Standard errors and confidence intervals in within-subject designs: Generalizing Loftus & Masson (1994) and avoiding biases of alternative accounts

Volker H. Franz (Universität Hamburg, Hamburg, Germany)

Geoffrey R. Loftus (University of Washington, Seattle, Washington),

Psychonomic Bulletin & Review (in press)

RUNNING HEAD: Confidence intervals in within–subject designs

Correspondence should be addressed to:

Prof. Dr. Volker Franz

Universität Hamburg

von Melle Park 5

20146 Hamburg

Phone: ++0049 (0)40 42838 2551

Fax: ++0049 (0)40 42838 5492

Email: volker.franz@uni-hamburg.de

Abstract

Repeated-measures designs are common in experimental psychology. Because of the correlational structure in these designs, calculation and interpretation of confidence intervals is nontrivial. One solution was provided by Loftus and Masson (1994). This solution, although widely adopted, has the limitation of implying the same-size confidence intervals for all factor levels and therefore does not allow assessment of variance homogeneity assumptions (i.e., the circularity assumption, which is crucial for the repeated measures ANOVA). This limitation and the method's perceived complexity has sometimes led scientists to use a simplified variant, based on a per–subject normalization of the data (Morrison & Weaver, 1995; Bakeman & McArthur, 1996; Cousineau, 2005; Morey, 2008). We show that this normalization method leads to biased results and is uninformative with regard to circularity. Instead, we provide a simple, intuitive generalization of the Loftus and Masson method that allows assessment of the circularity assumption.

Confidence intervals are important tools for data analysis. In Psychology, confidence intervals are of two main sorts. In between–subjects designs, each subject is measured in only one condition such that measurements across conditions are typically independent. In within–subjects (repeated–measures) designs, each subject is measured in multiple conditions. This has the advantage of reducing variability caused by differences among subjects. However, the correlational structures in the data cause difficulties in specifying confidence-interval size.

Figure 1a shows hypothetical data from Loftus and Masson (1994). Each curve depicts performance of one subject in three exposure-duration conditions. Most subjects show a consistent pattern — better performance with longer exposure duration — which is reflected by a significant effect in repeated-measures ANOVA ($F(2,18)=43$, $p<.001$).

Figure 1 here

However, this within–subjects effect is not reflected by traditional standard errors of the mean (*SEM*; Figure 1b) as calculated with the formula

$$SEM_j^{betw} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (y_{ij} - \overline{y_{.j}})^2}$$

where SEM_j^{betw} is the *SEM* in condition *j*, *n* the number of subjects, y_{ij} the dependent variable (*DV*) for subject *i* in condition *j*, and $\overline{y_{.j}}$ the mean *DV* across subjects in condition *j*.

The discrepancy occurs because the SEM^{betw} include both the subject-by-condition interaction variance — the denominator of ANOVA's F-ratio — and in addition the between–subjects variance that is irrelevant in the F-ratio. In our example, subjects show highly variable overall performances which hides the consistent pattern of within–subject effects. This is common: the between–subjects variability is typically larger than the subject-by-condition interaction variability. Therefore, the SEM^{betw} is inappropriate for assessing within–subject effects. Before discussing solutions to this shortcoming, we offer some general

comments about error bars.

Error bars

Error bars reflect measurement uncertainty and can have different meanings. For example, they can correspond to *SEM*, standard deviations, confidence intervals, or the recently proposed inferential confidence intervals (Goldstein & Healy 1995; Tryon 2001). Each of these statistics stresses one aspect of the data and each has its virtues. For example, standard deviations might be the first choice in a clinical context where the focus is on a single subject's performance. In experimental psychology the most used statistic is the *SEM*. For simplicity, we will therefore focus on the *SEM*, although all our results can be expressed in terms of any related statistic.

To better understand the *SEM*, it is helpful to recapitulate two simple "rules of eye" for the interpretation of *SEM*s. The rules, which we will call the 2- and 3-*SEM* rules, respectively, are equivalent to Cumming and Finch's (2005) rules 6 and 7. First, if a single mean (based on $n \geq 10$ measurements) is further from a theoretical value (typically zero) than ~2 *SEM*, then this mean is significantly different (at $\alpha = .05$), from the theoretical value. Second, if two means (both based on $n \geq 10$ measurements) in a between-subjects design with approximately equal *SEM*s are further apart than ~3 *SEM*, then these means are significantly different from one another (at $\alpha = .05$)[1].

Loftus & Masson method

Loftus and Masson (1994) offered a solution to the problem that SEM^{betw} hide within–subject effects (Figure 1c). The $SEM^{L\&M}$ are based on the pooled error term of repeated-measures ANOVA and constructed such that the 3-*SEM* rule can be applied when interpreting

[1] For simplicity, the 3-*SEM* rule treats all comparisons as *a-priori* contrasts and does not take into account problems of multiple testing. Below we provide an example of Bonferroni correction for post-hoc testing. Similarly, one could calculate confidence intervals based on Tukey's range test or similar statistics.

differences between means. This central feature makes the $SEM^{L\&M}$ in a repeated-measures design behave analogously to the SEM^{betw} in a between–subjects design[2].

Normalization method

Although widely accepted, Loftus and Masson's (1994) method has two limitations: (a) by using the pooled error term, the method assumes *circularity* which is to a repeated-measures design what homogeneity of variance (HOV) is to a between-subjects design. Consequently, all $SEM^{L\&M}$ are of equal size. This is different from between–subjects designs where the relative sizes of the SEM^{betw} allow judgment of the HOV assumption. (b) the formulas by Loftus and Masson (1994) are sometimes perceived as unnecessarily complex (Bakeman & McArthur, 1996).

Therefore, Morrison & Weaver (1995), Bakeman and McArthur (1996), Cousineau (2005), and Morey (2008) suggested a simplified method which we call normalization method. It is based on an illustration of the relationship between within– and between–subjects variance used by Loftus and Masson (1994)[3]. Proponents of the normalization method argue that it is simple and allows judgement of the assumption of circularity.

The normalization method consists of two steps. First, the data are normalized (Figure 1d). That is, the overall performance levels for all subjects are equated without changing the pattern of within–subject effects. Normalized scores calculate as

$$w_{ij} = y_{ij} - \overline{(y_{i.} - y_{..})}$$

[2] Note that the $SEM^{L\&M}$ only provides information about the *differences* among within–subject levels. It does not provide information about the *absolute value* of the *DV*, for which SEM^{betw} would be appropriate. It is, however, rare in Psychology that absolute values are of interest.

[3] Unfortunately, this illustration led to some confusion. Although it provides a valid description of the error term in the repeated measures ANOVA, it suggests that the Loftus and Masson method was based on normalized scores, which is not true. Therefore, the normalization method is not a generalization of the Loftus and Masson method. Also the critique based on the assumption that the Loftus and Masson method used normalized scores (Blouin & Riopelle, 2005) does not apply.

where *i* and *j* index subject and factor level, w_{ij} and y_{ij} represent normalized and raw scores respectively, $\overline{y_{i.}}$ is the mean score for subject *i* (averaged across all conditions), and $\overline{y_{..}}$ is the grand mean of all scores. Second, the normalized scores w_{ij} are treated as if they were from a between–subjects design. The rationale is that the irrelevant between–subjects differences are removed such that now standard computations and the traditional *SEM* formula can be used on the normalized scores

$$SEM_j^{norm} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (w_{ij} - \overline{w_{.j}})^2}$$

with SEM_j^{norm} being the SEM^{norm} in condition *j* and *n* the number of subjects. These SEM^{norm} are shown in Figure 1e.

The normalization method seems appealing in its simplicity. All that is required is normalizing the within–subjects data and then standard methods from between–subjects designs can be used. However, the method underestimates the *SEM* and does not allow assessment of circularity.

Problem 1 of the normalization method: *SEM* are too small

Figures 1c and 1e illustrate this problem: all SEM^{norm} are smaller than the $SEM^{L\&M}$. This is a systematic bias that occurs because the normalized data, although correlated, are treated as uncorrelated. Consequently, the pooled SEM^{norm} underestimates the $SEM^{L\&M}$ by a factor of $\sqrt{\frac{J-1}{J}}$ (J being the number of factor levels)[4]. Morey (2008) derived this relationship and also suggested that the SEM^{norm} be corrected. However, this is not a complete solution because the

[4] That the normalization method is biased might confuse some readers because they remember that we can represent a within–subjects ANOVA as a between–subjects ANOVA on the normalized scores (Maxwell & Delaney, 2000, p. 472, footnote 5 of chapter 11). However, to obtain a correct F–test we would need to deviate from the between–subjects ANOVA by adjusting the degrees of freedom (Loftus & Loftus, 1988, digression 13–1, p. 426). This adjustment takes into account that the normalized data are correlated and is not performed by the normalization method.

method also leads to an erroneous view of what circularity means.

Circularity

Between-subjects ANOVA assumes HOV and we can assess its plausibility by judging whether the SEM^{betw} are of similar size. The corresponding assumption for repeated–measures ANOVA is *circularity* (Huynh & Feldt, 1970, Rouanet & Lepine, 1970).

Consider the variance–covariance matrix Σ of a repeated-measures design. Circularity is fulfilled if and only if an orthonormal matrix M exists that transforms Σ into a spherical matrix (i.e., with λ on the main diagonal and zero elsewhere), such that

$$M\Sigma M' = \lambda I$$

where λ is a scalar, and I is the identity matrix (cf. Winer, Brown, & Michaels, 1991). Because of this relationship to sphericity, the circularity assumption is sometimes called the sphericity assumption.

We can reformulate circularity in a simple way: circularity is fulfilled if and only if the variability of all pairwise differences between factor levels is constant (Huynh & Feldt, 1970, Rouanet & Lepine, 1970). Therefore we can assess circularity by examining the variance of the differences between any two factor levels. Depicting the corresponding *SEM*, which we describe below, is an easy generalization of the Loftus and Masson method. Before describing this method, we show that the normalization method fails to provide correct information about circularity.

Problem 2 of the normalization method: erroneous evaluation of circularity

There are different reasons why the normalization method cannot provide a visual assessment of circularity. For example, testing for circularity requires evaluating the variability of all $J(J-1)/2$ pairwise differences (J being the number of factor levels), while the normalization method yields only J SEM^{norm} to compare. Also, we can construct examples

showing clear violations of circularity that are not revealed by the normalization method.

Figure 2 shows such an example for one within–subjects factor with 4 levels. The pairwise differences (Figure 2d) show small variability between levels A,B and levels C,D, but large variability between levels B,C. The normalization method does not indicate this large circularity violation (Figure 2c). The reason can be seen in Figure 2b: normalization propagates the large B,C variability to conditions A and D. Because conditions A and B don't add much variability themselves, the normalization method creates the wrong impression that circularity holds.

Figure 2 here

It is instructive to evaluate this example using standard measures of circularity. The Greenhouse–Geisser epsilon (Box, 1954b, 1954a, Greenhouse & Geisser, 1959) attains its lowest value at maximal violation (here $\varepsilon_{min} = 1/(J-1) = 0.33$) while a value of $\varepsilon_{max} = 1$ indicates perfect circularity. In our example, $\varepsilon = 0.34$, showing the strong violation of circularity (Huynh & Feldt's, 1976, epsilon leads to the same value). The Mauchly (1940) test also indicates a significant violation of circularity ($W = 0.0001$; $p<.001$) and a repeated–measures ANOVA yields a significant effect ($F(3, 57) = 3$, $p= .036$) but only if we — erroneously — assume circularity. If we recognize this violation of circularity and perform the Greenhouse–Geisser or Huynh–Feldt corrections, then the effect is not significant (both p's = 0.1). A multivariate ANOVA (MANOVA), also leads to a nonsignificant effect ($F(3,17)=1.89$, $p=.17$). In summary, our example shows that the normalization method can hide serious circularity violations. A plot of the *SEM* of the pairwise differences, on the other hand, clearly indicates the violation.

A better approach: picturing pairwise differences

As a simple and mathematically correct alternative to the normalization method, we

suggest to show all pairwise differences between factor levels with corresponding *SEM* ($SEM^{pairedDiff}$), as shown in Figures 1g and 2d. To the degree that these $SEM^{pairedDiff}$ are variable there is evidence for violation of circularity. Figure 1g shows that for the Loftus and Masson (1994) data, all $SEM^{pairedDiff}$ are similar, suggesting no serious circularity violation (which is consistent with standard indices; Greenhouse–Geisser ε = 0.845; Huynh–Feldt ε = 1; Mauchly test $W = 0.817$; $p = .45$).

The $SEM^{pairedDiff}$ are easy to compute because only the traditional formulas for the *SEM* of the differences are needed. Consider the levels *k* and *l* of a repeated measure factor. We first calculate the pairwise differences for each subject $d_i = y_{ik} - y_{il}$, then use the traditional formula to calculate the *SEM* of the mean difference

$$SEM_{kl}^{pairedDiff} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(d_i - \overline{d}_{\cdot})^2}$$

This approach is consistent with the Loftus and Masson method because pooling the $SEM^{pairedDiff}$ results in $\frac{1}{\sqrt{2}}SEM^{L\&M}$ (Appendix 1). Therefore, we can use this relationship to calculate the $SEM^{L\&M}$ without the inconvenience of extracting the relevant ANOVA error term from the output of a statistical program (another critique of the Loftus & Masson method; Cousineau, 2005; Morey, 2008)

Picturing pairwise differences can supplement numeric methods

Figure 3 illustrates how evaluating $SEM^{pairedDiff}$ can lead to a surprising result, thereby showing the virtues of our approach. Repeated measures ANOVA shows for these data a clearly non-significant result, whether we correct for circularity violation or not ($F(3,117) = 1.2$; p=.32, Greenhouse–Geisser ε = .50; p = .30; Huynh–Feldt ε = .51; p = .30). We show that our method nevertheless detects a strong, significant effect and will guide the researcher to the in this case more appropriate multivariate methods.

Figure 3 here

Inspecting Figure 3c for circularity violations shows that between conditions D and C there is a very small $SEM^{pairedDiff}$, indicating that the pairwise difference between these conditions has much less variability than all the other pairwise differences. Applying the 2-*SEM* rule indicates that the corresponding difference differs significantly from zero, while no other differences are significant. This is also true, using the Bonferroni correction[5] for multiple testing, as suggested by Maxwell and Delaney (2000).

In short, $SEM^{pairedDiff}$ indicate that there is a strong circularity violation and a strong effect. Univariate repeated measures ANOVA does not detect this effect even when correcting for circularity violations. MANOVA on the other hand, detects the effect ($F(3,37)=98$, $p<.001$) and is thereby consistent with the result of our approach[6].

This example shows that the $SEM^{pairedDiff}$ convey important information about the correlational structure of the data that can prompt the researcher to use more appropriate methods. No other method discussed in this article would have achieved this.

Practical considerations when picturing pairwise differences

The above example shows that our approach can help the researcher during data analysis. When presenting data to a general readership, a more compact way of presenting the $SEM^{pairedDiff}$ might be needed, especially for factors with many levels (because the number of pairwise differences can become large; J factor levels result in $J(J-1)/2$ pairwise differences). If a plot of pairwise differences would be overly tedious one could (a) present the data as an upper triangular matrix either in numerical form or as a color-coded heat-map;

---

[5] The Bonferroni correction is this: We have 6 possible comparisons. Therefore, we need the (100-5/6)%=99.12% criterion of the t-distribution with (40-1)=39 degrees of freedom; which is: $t_{crit} = 2.78$. Therefore, all *SEM* need to be multiplied by this value (instead of 2 as in the 2-SEM rule).

[6] In our example, MANOVA is more appropriate because it does not rely on the assumption of circularity. It has, however, other limitations (mainly for small sample sizes) such that it cannot simply replace univariate ANOVA in general.

(b) present the $SEM^{pairedDiff}$ together with the $SEM^{L\&M}$ in one single plot, as shown in Figure 1f. In this plot, the error bars with short crossbar correspond to the $SEM^{pairedDiff}$ (scaled, see below) and the error bars with long crossbar correspond to the $SEM^{L\&M}$. The plot gives a correct impression of circularity by means of the scaled $SEM^{pairedDiff}$ (if circularity holds, all scaled $SEM^{pairedDiff}$ will be similar to $SEM^{L\&M}$) and allows application of the 3-SEM rule to interpret differences between means. The downside is, that it is not immediately apparent which error bars belong to which pair of means. The researcher needs to decide whether compactness of presentation outweighs this limitation.

To create a plot like Figure 1f, each $SEM^{pairedDiff}$ is multiplied by $\frac{1}{\sqrt{2}}$ and then plotted as error bar for each of the two means from which the difference was calculated. The scaling is necessary because we go back from a difference of two means to two single means. The scaling gives us, for each mean, the *SEM* that would correspond to the *SEM* of the difference if the two means were independent and had the same variability, such that the 3-*SEM* rule can be applied and the scaled $SEM^{pairedDiff}$ are compatible with the $SEM^{L\&M}$ (Appendix 1).

Generalization to multi–factor experiments.

(a) Only within-subjects factors

So far, we discussed only single–factor designs. If there is more than one repeated–measures factor, the $SEM^{pairedDiff}$ should be calculated across all possible pairwise differences. This is a simple method that is consistent with the Loftus and Masson method, which also reduces multiple–factors to a single–factor (e.g., a 3 x 5 design is treated as a single-factor design with 15 levels).

With regard to circularity, our generalization is slightly stricter than necessary because we consider the pairwise differences of the variance–covariance matrix for the full comparison

(by treating the design as a single–factor design). If the variance–covariance matrix fulfills circularity for this comparison, then it also fulfills it for all sub-comparisons, but not vice–versa (Rouanet & Lepine, 1970, corollary 2). Therefore, it is conceivable that the $SEM^{pairedDiff}$ indicate a violation of circularity, but that a specific sub-comparison corresponding to one of the repeated–measures factors does not. However, we think that the simplicity of our rule outweighs this minor limitation.

(b) Mixed designs (within- and between-subjects factors)

In mixed designs, an additional complication arises because each group of subjects (i.e., each level of the between subjects factors) has its own variance-covariance matrix, all of which are assumed to be homogeneous and circular. Thus, there are two assumptions, HOV and circularity. As mentioned by Winer et al. (1991, p. 509), "these are, indeed, restrictive assumptions"; hence even more need for a visual guide to evaluate their plausibility.

Consider one within-subjects factor and one between-subjects factor, fully crossed, with equal group sizes. For each level of the between-subjects factor we suggest a plot with the means and SEM^{betw} for all levels of the within-subjects factor, along with a plot showing the pairwise differences and their $SEM^{pairedDiff}$ (Figure 4 and Appendix A2). To evaluate the homogeneity and circularity assumptions, respectively, one would gauge whether all SEM^{betw} corresponding to the same level of the within-subjects factor were roughly equal and whether all possible $SEM^{pairedDiff}$ were roughly equal.

Figure 4 here

Inspecting Figure 4a shows that group 2 has higher SEM^{betw} than the other groups, suggesting a violation of the HOV assumption. And indeed, the four corresponding Levene (1960) tests, each comparing the variability of the groups at one level of the within-subjects factors, show significant deviation from HOV (all $F>27$, all $p<.001$). Our approach reveals

that this is due to the higher variability of group 2. Inspecting Figure 4b shows that the $SEM^{pairedDiff}$ are similar, suggesting that circularity is fullfilled. This, again, is consistent with standard repeated–measure methods (Greenhouse–Geisser $\varepsilon = 0.960$; Huynh–Feldt $\varepsilon = 1$; Mauchly test $W = 0.944$; $p = .25$).

Precautions

Although we believe our approach to be beneficial, it needs to be applied with caution (as any statistical procedure). Strictly speaking, the method only allows judgements about pairwise differences and the circularity assumption. It does not allow judgments of main effects or interactions. For this we would need pooled error terms and overall averaging, as used in ANOVA. Also, our use of multiple estimates of variability (i.e., for each pairwise difference a different $SEM^{pairedDiff}$) makes each individual $SEM^{pairedDiff}$ less reliable than an estimate based on the pooled error term. In many situations, however, neither restriction is a serious limitation.

For example, consider Figure 1g. The $SEM^{pairedDiff}$ are consistent, such that the *SEM* based on the pooled error term will be similar (Appedix A1) and that the inherently reduced reliability of the $SEM^{pairedDiff}$ is no problem. Each pairwise difference suggests a significant difference from zero, be it interpreted as a-priori test, as post-hoc test,[7] or by applying the 2-*SEM* rule of eye. Therefore, a reader seeing only this figure will have an indication that the main effect of the ANOVA is significant. This example again shows how our method can supplement (though not supplant) traditional, numerical methods.

[7] As an example, let us calculate the CI for the difference "2s-1s": (a) A-priori test: The 95% critical value of the t-distribution is: $t_{crit\ 95\%}(9)=2.26$, resulting in a CI of 2 +/- (0.33*2.26) = [1.25,2.75]. (b) Post-hoc test with Bonferroni correction: With J=3 pairwise comparisons, we need the (100-5/3)=98.33% criterion of the t-distribution, which is $t_{crit\ 98.33\%}(9)=2.93$ and the CI calculates as 2 +/- (0.33*2.93) = [1.03,2.97].

Conclusions

We suggest a simple method to conceptualize variability in repeated–measures designs: calculate the $SEM^{pairedDiff}$ of all pairwise differences and plot them. The homogeneity of the $SEM^{pairedDiff}$ provides an assessment of circularity and is (other than the normalization method) a valid generalization of the well–established Loftus and Masson (1994) method.

References

Bakeman, R., & McArthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison, and others. *Behavior Research Methods, Instruments, & Computers*, *28*, 584–589.

Blouin, D. C., & Riopelle, A. J. (2005). On confidence intervals for within–subjects designs. *Psychological Methods*, *10*, 397–412.

Box, G. E. P. (1954a). Some theorems on quadratic form applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two–way classification. *Annals of Mathematical Statistics*, *25*, 484–498.

Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. effect of inequality of variance in the one–way classification. *Annals of Mathematical Statistics*, *25*, 290–302.

Cousineau, D. (2005). Confidence intervals in within–subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–46.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 581*, 175-177.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split–plot designs. *Journal of Educational Statistics*, *1*, 69–82.

Huynh, L., & Feldt, S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F–distributions. *Journal of the American Statistical Association*, *65*, 1582–1589.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278-292). Palo Alto, CA: Stanford University Press.

Loftus, G. R., & Loftus, E. F. (1988). *Essence of Statistics* (2 ed.). New York: McGraw–Hill.

Loftus, G. R., & Masson, E. J. M. (1994). Using confidence intervals in within–subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics, 11*, 204-209.

Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective.* Lawrence Erlbaum: New Jersey.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61–64.

Morrison, G. R., & Weaver, B. (1995). Exactly how many p–values is a picture worth—a commentary on Loftus's plot–plus–error–bar approach. *Behavior Research Methods Instruments & Computers*, *27*, 52–56.

Rouanet, H., & Lepine, D. (1970). Comparison between treatments in a repeated–measurement design—ANOVA and multivariate methods. *British Journal of Mathematical & Statistical Psychology*, *23*, 147–163.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371-386.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3 ed.). McGraw–Hill: Boston.

Appendix

A1. Relationship between $SEM^{pairedDiff}$ and $SEM^{L\&M}$

We show the $SEM^{L\&M}$ is equal to the pooled and scaled $SEM^{pairedDiff}$ in the following way

$$SEM^{L\&M} = \sqrt{\overline{\left(\frac{1}{\sqrt{2}} SEM^{pairedDiff}{}_{\cdot\cdot}\right)^2}}$$

This notation is similar to Winer et al. (1991): The horizontal line and the two dots indicate that all corresponding $SEM^{L\&M}$ are pooled. For example, in Figure 1g the $SEM^{pairedDiff}$ are: 0.3333, 0.2906, 0.4163 such that

$$\begin{aligned} SEM^{L\&M} &= \sqrt{\frac{\left(\frac{1}{\sqrt{2}} SEM^{pairedDiff}{}_{12}\right)^2 + \left(\frac{1}{\sqrt{2}} SEM^{pairedDiff}{}_{13}\right)^2 + \left(\frac{1}{\sqrt{2}} SEM^{pairedDiff}{}_{23}\right)^2}{3}} \\ &= \sqrt{\frac{0.2357^2 + 0.2055^2 + 0.2944^2}{3}} = 0.2480 = SEM^{L\&M} \end{aligned}$$

For the proof, consider a factor with $J = 3$ levels first. For a single–factor, repeated measures ANOVA $MSE = \overline{var_{\cdot}} - \overline{cov_{\cdot\cdot}}$ (Winer et al., 1991; p. 264). Because $SEM^{L\&M} = \sqrt{\frac{MSE}{n}}$, we obtain

$$SEM^{L\&M2} = \frac{MSE}{n} = \frac{\overline{var_{\cdot}} - \overline{cov_{\cdot\cdot}}}{n} = \frac{var_1 + var_2 + var_3 - cov_{12} - cov_{13} - cov_{23}}{3\,n}$$

The SEM for the difference between levels k and l is $SEM_{kl}^{pairedDiff} = \sqrt{\frac{var_k - 2\,cov_{kl} + var_l}{n}}$. Multiplying by $\frac{1}{\sqrt{2}}$ and pooling gives

$$\begin{aligned} \overline{\left(\frac{1}{\sqrt{2}} SEM^{pairedDiff}{}_{\cdot\cdot}\right)^2} &= \frac{1}{3}\left(\frac{1}{2} SEM_{12}^{pairedDiff2} + \frac{1}{2} SEM_{13}^{pairedDiff2} + \frac{1}{2} SEM_{23}^{pairedDiff2}\right) \\ &= \frac{var_1 - 2\,cov_{12} + var_2 + var_1 - 2\,cov_{13} + var_3 + var_2 - 2\,cov_{23} + var_3}{6\,n} \\ &= \frac{var_1 + var_2 + var_3 - cov_{12} - cov_{13} - cov_{23}}{3\,n} = SEM^{L\&M2} \end{aligned}$$

Generalization to a factor with more than 3 levels: There are $\frac{J(J-1)}{2}$ pairwise differences,

$\frac{J(J-1)}{2}$ covariances, and J variances. This gives

$$\begin{aligned}
\overline{\left(\frac{1}{\sqrt{2}}SEM^{pairedDiff}{}_{\cdot\cdot}\right)^2} &= \frac{2}{J(J-1)}\sum_{k=1}^{J-1}\sum_{l=k+1}^{J}\frac{1}{2}SEM_{kl}^{pairedDiff\,2} \\
&= \frac{2}{J(J-1)}\frac{1}{2n}\sum_{k=1}^{J-1}\sum_{l=k+1}^{J} var_k - 2\,cov_{kl} + var_l \\
&= \frac{1}{n}\left(\frac{1}{J(J-1)}\sum_{k=1}^{J}(J-1)\,var_k - \frac{1}{J(J-1)}\sum_{k=1}^{J-1}\sum_{l=k+1}^{J} 2\,cov_{kl}\right) \\
&= \frac{1}{n}\left(\frac{1}{J}\sum_{k=1}^{J} var_k - \frac{2}{J(J-1)}\sum_{k=1}^{J-1}\sum_{l=k+1}^{J} cov_{kl}\right) \\
&= \frac{1}{n}\left(\overline{var_{\cdot}} - \overline{cov_{\cdot\cdot}}\right) = SEM^{L\&M\,2}
\end{aligned}$$

A2. Mixed designs

We treat all within– and between–subjects factors of a mixed design as single factors, such that we reduce the problem to one between– and one within–subjects factor. In such a two–factor, mixed design there is for each level of the between–subjects factor a different variance–covariance matrix for the within–subjects factor, which all have to be homogeneous and circular (Winer et al., 1991, p. 506). If group sizes are equal, this can be assessed in three steps: (a) estimate for each level of the within–subjects factor, whether the corresponding SEM^{betw} are equal across all levels of the between–subjects factor. If this is the case, then the entries on the diagonal of the variance–covariance matrices (i.e., the variances) are equal. (b) estimate for each pair of within–subject levels, whether the corresponding $SEM^{pairedDiff}$ are equal across all levels of the between–subjects factor. This ensures that all off–diagonal elements of the variance–covariance matrices (i.e., the covariances) are equal, because we already know that the variances are equal and due to the relationship $SEM_{kl}^{pairedDiff} = \sqrt{\frac{var_k - 2\,cov_{kl} + var_l}{n}}$ the $SEM_{kl}^{pairedDiff}$ can only be equal if the covariances are equal. (c) estimate for each level of the between–subjects factor whether the $SEM^{pairedDiff}$ corresponding to all pairs of within–subject levels are equal. This ensures circularity of the variance–covariance matrices.

In short, we need to assess whether all SEM^{betw} at each level of the within–subject factor are similar and whether all $SEM^{pairedDiff}$ are similar. With unequal group sizes, we cannot use SEM because different n would enter the calculation. Therefore, we need to use standard

deviations instead.

Acknowledgments

Supported by grants DFG-FR 2100/2,3,4-1 to VF and NIMH-MH41637 to GL. Calculations were performed in R (http://www.R-project.org).

Figure Legends

Figure 1: Hypothetical data of Loftus and Masson (1994). **a.** Individual data: Each subject performs a task under three exposure durations (1 s, 2 s, and 5 s). Although subjects vary in their overall performance, there is a clear within–subjects pattern: all subjects improve with longer exposure duration. **b.** The between–subjects SEM^{betw} don't reflect this within–subjects pattern because the large between–subjects variability hides the within–subjects variability. **c.** $SEM^{L\&M}$ as calculated by the Loftus and Masson method adequately reflect the within–subjects pattern. **d.** Normalization method: First, data are normalized **e.** Second, traditional SEM are calculated across the normalized values, resulting in SEM^{norm}. **f.** Our suggestion for a compact display of the data. Error bars with long crossbars correspond to $SEM^{L\&M}$ and error bars with short crossbars to $SEM^{pairedDiff}$ (scaled by the factor $\frac{1}{\sqrt{2}}$ see main text). The fact that the $SEM^{pairedDiff}$ are almost equal to the $SEM^{L\&M}$ indicate that there is no serious violation of circularity. **g.** Pairwise differences between all conditions and corresponding $SEM^{pairedDiff}$. Error bars depict ± 1 SEM, as calculated by the different methods. Numbers below the error bars are the numerical values of the SEM.

Figure 2: Example showing that the normalization method fails to detect serious violations of circularity. **a.** Simulated data for a within–subjects factor with four levels. **b.** Normalized data. **c.** The normalization method leads to similar SEM^{norm}, thereby not indicating the violation of circularity. **d.** Pairwise differences and corresponding $SEM^{pairedDiff}$ indicate a large violation of circularity. Error bars depict ± 1 SEM, as calculated by the different methods. Numbers below the error bars are the numerical values of the SEM.

Figure 3: Example demonstrating the virtues of our approach. **a.** Simulated data for a within–subjects factor with four levels. **b.** Means and corresponding SEM^{betw}. **c.** The pairwise differences and corresponding $SEM^{pairedDiff}$ indicate a large violation of circularity. Error bars depict ± 1 SEM, as calculated by the different methods. Numbers below the error bars are the numerical values of the SEM.

Figure 4: Generalization of our approach to mixed designs. The example has one

between–subjects factor with 3 levels (groups 1–3) and one within–subjects factor with 4 levels (conditions A–D) **a.** Means and corresponding SEM^{betw}. Group 2 has larger SEM^{betw} indicating a violation of the homogeneity assumption. **b.** Pairwise differences and corresponding $SEM^{pairedDiff}$ indicate no violation of circularity. Error bars depict ± 1 SEM, as calculated by the different methods. Numbers below the error bars are the numerical values of the SEM.

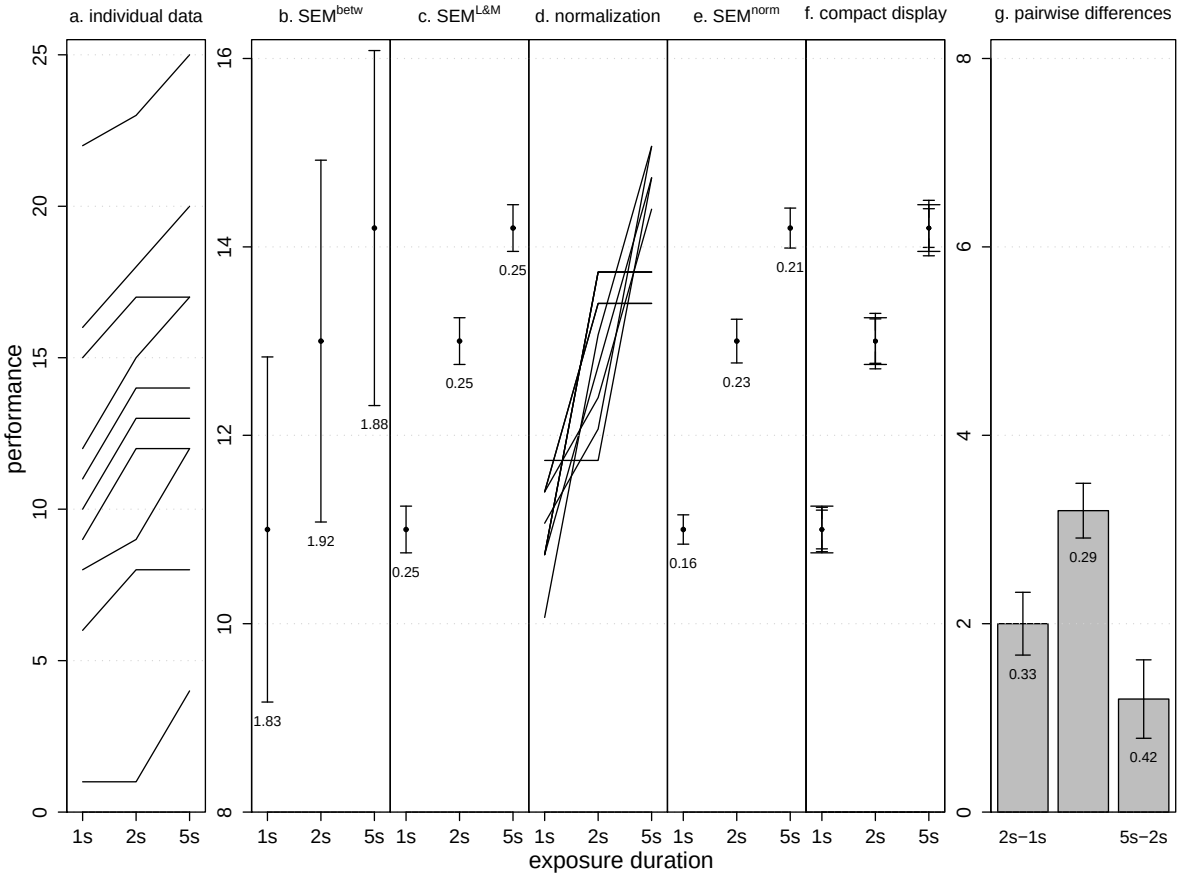
a. individual data
b. SEM^betw
c. SEM^L&M
d. normalization
e. SEM^norm
f. compact display
g. pairwise differences
performance
exposure duration
1s
2s
5s
1.83
1.92
1.88
0.25
0.25
0.25
0.16
0.23
0.21
0.33
0.29
0.42
2s−1s
5s−2s

Figure: 1

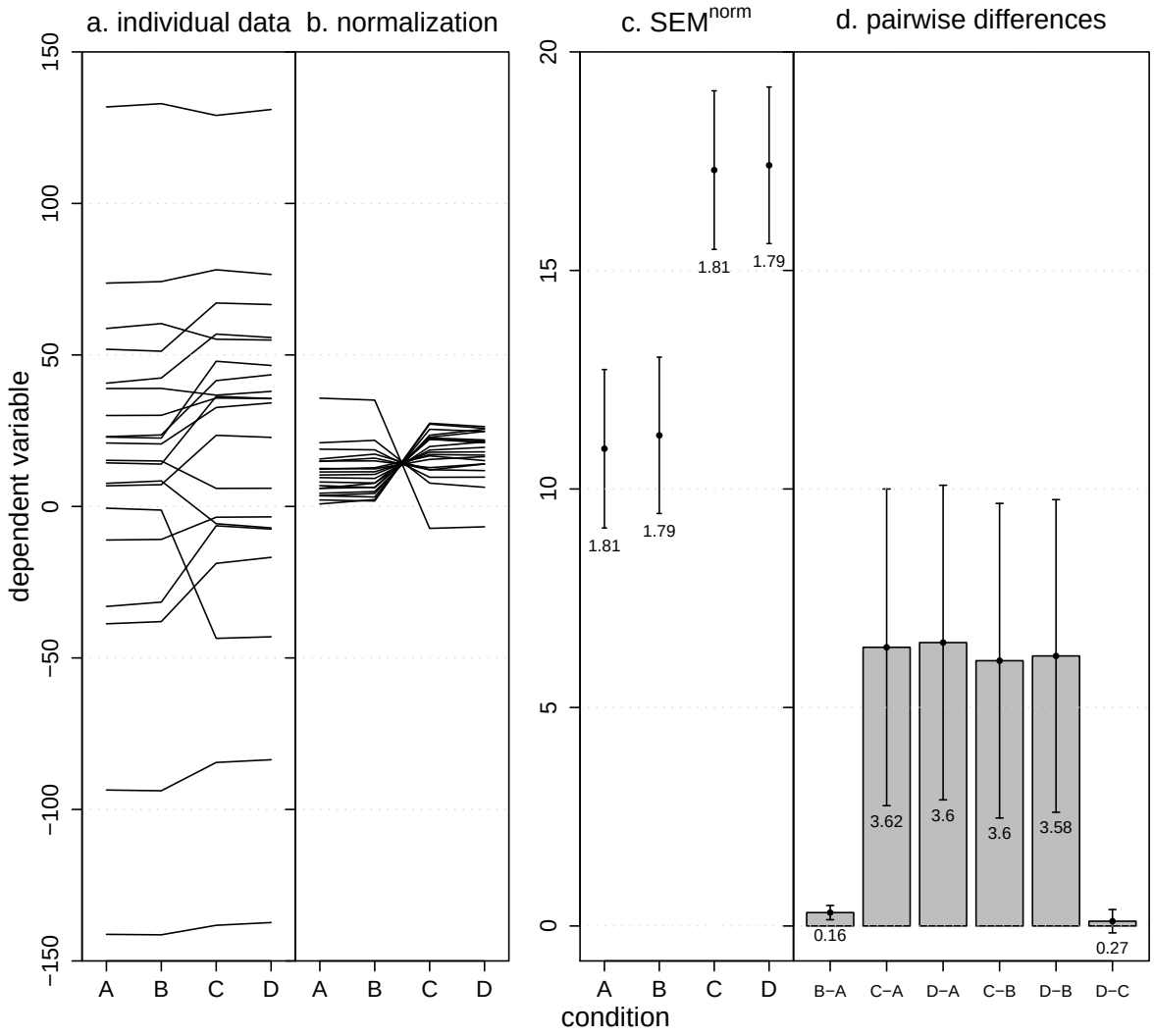
a. individual data
b. normalization
c. SEM^norm
d. pairwise differences
dependent variable
150
100
50
0
−50
−100
−150
A
B
C
D
20
15
10
5
0
1.81
1.79
1.81
1.79
3.62
3.6
3.6
3.58
0.16
0.27
B−A
C−A
D−A
C−B
D−B
D−C
condition

Figure: 2

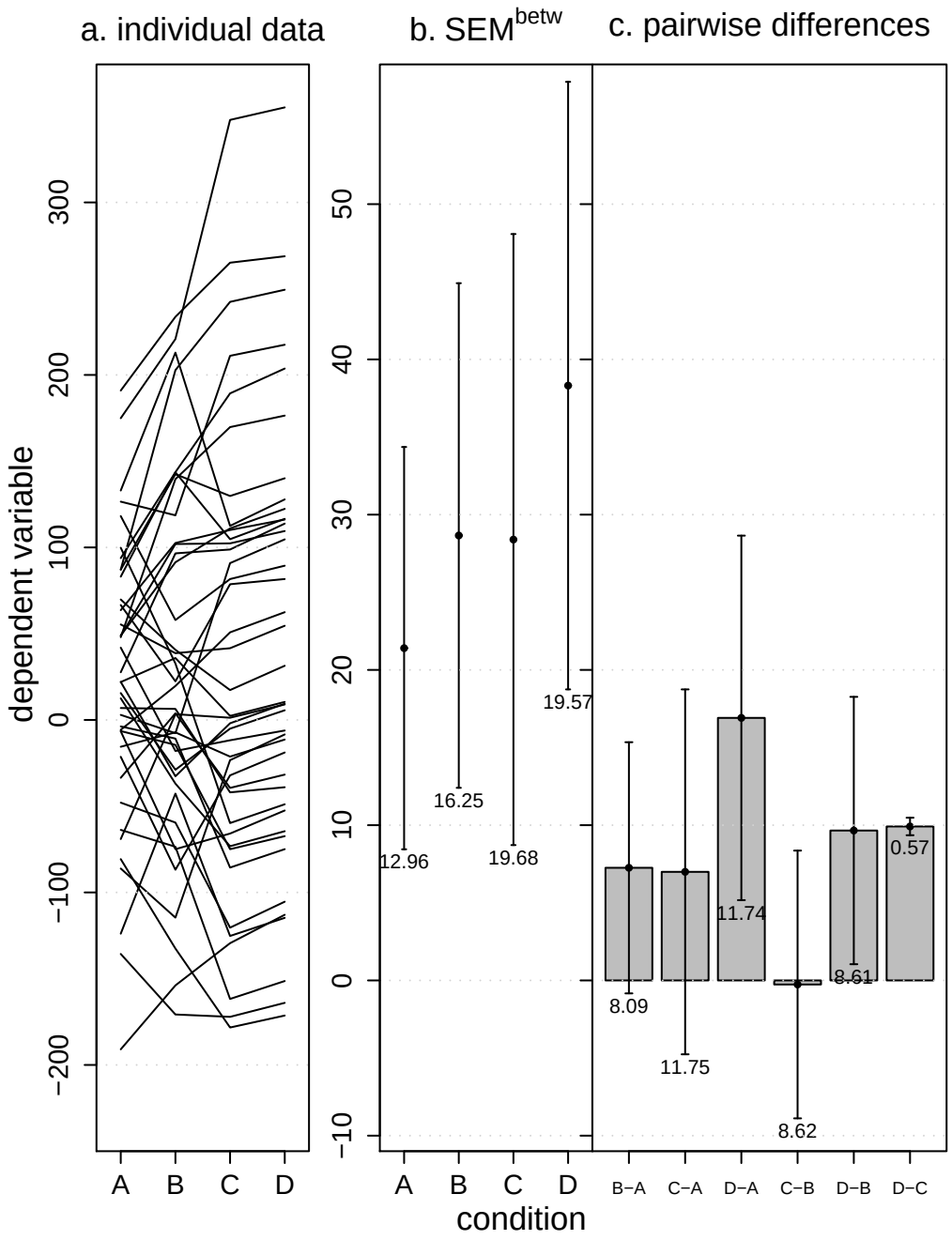
a. individual data
b. SEM^betw
c. pairwise differences
dependent variable
300
200
100
0
−100
−200
50
40
30
20
10
0
−10
A
B
C
D
condition
12.96
16.25
19.68
19.57
8.09
11.75
11.74
8.62
8.61
0.57
B−A
C−A
D−A
C−B
D−B
D−C

Figure: 3

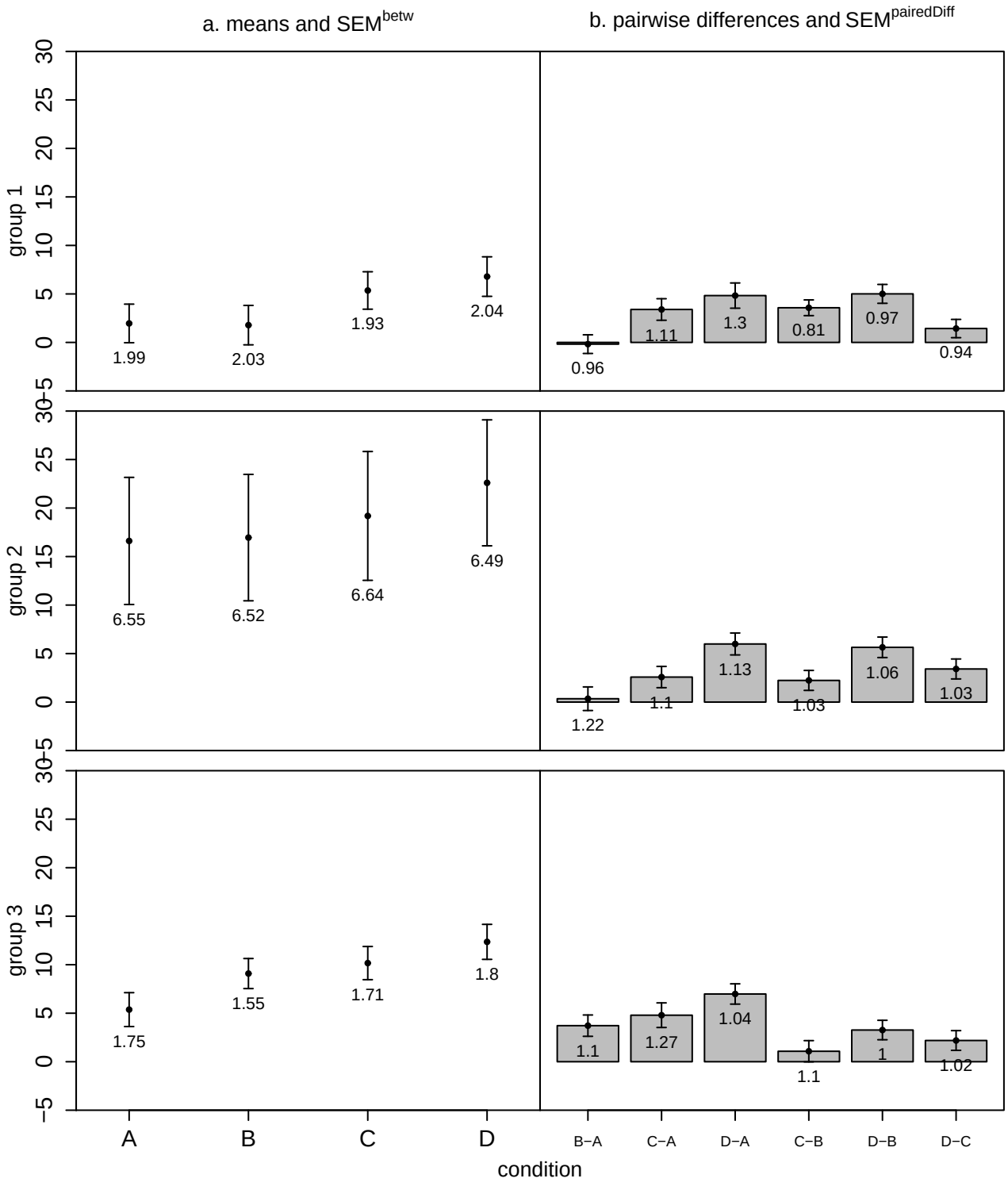
a. means and SEM betw
b. pairwise differences and SEM pairedDiff
group 1
group 2
group 3
1.99
2.03
1.93
2.04
6.55
6.52
6.64
6.49
1.75
1.55
1.71
1.8
0.96
1.11
1.3
0.81
0.97
0.94
1.22
1.1
1.13
1.03
1.06
1.03
1.1
1.27
1.04
1.1
1
1.02
A
B
C
D
B−A
C−A
D−A
C−B
D−B
D−C
condition

Figure: 4