Methods in Psychology 6 (2022) 100092

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.sciencedirect.com/journal/methods-in-psychology

Using linear mixed models to analyze learning processes within sessions improves detection of treatment effects: An exemplary study of chronometric mental rotation

Leonardo Jost *, Petra Jansen

Faculty of Human Sciences, University of Regensburg, Universitätsstraße 31, 93053, Regensburg, Germany

ARTICLE INFO

Keywords:
Practice effects
Linear mixed models
Mental rotation

ABSTRACT

Practice effects occur for many cognitive tasks. They are observed not only between repeated tests, but also within sessions. They can confound the detection of treatment effects, even when compared with control groups. We present an approach to reduce the impact of within-session practice effects through inclusion and manipulation of time in linear mixed models. With an exemplary mental rotation task, we demonstrate the possibility to investigate performance within sessions. We show how controlling for practice effects can improve comparisons between treatments. Through simulations, we demonstrate the impact of within-session practice effects and the usefulness of the presented approach.

1. Introduction

Traditionally, intervention studies using within-subjects repeated measures designs were commonly analyzed with ANOVAs. More recently, linear mixed models and general mixed models[1] became more widely spread, in part due to the increased availability of computing power and easy to use packages (Bates et al., 2015). The general advantages of mixed models are well documented but there exists some debate on the choice of specific parameters, especially the choice of random slopes (Barr et al., 2013; Bates et al., 2015; Brauer and Curtin, 2018; Matuschek et al., 2017). For overviews of the theoretical background and applications of mixed models, see, for example, Baayen et al., 2008, Barr et al. (2013), or Brauer and Curtin (2018), or a more beginner-friendly practical tutorial of Winter (2013).

In short, mixed models employ both random and fixed effects. There is no clear definition of random and fixed effects. Typically, random effects are expected to generalize to a more general sample (e.g., generalizing data of randomly chosen participants to the general population). Fixed effects include all items of interest (e.g., all treatments of interest in one study). Each random effect can furthermore be associated with random intercepts and random slopes. While random intercepts account for baseline differences, random slopes account for possible differences in the interaction of fixed and random effects. Some of the main advantages of mixed models are the treatment of multiple crossed or nested random effects, the possible inclusion of multiple covariates, and the possibility to analyze unbalanced and partially missing data, all while achieving higher statistical power.

Moreover, mixed models offer possibilities to analyze the progress during the time course of an experiment (Baayen et al., 2008; Mirman et al., 2008; Winter and Wieling, 2016). The change over time can be of primary interest but also an important covariate.

Here, we demonstrate how accounting for time within sessions in pre-posttest designs can improve control over practice effects and aid the interpretation of results. In traditional analyses, a treatment effect is combined with (unwanted) practice effects in a treatment group. This is compared to a control group, which should only be influenced by practice effects. However, the isolation of practice effects in control groups is still not optimal and the variance of practice effects between participants can impact the detection of treatment effects. Moreover, when performance differences between groups are detected, it is clear that the interpretability of posttest differences is impacted. The same should be clear for differences in practice effects between groups, which could in theory occur equally often but are rarely investigated. We present an approach, which through inclusion and manipulation of time within sessions estimates practice effects within sessions and has the following advantages:

* Corresponding author.
E-mail address: Leonardo.jost@ur.de (L. Jost).

[1] Often also called linear mixed effect models, hierarchical linear models, or other similar names.

https://doi.org/10.1016/j.metip.2022.100092
Received 18 February 2021; Accepted 28 April 2022
Available online 2 May 2022

First, it allows better detection of the magnitude and change of practice effects, which can in themselves be interesting.

Second, by the separating practice and treatment effects, it allows better estimation of the "true" treatment effect on test performance.

For this, we have performed multiple analyses on the openly available dataset of Jost and Jansen (2020) [2] of a chronometric mental rotation test. In this test, participants rotate abstract objects in the mind. This rotation takes longer if it is performed over a larger angular disparity (Shepard and Metzler, 1971). For these tasks, improvements by repetition have been shown within and between multiple sessions. These practice effects are an unwanted influence in repeated testing designs and are observable for many other cognitive tests (Calamia et al., 2012; Goldberg et al., 2015).

We have split the data to represent multiple testing sessions or a possible treatment session consisting of the repetition of the task. In the following analyses of these sessions, we investigate the detection of a treatment effect compared to no treatment between sessions or to correctly identify the null effect if no treatment was used. We compare an analysis using the suggested inclusion and manipulation of time within sessions to a more traditional analysis using only the testing sessions.

In the following, we will first introduce the general principle and show the possibilities of analyzing learning within one group and multiple testing sessions. In a second step, we will simulate two groups in a pre-post design and show how the analysis of time can aid interpretation of treatment effects even where traditional analyses might fail. We provide some constructed cases to demonstrate potential problems as well as simulations regarding their occurrences in practice. All code to recreate the demonstrated analyses is available as a GitHub repository: https://github.com/LeonardoJost/TimeAnalysis.

2. Method

2.1. Description of the dataset

We have used the publicly available dataset of Jost and Jansen (2020). In their results, reaction time shows easier interpretation and larger effects than accuracy and is typically the main variable of interest in mental rotation tasks. Thus, we will focus on the analysis of reaction time. The dataset contains a total of 15,525 observations of 41 participants and 16 different stimuli (item types) for the analysis of reaction time during an experimental session lasting 30 min. For each observation, the time of the answer since the start of the session (time) is reported. Next to the main variables of interest, the dataset also includes other variables (degree, side, experience), which are included as covariates but are not of primary interest here.

To simulate the effects of repeated testing, this dataset was split into three blocks of 10 min by the time of the answer of the participants. These blocks were analyzed as simulated pre- or posttests or as a training condition employing the repetition of the task. As a result, for comparison between adjacent blocks we expect no improvement. This comparison was used either as a control group or as a treatment without effect. For comparison between the first and third block, we expect to find an improvement by the repetition of the task in the second block, that is, the treatment.

2.2. Statistical analysis

Statistical analysis was performed similarly to Jost and Jansen (2020) with linear mixed models using lme4 package (version 1.1–23; Bates et al., 2015a,b) in R (version 4.0.3; R Core Team, 2018). Model parameters were estimated by maximum likelihood estimation using bobyqa algorithm wrapped by optimx package (version 2020–4.2; Nash and Varadhan, 2011) as optimizer. P-values were obtained by using likelihood ratio tests to test for improvement of model fit by inclusion of the fixed effect of interest and compared to a significance level of .05.

We do not report effect sizes because the relevant effect sizes for the interpretation have already been reported by Jost and Jansen (2020). For the application of the ideas presented here, the detection of effects is the deciding factor not the actual size of the effect. Although the existence of effects differs from significance (Amrhein et al., 2019) we will use significance to judge whether an effect will be found or how interpretable effects are.

The data contains two random effects: the participants and the item type and we used random intercepts and random slopes for degree and time by participant and random intercepts by item based on the suggestion of Matuschek et al. (2017). For the analysis of only blocks, we replaced the random slope by time by a random slope by block. For the fixed effects, we included the main effect of experience and the interaction between degree and side as these proved to be significant and possible covariates in the original dataset. For comparison purposes, these were included in all analyses regardless of their significance.

To compare the effects of time, we performed two analyses for each dataset. In one analysis, we excluded all effects of time and analyzed the interaction between degree and block. In the other analysis, we included the effects of time and analyzed the interaction between degree, block, and time. Degree was included as it is the main moderator of difficulty in mental rotation tasks and larger improvements for larger angles have been carved out (Jost and Jansen, 2020). For the comparison of different groups, we analyzed the additional interaction with group.

Starting from these models, non-significant fixed effects were stepwise removed from the model, such that effects which least decreased model fit were removed first and a model containing only significant fixed effects remained. This was performed only for the examples but not for the random simulations. The analysis of main effects contained in significant interactions was performed according to Levy (2014). Degree was centered such that main effects show the average improvement over all angles. Time was normalized such that time 0 was set to the end of the first block and main effects of block would indicate additional improvements between the first and the next block that could not be explained by improvements over time.

For the visualization of the data and the changes within sessions, we use generalized additive models as the possibly best representation of true effects. These were implemented using ggplot2 package (Wickham, 2016) in R (R Core Team, 2018). A visualization using linear improvements or block-wise approximation would be biased for one of the approaches discussed here.

3. Results

3.1. Analyses of performance within one group

3.1.1. Analysis of all blocks without treatment

First, we start with the full dataset. The overall improvement over time during the separate blocks is shown in Fig. 1.

For the analysis of only blocks, there is a significant improvement between blocks (χ^2 (2) = 48.77, p < .001). With the inclusion of time, the main effect of block is not significant anymore (χ^2 (2) = 1.97, p = .373). The improvement over time can explain all improvements over the blocks but, in this case, does not vary significantly between blocks (χ^2 (2) = 0.33, p = .848). As a result, the improvement over time can also interpolate the improvement during the second block.

Note that traditionally, one could include time as a covariate by centering time within each block. This reduces the variance within blocks by accounting for these improvements. However, this effect is quite small in the present data and does not change the significance of the results, as still the average performance within each block is compared. As the effect of time does not vary between blocks, the choice

[2] The data is available at https://osf.io/dr9mv/(DOI 10.17605/OSF.IO/DR9MV).

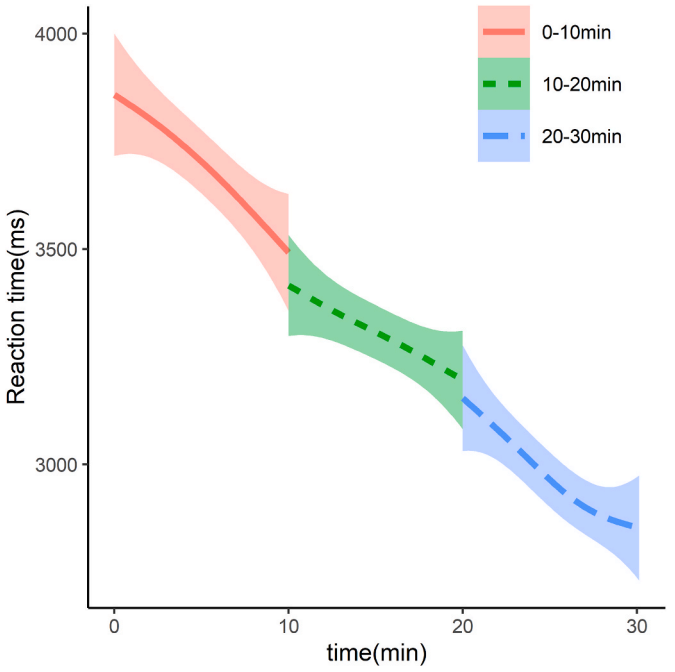

Fig. 1. Changes of reaction time over the full session, separated into three blocks.

of centering only influences the intercept. However, in the case that the effect of time differs between blocks, the choice of time 0 needs further consideration. In this case, the main effect of block describes the difference between blocks at time 0 (Levy, 2014). Thus, for the following analyses, time was normalized such that time 0 was set to the end of the first block and main effects of block would indicate additional improvements between the first and the next block that could not be explained by improvements over time.

3.1.2. Analysis of a treatment between two blocks

Now, we look at only the first and the third block, where the second block serves as the treatment. We remove the time gap between the blocks and move the third block 10 min forward in time. For the analysis of only blocks, there is a significant improvement by block (χ^2 (1) = 49.47, p < .001). With the inclusion of time, the main effect of block remains significant (χ^2 (1) = 46.95, p < .001). This shows that the repetition of the task in the second block causes an improvement, which can no longer be explained if the time of the second block is not accounted for. The effect of block in the time analysis now describes the difference between the end of the first block and the start of the third block. That is, when compared to a hypothetical situation without a second block there is a significant improvement, the treatment effect. This type of analysis will thus be used to evaluate the effectiveness of treatments.

3.2. Analysis of treatment effects between groups in a pre-post-design

Now, we will assign participants to groups and compare the detection of treatment effects. We will start by demonstrating the general principle and follow with special cases, where the groups differ in performance (as is often observed in real studies) and in their improvements within sessions (which seems equally probable but is rarely measured).

3.2.1. Analysis of a treatment between groups

We start with a case, where treatment and control group show a comparable pretest performance. For this, we duplicated the data and used one set for the treatment group and the other set for the control group. Both groups use the first block as pretest. The control group uses the second block as posttest and the treatment group uses the third block as posttest (thus including the second block as treatment). Again, block three is moved forward in time by 10 min. The improvement over time for this dataset is shown in Fig. 4. Not surprisingly, both types of analyses identify a treatment effect in the form of a significant interaction of block and group (χ^2 (1) = 12.36, p < .001 and χ^2 (1) = 22.94, p < .001). In such a case, on would typically perform separate analyses of each group. Here, these reiterate the results of the previous two analyses. Using only blocks for the analysis there is a smaller improvement in the control group while including time shows no significant improvement in the control group by block. Both analyses reveal, that the treatment indeed improves performance.

3.2.2. Analysis of a treatment and between-groups differences in pretest performance

Now, we assign groups to participants based on their performance in the first block (pretest). Typically, treatment effects are harder to interpret in these cases. On the one hand, one expects a larger improvement in the treatment group compared to the control group. On the other hand, one would expect smaller improvements for better performers.

3.2.2.1. Better performers in the treatment group. All participants with faster than median reaction time are assigned to the treatment group while the slower participants are in the control group. Both groups use the first block as pretest. As before, the control group uses the second block as posttest and the treatment group uses the third block as posttest. Block three is moved forward in time by 10 min. The improvement over time for this dataset is shown in Fig. 2.

For both analyses, the interaction between block and group is not significant (χ^2 (1) = 1.42, p = .234 and χ^2 (1) = 0.08, p = .778). As expected, the improvement is smaller for better performers and this cancels out the treatment effect. The analysis of time shows a significant main effect of block and a significant interaction of time, block, and group. These could be investigated further to identify possible reasons for the null effect of the treatment. However, both analyses should reach the same conclusion: More research is necessary in such a case.

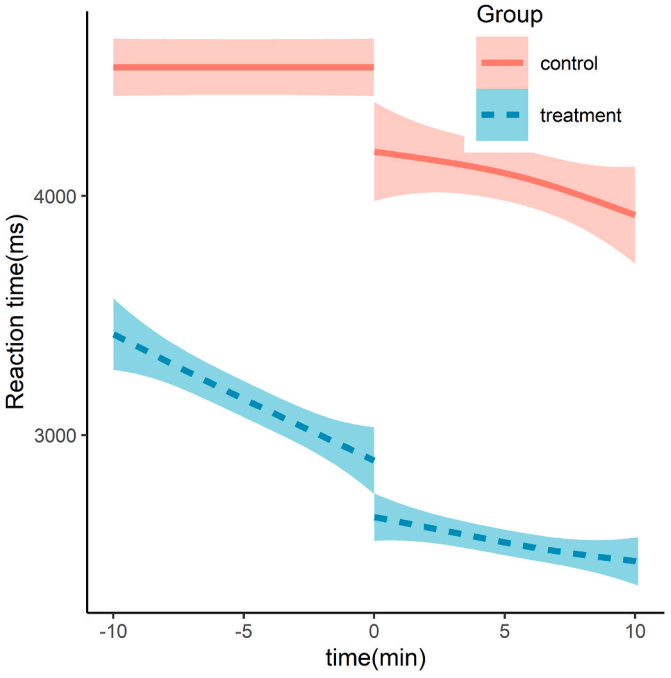

Fig. 2. Changes of reaction time within sessions, separated by groups. Groups differ in pretest performance with better performers in the treatment group. Pre- and posttest are separated at time 0.

3.2.2.2. Better performers in the control group. This dataset resembles the previous one but the slower participants are now in the treatment group and the faster participants are in the control group (see Fig. 3). Both analyses show significant interactions of block and group with larger improvements for the treatment group (χ^2 (1) = 13.228, p < .001 and χ^2 (1) = 37.75, p < .001). In the analysis of the control group separately, the inclusion of time can explain the improvement of the control group (χ^2 (1) = 0.03, p = .864), whereas the analysis without time also shows significant improvement in the control group (χ^2 (1) = 20.49, p < .001). This indicates, that the improvement in the control group is only due to a practice effect and indicates a treatment effect. However, again, in such a case more research should be conducted.

3.2.3. Analysis of a treatment and between-groups differences in practice effects

Before, we separated the participants by their performance. Now, we will separate them by their improvement from the first to the second block, that is, their practice effect. These are very constructed worst cases for the analysis of only blocks because it cannot distinguish between practice and treatment effects.

We will start with assigning those participants to the control group, which show the largest practice effects (see Fig. 4). In this case, the analysis without time shows a significant interaction of block and group (χ^2 (1) = 4.66, p = .031). The coefficients however lead to the wrong conclusion: They reveal a larger improvement in the control group. The analysis including time on the other hand can correctly identify the larger improvement for the treatment group (χ^2 (1) = 4.17, p = .041).

In another example, we assign the participants with larger practice effects to the treatment group but without an actual treatment. That is, we use the first two blocks for both the control group and the practice group and there should be no treatment effect (see Fig. 5). The analysis without time nevertheless shows a significantly larger improvement for the treatment group (χ^2 (1) = 30.29, p < .001), which can be explained with the inclusion of time (χ^2 (1) = 0.12, p = .732).

3.2.3. Random group allocations

The previous cases were all constructed extreme cases. To show possible occurrences in reality, we have performed two times 1000 random simulations of group allocations. One set of simulations used the second block as treatment (block one and three as pre- and posttest) and the other set used no treatment (block one and two as pre- and posttest). For each of the simulations, 20 or 21 of the 41 participants were randomly selected and assigned to the treatment group. The other participants were assigned to the control group. For every simulation, we

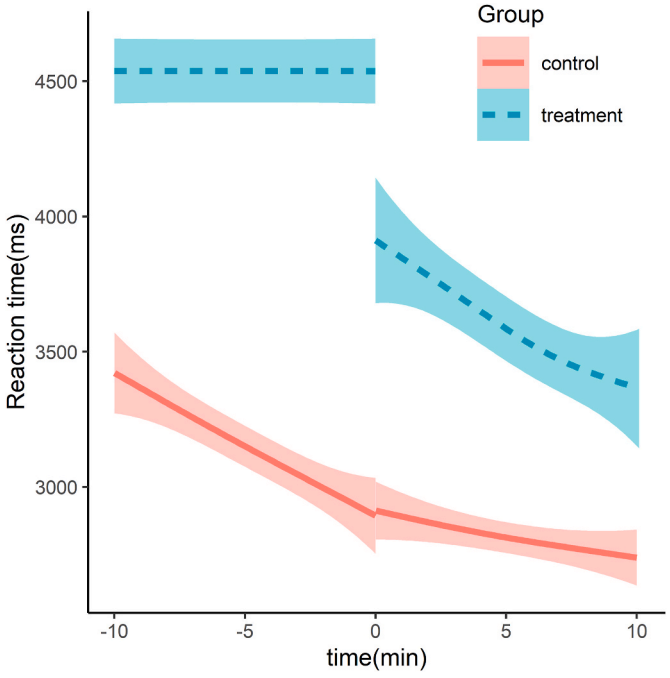

Fig. 3. Changes of reaction time within sessions, separated by groups. Groups differ in pretest performance with better performers in the control group. Pre- and posttest are separated at time 0.

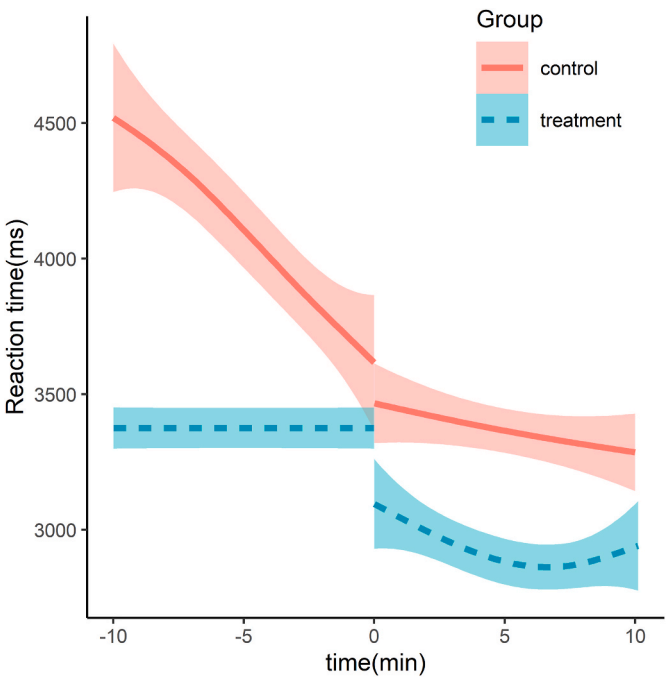

Fig. 4. Changes of reaction time within sessions, separated by groups. Groups differ in improvement with better learners in the control group. Pre- and posttest are separated at time 0.

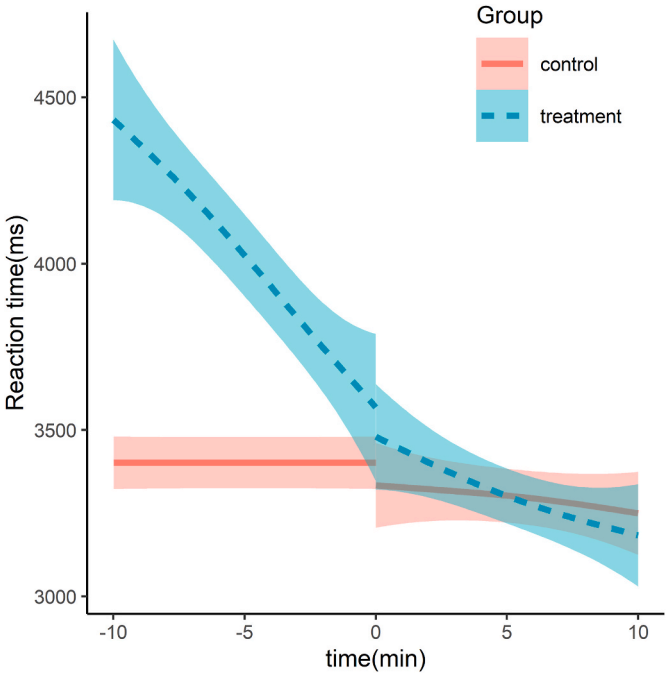

Fig. 5. Changes of reaction time within sessions, separated by groups. Groups differ in improvement with better learners labeled as treatment group but without an actual treatment. Pre- and posttest are separated at time 0.

compared the detection of a significant treatment effect between groups for both discussed types of analyses. In cases of convergence issues in any analysis, a new random allocation was generated. A summary of the significance of p-values is shown in Table 1. For comparison, we also included an analysis employing the change over time within sessions as a covariate by centering within each block. This analysis however only differed marginally from the analysis using only the blocks (differing in significance for only six out of the 2000 simulations). This shows, that the differences in the results of the proposed approach are not simply due to the inclusion of time but also the choice of the manipulation of time between blocks.

3.2.3.1. Simulation of treatment. Overall detection rates of a significant treatment effect were 75–80% in each analysis in line with sufficiently powered experiments. In no case was a significant treatment effect favoring the control group detected. For a total of 269 simulations the analyses disagreed on the significance of the block*group interaction. As we expected an improvement, we have looked at one case each, where at least one analysis did not detect a treatment effect. These were selected by the maximal individual p-values and the maximal sum of p-values (Fig. 6). Similar to the examples presented before, these cases suggest that an improvement is not always observable, but also that an analysis without time can lead to a wrong interpretation of treatment effects. Effects might wrongly be attributed to a treatment if improvements within the pre- or posttest are large in the treatment group. On the other hand, they might wrongly be disregarded if improvements within the pre- or posttest are large in the control group.

3.2.3.2. Simulation of no treatment. For the simulations without a treatment, all analyses show comparable type 1 error rates (0.05-0.06)[3] and differences in the detection of effects for the simulation of a treatment are thus not only due to varying type 1 error rates. For almost all detections of treatment effects (105 out of 108), the analyses disagreed on the significance of the block*group interaction. As before, we have looked at one case for both directions of the disagreement (selected by minimal p-values, Fig. 7). While any detection of a treatment is a type 1 error, these cases suggest a wrong detection of treatment effects due to practice effects for the traditional analysis. With the incorporation of time, treatment effects are detected because of actual fluctuations of performance between blocks.

3.3. Other applications

First, the approach is easily applicable to analyses of improvements in relative performance, improvements with the number of tasks finished instead of time spent on tasks, and, by the use of generalized mixed models, also for categorical response data. Changing the analysis to only number of tasks might not be advisable at this point as we have seen that slower participants improve equally much or more while handling fewer stimuli.

The demonstrated approach is of course also applicable to nonlinear changes over time within sessions, which can be approximated using polynomials or generalized additive models (Winter and Wieling, 2016). However, these cases might not be the most interesting for investigation of treatments in pre-post designs as treatment effects are less clear in meaning if there are already complex practice effects within sessions. Nevertheless, the demonstrated approach can be applied and moreover be adapted to compare different sessions at different time points. While in the case of linear improvements we set time 0 to the end of the pretest and the beginning of the posttest, for quadratic performance changes over time, time 0 might be more appropriately set closer to the peak performance within sessions.

Another application of the demonstrated approach could be piecewise linear approximation of changes over time within sessions. Compared with growth curve analyses and generalized additive models, this would be advantageous if effects are expected to be limited in time. For example, improvements due to task familiarization might only affect the first few minutes of a test and fatigue might only start to decrease performance after some time has passed. Piecewise linear approximation allows these effects to be restricted in time. An additional advantage is that coefficients are easily interpretable. However, there are further advantages and disadvantages of local versus global procedures, similar to considerations of piecewise interpolation or segmented regression compared with their global counterparts.

Table 1
Comparison of the significance of the block*group interaction in 1000 random simulations using the treatment and 1000 random simulations without a treatment for the analyses accounting for time (rows) and not using time (columns).

Time\no time	Treatment		No treatment	
	$p < .05$	$p > .05$	$p < .05$	$p > .05$
$p < .05$	641	131	3	51
$p > .05$	138	90	54	895

4. Discussion

The use of linear mixed models provides the opportunity to include practice effects within sessions in the analysis. We presented an approach, which through inclusion and manipulation of time within blocks can account for these practice effects instead of misinterpreting them as treatment effects. The usefulness of such an approach is demonstrated through examples. Through further random samples drawn from real data, it is demonstrated that such cases are not unrealistic.

With the analysis of performance within one group, we have shown that time can explain performance differences between repeated tests but also demonstrated the possibility to find treatment effects. By comparison with control groups, treatment effects can also be identified without an analysis of time if all groups show comparable performance. If groups differ in performance or, in the using traditional methods undetectable worst case, in learning speed, analyses without time might fail or wrongly attribute treatment effects. In these cases, the inclusion of time in the analysis provides significant improvements for the detection and interpretation of effects. While the examples were constructed from extreme values, their average performances also include cases which are conventionally interpreted as signs of (un-)successful treatments, such as differing or comparable pretest performance and comparable posttest performance between groups. At least in the examples, these conclusions would be wrong.

Overall, the inclusion of time in the analysis has shown to improve the analysis and detection of learning and treatment effects. The improved analysis of course comes with the cost of higher complexity. Using random simulations of group allocations, we have shown that differences in learning speed within sessions may pose a serious concern and influence regarding interpretation is not limited to constructed cases as over one quarter of the simulations was affected. This suggests that the increase in complexity is rewarding. We do note that in our sessions the treatment and testing sessions were comparable and the practice and treatment effects should be similar. In realistic experiments, the variance and magnitude of practice effects could be smaller compared with treatment effects. However, this is rarely investigated.

[3] This suggests that the choice of random slopes is overall acceptable. However, we do note that closer inspection of the distribution of p-values indicates a possibly higher type 1 error rates for the time analysis. We also note large fluctuations in power and type 1 error rates for different choices of random slopes. While this does not contradict the points discussed here, it supports the importance of the choice of random slopes.

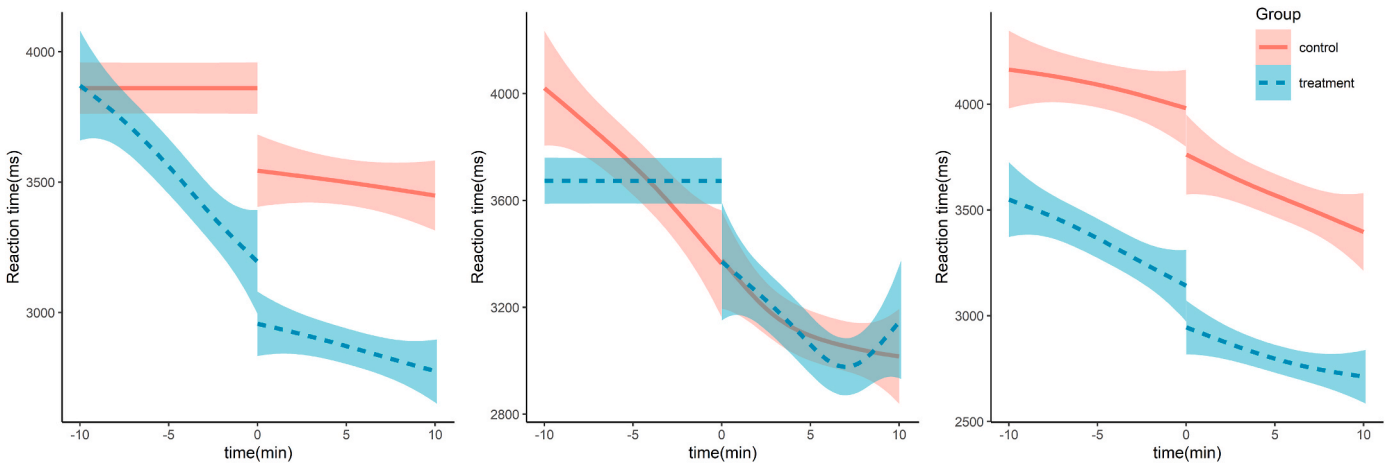

Fig. 6. Changes of reaction time within sessions for the random group allocation using a treatment, where at least one analysis does not show a significant treatment effect. Pre- and posttest are separated at time 0. Left: Analysis with time does not show a significant treatment effect. Center: Analysis without time does not show a significant treatment. Right: Neither analysis shows a significant treatment.

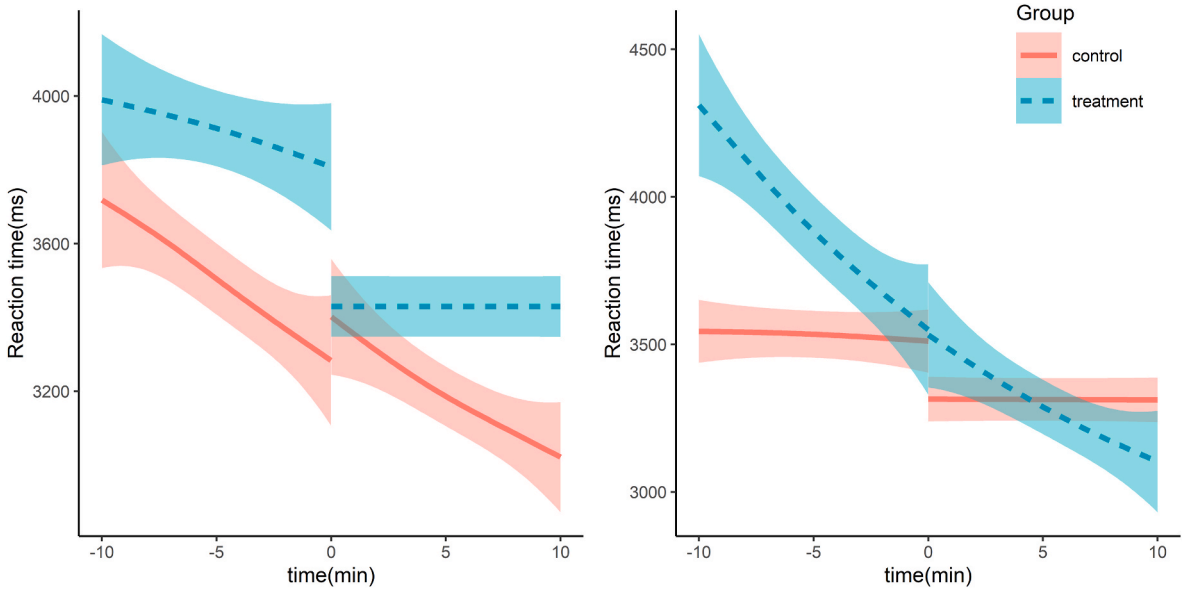

Fig. 7. Smoothed conditional means of reaction time for the random group allocation using no treatment, where at least one analysis shows a significant treatment effect. Blocks are separated at time 0. Left: Analysis with time shows a significant treatment effect. Right: Analysis without time shows a significant treatment.

While the suggested approach approximates practice effects, this is limited to effects within sessions and thus does not lessen the need for control groups, which can also account for other improvements between sessions. However, there is no consensus on which control activities optimally control for changes between sessions (Au et al., 2020). As such, the proposed analysis can supplement the control of placebo effects, especially if the comparability between control and treatment groups is hindered by performance or learning differences between groups. An advantage compared with the use of control groups is that no additional lab time or participants are required. This could be applicable, for example, in pilot studies or investigations of details of known treatments, where resources could better be invested in increasing power.

5. Limitations

While it allowed the construction and demonstration of multiple different cases, one major limitation is the simulation of groups and blocks from a dataset that did not originally include these. As all datasets were created from data of a single testing session there were no breaks between blocks, which might aid the linear continuation over time. This might transfer well to short experimental sessions where pre- and posttests are conducted within one hour. For multiple sessions spread out over days or years, the results might not be as applicable. But in these cases, the analysis of learning effects during single sessions and the approximation of performance toward the end of sessions might be of interest.

The demonstrated approach is also less useful for tasks, which produce only small practice effects within tests but larger effects between tests. For example, compared with quick adaptations in cognitive or coordinative tasks, strenuous physical exercise relies on rather slow morphological adaptation. While the investigation of changes within sessions might nevertheless be interesting, the incorporation into statistical analysis can only marginally control practice effects in such a case.

6. Conclusion and outlook

The use of linear mixed models provides the opportunity to include practice effects within sessions in the analysis. Through constructed examples and random simulations drawn from real data, we have shown the beneficial effects of the inclusion and manipulation of time in the statistical analysis with linear mixed models. The statistical model is improved, allows a more accurate analysis of learning effects, and a better detection of treatment effects. The proposed analysis can supplement comparisons with control groups and aid interpretation of treatment effects especially if such comparisons are hindered by pretest differences between groups.

Whereas we and many experiments focus on improvements of performance due to treatments, the analysis of learning within sessions also allows further interpretation. For example, an increased learning speed in the posttest could indicate a useful interaction of the treatment and the repetition of the task. On the other hand, a reduced learning speed in the posttest could indicate ceiling effects as a possible reason for null effects of treatments.

Author contributions

LJ: Conceptualization, methodology, software, analysis, writing. **PJ**: Conceptualization, supervision, writing.

Declarations

Open Practices Statement. No additional experimental data was collected for this study. All program code is available at https://github.com/LeonardoJost/TimeAnalysis. The study was not preregistered.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567 (7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9.

Au, J., Gibson, B.C., Bunarjo, K., Buschkuehl, M., Jaeggi, S.M., 2020. Quantifying the difference between active and passive control groups in cognitive interventions using two meta-analytical approaches. Journal of Cognitive Enhancement 4 (2), 192–210. https://doi.org/10.1007/s41465-020-00164-6.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59 (4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015a. Parsimonious Mixed Models. *ArXiv: 1506.04967*. http://arxiv.org/abs/1506.04967.

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2015b. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1). https://doi.org/10.18637/jss.v067.i01.

Brauer, M., Curtin, J.J., 2018. Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychol. Methods 23 (3), 389–411. https://doi.org/10.1037/met0000159.

Calamia, M., Markon, K., Tranel, D., 2012. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin. Neuropsychol. 26 (4), 543–570. https://doi.org/10.1080/13854046.2012.680913.

Goldberg, T.E., Harvey, P.D., Wesnes, K.A., Snyder, P.J., Schneider, L.S., 2015. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimer's Dementia: Diagnosis, Assessment and Disease Monitoring 1 (1), 103–111. https://doi.org/10.1016/j.dadm.2014.11.003.

Jost, L., Jansen, P., 2020. A novel approach to analyzing all trials in chronometric mental rotation and description of a flexible extended library of stimuli. Spatial Cognition & Computation 20 (3), 234–256. https://doi.org/10.1080/13875868.2020.1754833.

Levy, R., 2014. Using R Formulae to Test for Main Effects in the Presence of Higher-Order Interactions. *ArXiv:1405.2094*. http://arxiv.org/abs/1405.2094.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D., 2017. Balancing Type I error and power in linear mixed models. J. Mem. Lang. 94, 305–315. https://doi.org/10.1016/j.jml.2017.01.001.

Mirman, D., Dixon, J.A., Magnuson, J.S., 2008. Statistical and computational models of the visual world paradigm: growth curves and individual differences. J. Mem. Lang. 59 (4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006.

Nash, J.C., Varadhan, R., 2011. Unifying optimization algorithms to aid software system users: optimx for R. J. Stat. Software 43 (9), 1–14.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Shepard, R.N., Metzler, J., 1971. Mental rotation of three-dimensional objects. Science (New York, N.Y.) 171. https://doi.org/10.1126/science.171.3972.701 (FEBRUARY), 701–703.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer.

Winter, B., 2013. Linear Models and Linear Mixed Effects Models in R with Linguistic Applications. *ArXiv:1308.5499*. http://arxiv.org/abs/1308.5499.

Winter, B., Wieling, M., 2016. How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. Journal of Language Evolution 1 (1), 7–18. https://doi.org/10.1093/jole/lzv003.