difficult to evaluate evidence of a non-significant treatment effect – did the researcher try hard enough to find a suitable model?

**Levy**: Yes, but this has \*always\* been the case -- e.g., in ANOVA analyses of reading-time studies for syntactic comprehension whether you use raw reading times or first residualize RT on word length.

**KIF:**  There is a another related issue here. I am sometimes asked whether MEM could do the same job as ANCOVA. Suppose that frequency is confounded with a main effect – could one remove this confound by including frequency as a covariate, or by allowing random slopes of frequency for items?

**1.5  Should we start with more complex models and prune them, or go from simple to complex?**

**Raaijmakers:** My own opinion is quite different: do neither unless there are problems with the precision (=degrees of freedom) with which a crucial factor is measured.

**Quené:** My recommendation is to work from simple to complex. If a main effect is added, then start by including all its interactions with the previous effects already in the model. Remove interactions that are not significant. If any term is included or maintained in the fixed part, then you should subsequently assess whether it should also be included in the random part (i.e. whether it is indeed constant=fixed over subjects and/or over items). However this is not easy to assess using lmer...

**Levy**: Actually, this question contains several implicit presuppositions that I am not sure I am ready to accept:

\* we should determine the final model on whose basis we report results by a process of model selection;

\* that model selection procedure should either be forward stepwise or backwards stepwise;

\* we are all sure what the right criterion is for whether to take a step forward or backwards, and which step to take.

I think that model selection is one of the most fraught issues out there for data analysis. I personally am coming to favor methods that avoid model selection whenever possible, but it's not clear that model selection can always be avoided. So this is an open issue to me.

**1.6  What happens when a main effect disappears when we ask for an interaction?**

**Raaijmakers: “**Normally (i.e., in balanced designs) the interaction will not steal variance away from the main effects. After all, the SS for the interaction is what is left over after subtracting the SS for the main effects. What probably happens in such cases is that the interaction is very small (explains little) but steals degrees of freedom from the error term, resulting in a higher MSe.”

**KIF:** That could explain such an outcome, although probably not in this case. I should have provided more detail. The task was semantic categorization. The main effects were 'prime' with three levels (ID, congruent, and incongruent), and 'cat' (exemplar - nonexemplar). The covariate 'pleng' was the length of the prime in letters. Because I wondered whether the problem had something to do with 'pleng', I dichotomized it as 'plengf'. The random effects were 'subj' and 'itemN'. When I asked for the prime\*pleng interaction, the effect of prime was negligible. But when I remove the interaction term, the effect of prime was significant.

Roger Levy kindly provided a straightforward explanation:

**Levy:** I believe that the prime mover in your example is that pleng is not centered. In general, if one or more of the predictors entered into an interaction is not centered, then the main effects of those predictors lose their original interpretability. In your specific case, when you have prime\*pleng, the main effect of, e.g., the main effect of the cong<->ID contrast means the hypothetical effect of this contrast for a length-zero word. If you center pleng before doing the analysis then the meaning of the cong<->ID contrast becomes the effect of this contrast for an \*average-length\* word:

words2$Cpleng <- scale(words2$pleng)

and now the main effects for prime look basically the same as they do for the model with the interaction (I believe the slight discrepancies between the main-effect results for the two models arise from the slightly different distributions of pleng among the prime conditions.)

**KIF**: So we should always center a covariate – that much I understand. But I still don’t quite understand. Centering means converting the raw values to z-scores (that’s what scale() does). But this is what happens anyway when you compute a covariance, so why does that make a difference?

**Levy**: In general, checking that the predictors in question have been centered is the first place I would suggest looking when one finds that lower-order effects change when a higher-order interaction has been introduced. And one must always keep in mind that the \*meaning\* of a "main effect" really does change when a higher-order interaction is included! It is only in the special case of centered predictors with fully balanced data and linear models that the meaning of "main effect" doesn't change when you include the interaction.

**Kathy Rastle** had a similar problem, where including an interaction wiped out the original main effect, but centering didn’t help. She was looking at stress assignment in reading aloud (i.e., was it first or second syllable stress?). This could be expressed as proportion and analysed using linear models but the right way to do it is treat it as categorical and analyse using “family = binomial”. But this is where the interaction problem rears its head.

**1.7  Model-based trimming of outliers.**

**Raaijmakers: As best as I know, ALL methods of trimming have their problems. E.g. trimming based on the within-cell standard deviation may not catch all outliers simply because the outliers increase the within-cell standard deviation.**

**Levy** agrees with my view that one should not include the experimental conditions in the trimming model.

**Levy**: Yes, in general using experimental-condition-specific criteria for outlier removal will lead to anti-conservativity. You can show this through simulations. This issue is orthogonal to MEM versus ANOVA though -- e.g., some papers report condition-specific outlier removal criteria in ANOVA analysis of self-paced reading data, and this is also anti-conservative.

**KIF**: Isn’t it the case that if you applied condition-specific outlier removal iteratively, you would eventually manufacture any effect you wanted?

**1.8  Violations of the assumption of normality.**

**Raaijmakers: This discussion has a long history. Some experts have warned against the danger that nonlinear transformations will distort interaction effects, others have claimed that there is nothing wrong with transformations since "the numbers do not know where they came from" and all that is required for a valid test is that the numbers conform to the model used to analyze the data.**

# New Questions and Issues

As I anticipated, new questions are emerging as we go along. Here is the first batch.

**1.10 How do we evaluate and report effect size?**

**Quené**: One additional question is about effect size. How do we evaluate and report effect size? I've been searching the literature for useful suggestions but found very little. Effect sizes are difficult to assess in MEM because there are multiple random components (error components) to consider, not just a single one. It makes sense that the effect size depends on the random part of the MEM, but there are fewer established conventions regarding the composition of the random part in MEM than there are in RM-ANOVA. Moreover, there may be good reasons to expand or complicate the random part of the MEM.

**1.11 What is the best way to run multiple comparisons to follow-up on an omnibus test?**

**David Plaut**: “What is the best way to run multiple comparisons to follow-up on an omnibus test?  For instance, imagine you have a 2x2 factorial design with factors A, B, and AxB.  how would you explore the nature of a significant interaction term in an omnibus model to determine if it is due to A1 and A2 differing only on B1, or if they also differ significantly at B2?  It seems that there are two main issues to consider when answering this question: 1) how to leverage information from the omnibus model to conduct more sensitive/powerful analyses, and 2) how to control type-I error rates when conducting multiple comparisons.   
  
With respect to 1) in a standard ANOVA, there are several options available, including simple effects or t-tests using the presumably more reliable variance from the omnibus test.  What comparable options are available (in practice as implemented in R and/or in theory) for LMERs?  It seems that they would be ideally suited for leveraging omnibus-level information to make more sensitive tests (e.g., reliable estimates of participant slopes/intercepts to name but one), but how to do so exactly seems unclear.  For now, most others I've talked to suggest that the best option presently is to re-analyze from scratch each subset of the data with a restricted version of the omnibus model (e.g., only the main effect of A when examining differences in A across a single level of B) but everyone also seems to agree that this is not ideal.   
  
With respect to 2) it is unclear how to properly correct for multiple comparisons in the standard fashion (e.g., multistage bonferonni) since the degrees of freedom are undefined for LMERs.  Using the conservative t-test approximation, while not unreasonable for conducing a single comparison in a large data set, intuitively seems problematic in this case because small imprecisions can scale rapidly -- perhaps even exponentially depending on the correction -- across tests with fewer observations.  Is there a way to use the Monte Carlo sampling techniques or other techniques to obtain more accurate p-values? We're also not sold on the idea of dropping non-significant but theoretically relevant factors from the model to boost power on other effects -- it seems a little bit too much like biasing the analysis to find what one is looking for....”

**1.12 The importance of analyzing the random effects**

Alex Fine raises a query about what appears to be standard practice.

**Fine**: A \*major\* problem with the way MEMs are being used is that many researchers confidently report that they used "mixed effects linear regression with subjects and items as random effects". This is a terrible general practice, and I would bet a large sum of money that it is leading to spurious results being accepted by the community. The reason is that this random effects (RE) structure only adjusts for   
differences in the average (say) reaction time for subjects and items but does not address the possibility--somewhat crudely speaking--that an effect of predictor X could by driven by, e.g., a subset of the subjects. A random by-subject slope for X would address this, but very few people who use MEMs right now do this. Reviewers at journals, etc., \*must\* set a higher standard for the treatment of random effects. When I use these models, I always include "the maximal random effects structure justified by the data based on model comparison". In my (limited) experience, 9 times out of 10 this RE structure is NOT just random intercepts for subject and item and, moreover, adding random slopes very often weakens or completely gets rid of effects that were there when the model just had random intercepts.

**KIF**: Harald Baayen has made the same point, and will no doubt add his voice to this chorus.

**1.13 Obtaining p values for complex models**  
  
**Fine**: A lot of people don't like to explore the random effects structure because, currently, it's hard to get p-values with MEMs that have RE structures more complex than just random intercepts. And of course you can't do anything in this field without p-values. You always get t-values, and for large data sets you can safely assume that a t-value of 1.96 or above is significant at the .05 level. I've had some luck reporting significance this way in proceedings papers, but have a feeling it may not fly in journals. So we need to either loosen our adherence to the p-value as the criterion for statistical reliability, or we should make a priority of implementing the same kinds of MCMC sampling algorithms for getting p-values from models with more complex RE structures that are currently implemented for models with only random intercepts.

**KIF**: Do this mean that the p values that we get from **pvals.fnc** are not appropriate for models with more complex RE structures?

**1.14 The effect of transformations**

If a log transform of the raw RTs is used, the estimates of the effect of a given variable are tiny when converted back to msec values. Obviously, something has happened to the original scale as a result of the computations involved in MEM. Is there a solution for this? Or should we report estimates based on untransformed RTs?

KIF

4/18/2011