Special issue article: Methods and statistics in social psychology: Refinements and new developments

Predicting variability: Using multilevel modelling to assess differences in variance

TOON KUPPENS¹* AND VINCENT Y. YZERBYT²
¹University of Groningen, The Netherlands; ²Université catholique de Louvain, Belgium

Abstract

Researchers are sometimes interested in the variability of data rather than in their absolute or relative values. An important example of such a situation in social psychology is consensus: the fact that people are more similar to each other in certain conditions. Currently, methods to assess differences in consensus (or variability in general) are not very well-developed or widely known. We describe an existing tool that allows testing the extent to which variability depends on one or several predictor variables. We explain how multilevel modelling’s capacity to model heterogeneity in residual variance can be used to test substantive hypotheses about differences in variance. We illustrate the procedure and warn against potential misuses of multilevel modelling to analyse differences in variability. We also provide MLwiN and SAS PROC MIXED syntax necessary to run this kind of analyses. In some specific, simple cases, SPSS MIXED can also be used. Copyright © 2014 John Wiley & Sons, Ltd.

Most statistical analyses of research data focus on mean differences between naturally occurring or experimentally created categories, the relations between continuous measures, or some combination of both. However, there are also research questions that concern the variability of the data rather than group means or relations between variables. Currently, ways to analyse variability are not very well known in the larger scientific community. They are only rarely part and parcel of the student training in statistics and are seldom exploited by more established researchers. Here, we present multilevel modelling as a convenient method to model the variability of the data as a function of predictor variables. We also provide guidelines on how to use this tool in practice.

Heterogeneity of variance is often seen or treated as a problem rather than an interesting phenomenon in and of itself. Homogeneity of variance (also called homoscedasticity) is a well-known data assumption in the general linear model. When assessing the relation between a predictor and an outcome variable, the residual variance at different values of the predictor should be equal. If there is heterogeneity of variance (also called heteroscedasticity), conventional tests of statistical inference are biased. For categorical predictors, the so-called “Levene’s test” (see Conover, Johnson, & Johnson, 1981) is not really one test but rather an analysis strategy generating several, related tests that all test the null hypothesis that the variance is the same across groups. This strategy is often used to test the homoscedasticity assumption, even if a recent survey of published psychological research showed that heteroscedasticity is still prevalent and uncorrected for in many cases (Ruscio & Roche, 2012). Rosopa, Schaffer, and Schroeder (2013) gave a comprehensive overview of procedures for statistical inference when the homoscedasticity assumption is violated (see also Grissom, 2000 and Sharma & Kibria, 2013). However, they have focused exclusively on heterogeneity of variance as a problem that needs to be corrected. Others have argued that heterogeneity of variance between conditions indicates that important interactions between condition and individual characteristics have not been included in the model (Bryk & Raudenbush, 1988; Kim & Seltzer, 2011). Although this approach treats heteroscedasticity as substantively interesting, it still regards it as a problem that needs to be corrected. Furthermore, the method for assessing heterogeneity discussed by Bryk and Raudenbush (1988) is not readily available in statistical software.

In multilevel modelling, heterogeneity of variance can be taken into account in the model. This means that rather than resorting to alternative methods when the assumption of homoscedasticity has been violated, one can simply account for it in the model. In the same way that the scores on the dependent variable are modelled to depend on a range of predictor variables, the residual variance can also be modelled as a function of one or more predictors. In addition to correcting for heteroscedasticity, this feature of modelling heterogeneity of variance can be used to test substantive hypotheses about (differences in) variance. That is what this article sets out to explain.

An interest in variability might take many forms. One situation in which researchers should be interested in the variability of data rather than, or in addition to, the pattern of means is in the study of consensus. Consensus means that people tend to have similar opinions, attitudes, or behaviours. In other words, there is limited inter-individual variability. Studying consensus is thus one example of an interest in variability and a particularly relevant one for social psychology. It is easy to think of theoretical questions or applied settings in which consensus is of interest. For example, it is possible that...
the more strongly one identifies with a group, the more similar one’s opinions will be to those of other group members. Another example is whether actual interactions among group members make their opinion more similar or bring them to converge to the group average, thereby increasing group consensus. Ostroff and Fulmer (2014) presented a detailed discussion of ways in which variance (rather than, or in addition to means) can be a worthwhile subject of analysis. These authors distinguish intra-individual, inter-individual, and intergroup variability showing that each one of those has interesting applications in social or organisational psychology. They treat variability both as a predictor and as an outcome variable. We focus here more specifically on a statistical method to analyse differences in variance of the dependent or outcome variable.

**Classical Approaches to Consensus as an Example of Studying Variability**

How has variability heretofore been assessed in the literature? As we already mentioned, inter-individual consensus is a well-known example. One way by which consensus among participants has been approached builds upon the seminal work by Katz and Braly (1933) and is known as the measure of uniformity. In this measure, respondents are invited to describe a target group with a given number of traits, often taken from a larger list of traits. Researchers then examine the number of traits that is needed to include 50% of all the selections made by all participants. The fewer traits that are needed to reach this critical percentage, the more there is consensus among respondents. This measure was refined by Haslam et al. (1998) to compute a coefficient of agreement. Given that each participant is requested to select a specified number of traits from a list, one counts the number of times that each trait selected by a participant is also selected by another participant in the same condition, divided by the maximum number of times this could be the case. Of course, one can think of a variety of measures of within-group variability in order to materialize the consensus or homogeneity of the people belonging to a given group or condition (Klein, Conn, Smith, & Sorra, 2001).

What is common about all these approaches to studying consensus is that the analysis occurs at the level of the group. That is, some sort of index of variability or consensus is calculated for the group as a whole, and this group-level construct is related to group-level predictor or outcome variables. Clearly, when the interest of the researchers, and thus the unit of analysis, is the group or differences between groups, this strategy proves satisfactory. At the same time, such a group-level approach does not deliver much information with respect to what happens at the within-group level. The method for studying variability that we will present later is based on multilevel modelling, and allows studying groups and individuals at the same time without sacrificing one level or the other.

**Multilevel Approaches to Studying Variability**

In general, multilevel modelling is a well-suited technique to analyse nested data at different levels (and it can also be extended to non-nested data such as cross-classified or multiple membership data). Regarding consensus in particular, the technique allows a study of consensus at a lower level (i.e. within-group rather than between-group) than in the approaches discussed earlier. The same idea, however, can be applied to intra-individual variability: multilevel modelling can analyse intra-individual and inter-individual variability at the same time. Multilevel modelling thus seems to be a very promising technique for studying variability. Before we turn to a more detailed presentation, it is important to realize that the use of multilevel modelling by no means guarantees that one is posing the problems in the right manner. Indeed, the use of multilevel modelling for studying variability has not always been correct.

One example of a problematic application of multilevel modelling to the study of variability can be found in recent studies on emotions (Seger, Smith, & Mackie, 2009; Smith, Seger, & Mackie, 2007). Smith and colleagues (2007) predicted that emotions related to the group would be less variable than individual emotions. More specifically, these authors wanted to know whether a series of emotions that people felt “as a group member” were socially shared within a group and more so than the same individual emotions. Smith and colleagues (2007) predicted that group emotions would be less variable than individual emotions. More specifically, these authors wanted to know whether a series of emotions that people felt “as a group member” were socially shared within a group and more so than the same individual emotions. Of course, one can think of a variety of measures of within-group variability in order to materialize the consensus or homogeneity of the people belonging to a given group or condition (Klein, Conn, Smith, & Sorra, 2001).

What is common about all these approaches to studying consensus is that the analysis occurs at the level of the group. That is, some sort of index of variability or consensus is calculated for the group as a whole, and this group-level construct is related to group-level predictor or outcome variables. Clearly, when the interest of the researchers, and thus the unit of analysis, is the group or differences between groups, this strategy proves satisfactory. At the same time, such a group-level approach does not deliver much information with respect to what happens at the within-group level. The method for studying variability that we will present later is based on multilevel modelling, and allows studying groups and individuals at the same time without sacrificing one level or the other.

One instructive consequence of the erroneous use of the mean group emotion as a predictor is that, as long as the scale of the mean emotion is not transformed (i.e. the mean is not multiplied or logged), the regression coefficient of the mean group emotion will be exactly 1 whenever it is uncorrelated with other predictors in the model. We applied this model (i.e. predicting a variable by the mean of subgroupings of the same variable) to randomly generated data and confirmed that the regression coefficient for the “mean group emotion” is exactly 1 (Table 1). Moreover, the effect of the “mean group emotion” is highly significant (p < .001), even in these randomly generated data. This shows that this effect is

---

1These authors use the term “convergence” rather than “variability” or “consensus.” They predict that group emotions will converge to the average profile of group emotions, in comparison with individual emotions. Converging to the average is of course the same as being less variable, and we will therefore not use the term “convergence” here.
inevitably implied by the model. In addition, the results for the randomly generated data show that (i) the emotion mean explains all between-emotions variance in the data, even when this variance is very large, and that (ii) this does not only hold for the emotion mean but also for the participant mean. Using the participant means to predict the scores is an error that is intuitively easier to understand because researchers are used to think in terms of individuals as a higher-level classification. In repeated-measures ANOVA, for example, it is rather easy to recognize that the effect of the participant mean would explain all the between-subjects variance in the sample. If someone suggested that such an effect should be substantively interpreted (e.g. as evidence for less variability of scores around the participant mean), any reader would be quick to spot the mistake. What needs to be appreciated here is that taking the emotion means to predict the scores is in fact exactly the same mistake.

How then should one assess differences in variance? Multilevel modelling, when used correctly, can provide an answer. It can test whether variances differ between conditions or whether variance depends on continuous measures. The way in which multilevel modelling achieves this is by modelling the residual variance (Raudenbush & Bryk, 2002, p. 131; Snijders & Bosker, 2012, chapter 8). In the same way that the score on the dependent variable can be predicted by a range of variables, so too can the residual variance of the model be a function of predictor variables. When done correctly, this residual variance can be used to test substantive hypotheses about the variance in the data. Although multilevel modelling is increasingly being used in social psychology, this particular feature of multilevel models has not received the proper level of attention yet.

The reader may wonder why it is useful to turn to the residual variance in a context when one is interested in consensus. The rationale underlying the examination of residual variance can be better understood when one realizes that testing whether the residual variance differs between groups or between the values of a continuous predictor is the same as a test of the homoscedasticity assumption of the general linear model. In the case of a categorical predictor, such a homoscedasticity test is routinely carried out with a version of the so-called Levene’s test. Levene’s test consists in taking the absolute values of the deviations from the group mean or group median, and subjecting those to a one-way ANOVA (Brown & Forsythe, 1974; Conover et al., 1981; Fox, 2008). A statistically significant Levene’s test means that you have detected a difference in (residual) variation between conditions. By using deviations from the group mean or median, one eliminates differences between groups from the data and analyses residual variation rather than raw variation. The multilevel method we describe here is, compared with the Levene’s test, a more elegant and more generally applicable method to test homoscedasticity. What the Levene’s test achieves using the workaround of an ANOVA of absolute deviations is formally incorporated in the multilevel model by allowing one to model the residual variance. Not only does this approach work with continuous predictors in addition to categorical ones, it can also test whether the residual variance depends on main effects and interactions.

A note about significance testing is warranted when one uses multilevel modelling to test heterogeneity of variance. A significance test for whether the residual variance is a function of one or more independent variables can be done using a likelihood ratio or deviance test based on the full or restricted maximum likelihood deviance. In this test, the deviance (which is another name for minus two times the log of the likelihood of the model or $-2\log\text{Likelihood}$) of two models is compared. These two models are the models with versus without a predictor for the variance. The difference in deviance between those models follows a chi-squared distribution with the difference in the number of parameters between the models as the number of degrees of freedom. Say that we have a first model that assumes homogeneity of variance and a second model that lets the residual variance be predicted by a continuous variable. In that case, the second model has one extra parameter, and therefore, the likelihood ratio test is calculated with one degree of freedom. In case the variance is very small, a deviance test for a difference in variance may become problematic because the variance lies close to its boundary zero (Stoel, Garre, Dolan, & van de Wittenboer, 2006). However, we believe that such cases of very small variance (at level 1) are rare. A more general cautionary remark is that the validity of the deviance test depends on whether the model is well specified (e.g. all important fixed or random effects have been included) and the data fit the assumptions (e.g. a normal distribution in the case of a normal response model).

If the deviance test is used to compare models that differ in the fixed effects, then only the full maximum likelihood deviance should be used.
While the deviance test is an easy and appropriate way to test whether the heterogeneity of variance is statistically significant, it is important to point out that more advanced methods are available too. Bootstrapping or a Bayesian approach such as Markov chain Monte Carlo methods are arguably even better methods to test hypotheses about differences in variability. However, the complexity of such techniques puts them beyond the scope of this article.

Before we illustrate the deviance test in more detail, let us turn to a few examples from the literature.

EXAMPLES

There are some examples in the literature of using the aforementioned method to test substantive hypotheses about variance. Multilevel modelling was partly developed in education research to account for the nesting of individuals within schools (Goldstein, 1987). In that same context, the possibility of modelling heterogeneity of variance has been used to analyse the variance in educational outcomes. Snijders and Bosker (2012, 124–128) provided a detailed description of such an analysis. One of the relevant findings is that the higher the socioeconomic status (SES) of pupils or school, the less variability there is in the educational outcomes of pupils. In other words, outcome diversity is lower for pupils or schools with higher SES. The SES-dependent variance here is clearly a factor of interest rather than merely a nuisance.

A second example comes from our own research on group-based emotions. Group-based emotions are emotional reactions to group concerns (rather than individual concerns). In one study, we wanted to test the idea whether group-based emotions could be strengthened by small group discussions that focused on the group concern that elicited the emotion (Kuppens, Yzerbyt, Dandache, Fischer, & van der Schalk, 2013, Experiment 2). Group-based indignation was indeed stronger when the emotion-eliciting topic had been discussed in small groups, compared with when an irrelevant topic was discussed. Another question, however, is whether interaction in small groups increases the consensus among group members. Do group members report more similar levels of indignation after a discussion on a relevant rather than an irrelevant topic? We tested this idea by comparing a model in which homogeneity of variance across the two conditions was assumed, with a model in which variances were allowed to differ. We found that the residual variance in indignation after the group discussion was smaller when the group had discussed the event that had elicited the indignation rather than when the group had discussed an irrelevant event. A deviance test of the difference between the models gave a $p$-value of .004, meaning that the model with separate variances per condition was significantly better.\(^3\)

**Variance, Fixed Effects, and Residual Variance**

When using this technique of modelling the residual variance, it is important to be aware that it models the residual variance, that is, after taking into account all the other effects in the model. Remember that the Levene’s test also eliminates group differences by using deviation scores. It is therefore crucial to carefully consider what kind of variance needs to be analysed and which effects (on the mean level of the dependent variable rather than the variance) need or do not need to be taken into account. We illustrate this with an example that we have touched upon earlier, namely the consensual nature of group emotions. Let us assume that participants rated a series of emotions twice: once as an individual and once as a group member. The question is whether group emotions are more consensual than individual emotions. Clearly, a greater consensus on group emotions than on individual emotions could be a consequence of group members thinking about similar issues or concerns when they are rating emotions “as a group member,” something that is less likely to occur when they are rating emotions “as an individual.”

Here, we use the data from Smith et al.’s (2007) Study 1.\(^4\) These authors had 101 participants and measured their individual emotions as well as their group emotions as an Indiana University student (a series of 12 emotions were measured for each type of emotion). We fitted a model with both individual and group emotions as the dependent variable (this means 24 data points per participant). We included a covariance between individual and group emotions because ratings of an individual emotion (e.g., individual happiness) might be related to ratings of the same group emotion (e.g., group happiness). We also included a random effect for participants. Because we are not interested in between-emotions variance, we added fixed effects for emotion (the 12 emotions), type of emotion (individual versus group emotion), and their interaction. The fixed part of the model therefore explains all the between-emotions variance in the data; it is a saturated model of the emotion predictors. If we had not done that, the residual variance would be made up of both between-emotions and within-emotion variance, and the residual variance could not be used as an indicator of the variability of individual versus group emotions. This shows that it is important to consider the effects present in the model in order to understand and interpret the meaning of the residual variance.

So, after making sure that all the between-emotions variance in the data is explained by the fixed effects in the model,\(^5\) we then let the remaining residual variance be predicted by the type of emotion (individual vs group emotion). If group emotions are more consensual, the residual variance should be lower for group emotions than for individual emotions and the effect of type of emotion on the variance should be significant. The effect was contrary to Smith et al.’s (2007) prediction. Type of emotion had a significant effect on the variance ($p = .002$), but the residual variance was higher for emotions as an Indiana University student (1.46) than for individual emotions (1.23).\(^6\) Other data from Smith and colleagues (2007) as well as our own data are consistent with this result. Together, they provide strong evidence that group emotions

\(^{\text{3}}\)This particular finding (the difference in variance) was not actually discussed in the text of the paper, but the relevant parameter is at the bottom row of Table 3 of this same paper.

\(^{\text{4}}\)These data were kindly provided to us by Eliot Smith.

\(^{\text{5}}\)Alternatively, this could also be done by mean centring all 24 emotions separately.

\(^{\text{6}}\)More details on this analysis plus the SPSS, SAS PROC MIXED, and MLwiN syntax needed to run these models can be found in the online appendix (see Supporting Information).
as measured by Smith and colleagues (2007) show more rather than less variability in comparison with individual emotions.

MODELLING VARIANCE HETEROGENEITY IN SAS AND MLWIN

How is this procedure implemented by means of existing statistical software? Here, we provide the syntax for SAS software (PROC MIXED procedure) and MLwiN. MLwiN uses a graphical interface but also has a command language. For modelling variance heterogeneity using the graphical interface, there is an excellent step-by-step tutorial available in the user’s guide (Rasbash, Steele, Browne, & Goldstein, 2012, chapter 7, 89–106), which can be freely downloaded from the Centre for Multilevel Modelling website. Even if one does not use or plan to use MLwiN, we recommend reading this tutorial. Here, we present a way to build MLwiN models using the command language or syntax. SPSS also has a module for multilevel modelling, but it cannot model heterogeneity of variance, except for a few very specific cases that are discussed later. Table 3 contains a summary of the kind of models that can be handled by SAS, MLwiN, and SPSS using the methods we describe here. R also offers the possibility of modelling heterogeneity of variance (see Pinheiro & Bates, 2000, chapter 5), but we did not look into this issue in detail.

Using fictitious data (Table 2), we deal with three types of situations. First, we discuss the case in which one wants to predict variance with a categorical between-subjects variable. A second situation is when variance is predicted by a categorical within-subjects variable. Depending on whether there are only two or more within-subjects measurements, this is either the most simple or the most complex situation. In the third type of situation, variance depends on the values of a continuous predictor variable. After having presented the syntax for these three situations applied to our small fictitious dataset, we will analyse real data from a published study on small group interaction (Smith & Postmes, 2011).

Variance Predicted by a Categorical Between-subjects Variable

When residual variance depends on a categorical between-subjects variable, the syntax and procedure is quite straightforward. Imagine an outcome variable $Y$ and a predictor variable $X$ that has three categories. As an example of such a situation, we analyse the data presented in Table 2.

Table 2. Example data

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Z</th>
<th>Y</th>
<th>Xdum1</th>
<th>Xdum2</th>
<th>Xdum3</th>
<th>IDpart</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.71</td>
<td>7.91</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.49</td>
<td>7.24</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.68</td>
<td>6.91</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.82</td>
<td>6.45</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.13</td>
<td>9.80</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.92</td>
<td>7.96</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.57</td>
<td>7.79</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1.00</td>
<td>6.07</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1.54</td>
<td>6.12</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1.09</td>
<td>8.42</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1.64</td>
<td>6.92</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1.29</td>
<td>3.88</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>2.31</td>
<td>3.94</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>1.85</td>
<td>9.49</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>1.78</td>
<td>10.62</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>2.94</td>
<td>5.53</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>2.32</td>
<td>6.57</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>0.90</td>
<td>9.07</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>2.71</td>
<td>7.47</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>2.41</td>
<td>6.69</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>1.78</td>
<td>7.61</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>4.67</td>
<td>1.92</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>2.40</td>
<td>2.88</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>2.62</td>
<td>9.03</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>4.47</td>
<td>8.40</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>2.32</td>
<td>11.45</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>3.89</td>
<td>13.37</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>3.66</td>
<td>12.02</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>4.55</td>
<td>4.37</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>2.94</td>
<td>1.55</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Results for the SAS model where the variance depends on the categorical between-subjects variable $X$ (AIC=Akaike information criterion, AICC=Corrected Akaike information criterion, BIC=Bayesian information criterion)
Proc mixed data=example method=REML noclprint covtest;
Class X;
Model Y = X / solution ddfm=kr;

The deviance (or \(-2\)LogLikelihood) of the simple model is 141.2 (the output of the simple model is not presented here). The deviance of the model with heterogeneous variance is 127.20 (see the “Fit statistics” table in Figure 1; a more precise number can be found in the “Iteration history,” which is not presented here). The difference between these two is 14.00, and the difference in the number of parameters between the models is two (one residual variance in the simple model and three residual variances in the heteroscedastic model). A chi-squared test with a value of 14.00 and two degrees of freedom gives a \(p\)-value of .0009. Our conclusion therefore would be that the variance differs significantly between the three conditions.

**MLwiN**

In MLwiN, the procedure is similarly easy. Using the data from Table 2, the following syntax will fit the simple model with homogeneous variance:

```snack
iden 1 "id"
expl 1 "cons"
setv 1 "cons"
resp "y"
expl 1 "xdum2" expl 1 "xdum3"
```

where “cons” is a column existing entirely of 1s, “id” is a unique identifier for each case, and x dum2 and x dum3 are dummy variables indicating \(x=2\) and \(x=3\), respectively, with \(x=1\) as the reference category. If you hit “Start,” you will see that the ML deviance of this model is 146.003. To add the heterogeneous variance to this model, write

```snack
sete 1 "xdum2" "cons"
sete 1 "xdum3" "cons"
```

Now hit “More” or “Start,” and you will see that the ML deviance for the model with heterogeneous variance is 130.502 (Figure 2). The difference in deviance between the models is 15.501, and a deviance test with two degrees of freedom gives \(p=0.0004\). These numbers are slightly different to what we obtained with SAS PROC MIXED because MLwiN only presents the full maximum likelihood deviance whereas we had requested the restricted maximum likelihood deviance in SAS. These estimation methods give slightly different results. Both the full and the restricted maximum likelihood deviance can be used for a deviance test for random parameters (such as a difference in variance or a random slope), but only full maximum likelihood can be used for a deviance test for fixed parameters. In the case of random parameters, we advise the use of the deviance test based on the restricted maximum likelihood deviance when possible.

As mentioned before, SPSS cannot model heterogeneity of variance. In this particular case, however, we can trick SPSS into fitting the model that we want by making it believe that the variable “X” represents a within-subjects rather than a between-subjects factor. This is achieved with the following syntax:

```snack
MIXED Y by X /print = solution testcov gr 
/FIXED=X | SSTYPE(3) 
/METHOD=REML 
/Repeated=X | Covtype(diag) subject(ID).
```

The “/Repeated = X” command indicates that X represents repeated measurements nested within individuals, which is not actually the case here but is needed to estimate separate variances for the three between-subjects groups defined by X. “Covtype(diag)” requests a diagonal variance–covariance matrix, which means that the three variances are estimated separately and allowed to differ, but all covariances are set to zero (Figure 3). Covariances, of course, are impossible and meaningless between the groups of a between-subjects factor. The simpler model with homogeneous variance can be estimated by replacing “Covtype(diag)” with “Covtype(id),” which estimates a variance–covariance matrix in which all three variances are equal. The restricted maximum likelihood deviance for these models is 127.20 and 141.15, respectively. This is the same as in SAS PROC MIXED.

The downside of using SPSS is that its capabilities are limited to these kinds of simple models. If there was a real within-subjects design in these data (in addition to the between-subjects factor X), this could not have been modelled because we already used the Repeated statement to model the between-subjects factor. Furthermore, SPSS cannot let the residual variance depend on a continuous variable, which we will see later is the function that allows the most flexibility.

**Variance Predicted by a Categorical Within-subjects Variable**

When the categorical variable predicting the variance is a within-subjects variable, there is an additional issue to be
taken into account. The within-subjects variable defines several measurements that are nested within the same person. The measurements will therefore tend to be correlated, and this within-subject correlation is something the model should allow for. Note that there can be no correlation in the case of a between-subjects variable because every case can belong to one group or between-subjects condition only.

In some cases, this can be done easily using SPSS or SAS. When the within-subjects variable has only two categories (i.e., there are only two measurements for which the variance needs to be compared), SPSS or SAS can be used to test the difference in variance. For more information, see the analysis of the variance of individual versus general group emotions, presented in the online appendix (see Supporting Information) and discussed later. Another case in which this “easy” analysis in SPSS or SAS can be used is when the covariance structure between all within-subjects measurements follows some specific structure available in SPSS/SAS (e.g., compound symmetry, first-order autoregressive, or Toeplitz). In such cases, the test for different variances is a straightforward extension from what is explained in the online appendix (see Supporting Information). For example, when the within-subject variable is “time,” then the autoregressive or Toeplitz structures can be adequate.

A more general way to model this within-subject dependence, however, is by borrowing from the structure of the multivariate multilevel model (Goldstein, 1999, chapter 4; Hox, 2010, chapter 10; Snijders & Bosker, 2012, chapter 16). The multivariate multilevel model can be used to model all fixed and random effects separately for different outcome variables or within-subject measurement. However, we will use this technique here only to model the covariances between within-subject measurements. Going back to our fictitious data in Table 2, the variable X now indicates which within-subject measurement the data point refers to, rather than a between-subjects condition. “IDpart” identifies the participant. Modeling the covariances between the three measurements is achieved by including three dummy variables for the three measurements defined by X. All three dummies are given a random slope at level 2 (the level of individual participants), which creates a full variance–covariance matrix. Finally, the variance components of that matrix are set to zero so that only the covariances between the three measurements are estimated. This can be done in MLwiN. As far as we know, SAS does not offer this possibility in the simple way described here. More complex applications of modelling within-subjects variance using SAS software exist (Hedeker, Mermelstein, & Demirtas, 2008), but they require more technical and statistical knowledge than the models presented here. The syntax given here will fit these models in MLwiN, but for a full understanding, we recommend consulting Hox (2010, chapter 10) or Snijders and Bosker (2012, chapter 16).

### MLwiN

We first need to sort the data because they are out of order for this analysis. Remember that “IDpart” indicates the participants and that data in MLwiN should be ordered by the higher-order units. The following syntax will sort the data.

```plaintext
sort on 2 keys in "idpart" "x" carrying "id" "zi" "yi" "xdum1" "xdum2" "xdum3" results to "idpart" "x" "id" "zi" "yi" "xdum1" "xdum2" "xdum3"
```

Now, the model with homogeneity of variance can be built with this syntax:

```plaintext
clear
iden 1 "id"
iden 2 "idpart"
expl 1 "cons"
setv 1 "cons"
resp "y"
expl 1 "xdum1"
expl 1 "xdum2"
expl 1 "xdum3"
sete 2 "xdum1" "xdum2"
sete 2 "xdum1" "xdum3"
sete 2 "xdum2" "xdum3"

fpar 0 "xdum1"
```

Note that each covariance can be set separately in MLwiN using the `sete` command, which allows great flexibility. The deviance for this model is 142.502. Adding heterogeneity of variance according to X,

```plaintext
sete 1 "xdum2" "cons"
sete 1 "xdum3" "cons"
```

The deviance for the model with heterogeneity of variance is 127.896 (Figure 4). The difference in deviance is 14.606, and the deviance test with two degrees of freedom gives a p-value of .0007.

### Variance Predicted by a Continuous Variable

#### SAS PROC MIXED

In SAS PROC MIXED, the residual variance can also be made to depend on a continuous variable, say Z, using the `Repeated` statement (Littel, Milliken, Stroup, Wolfinger, & Schabenberger, 2006, p. 405). The statement

```plaintext
proc mixed;
  class subject time;
  model y = x z / s;  
  repeated time / subject = subject type = cs; 
  estimate 'Two-way interaction' x*z / s;
run;
```

The syntax given here will fit these models in MLwiN, but for a full understanding, we recommend consulting Hox (2010, chapter 10) or Snijders and Bosker (2012, chapter 16).
Repeated / local=exp(Z); will let the residual variance depend on the variable Z. This gives the following syntax:

\begin{verbatim}
Proc mixed data=example method=REML noclprint covtest;
Model Y = Z / solution ddfm=kr;
Repeated / local=exp(Z);
\end{verbatim}

The results for this model can be found in Figure 5. In the “Covariance parameter estimates” table in Figure 5, there is a significance test associated with the “EXP Z” parameter, but it is better to perform the deviance test. We again compare the deviance of the heteroscedastic model with the deviance of a simpler, homoscedastic model. The syntax of the simpler model is as follows:

\begin{verbatim}
Proc mixed data=example method=REML noclprint covtest;
Model Y = Z / solution ddfm=kr;
\end{verbatim}

The deviance of this model can be found in Figure 5. In the “Covariance parameter estimates” table in Figure 5, there is a significance test associated with the “EXP Z” parameter, but it is better to perform the deviance test. We again compare the deviance of the heteroscedastic model with the deviance of a simpler, homoscedastic model. The syntax of the simpler model is as follows:

\begin{verbatim}
Proc mixed data=example method=REML noclprint covtest;
Model Y = Z / solution ddfm=kr;
\end{verbatim}

The deviance of the simple model is 145.1 (output for this model is not presented). The deviance of the model with heterogeneous variance is 135.14 (see “Fit statistics” in Figure 5). The difference between these two is 9.96, and the difference in the number of parameters between the models is one (one residual variance in the simple model and one residual variance plus one parameter associated with differences in variance due to X in the heteroscedastic model). A chi-squared test with value 9.96 and one degree of freedom gives a p-value of .0015. Our conclusion therefore would be that there is a significant increase in variance when Z increases.

It is important to note that the “local=exp()” statement provides researchers with a very flexible means to address a variety of research questions. As a matter of fact, any relevant set of variables, whether continuous, dummy-coded, or contrast-coded variables, or even interactions, can be included between the parentheses. In the present case, one could think of including two contrast codes in order to further examine which specific conditions of the categorical variable differ from each other in terms of their residual variances. Hoffman (2007) discussed an example of a continuous variable predicting within-person variability, and she provided all the SAS syntax on her website (http://psych.unl.edu/hoffman/Sheets/MLM_HowTo.htm).

MLwiN

The following syntax builds the model with X as a continuous predictor and homogeneous variance:

\begin{verbatim}
clear
iden 1 "id"
expl 1 "cons"
setv 1 "cons"
resp "y"
expl 1 "x"
\end{verbatim}

The deviance of this model is 146.107. In order to make the residual variance depend on X, we run the following command:

\begin{verbatim}
sete 1 "cons" "X"
\end{verbatim}

Deviance for this model with heterogeneous variance is 131.889 (Figure 6). The difference in deviance with the simple model is 14.218, and a deviance test with one degree of freedom gives a p-value of .0002. Note again the difference with the SAS results; this is mainly due to the fact that SAS models an exponential function of the variance whereas MLwiN models a linear function.

As in SAS PROC MIXED, using continuous variables as predictors of the residual variance opens up a large range of possibilities. The graphical interface of MLwiN makes the models intuitively easy to understand.

The Power of Talk

We now turn to an example of real data. As mentioned before, we have found that small group discussions can decrease the
variability of ratings of group-based emotion, that is, an emotional reaction to group concerns (Kuppens et al., 2013). Apparently, the small group interaction not only increased the intensity of the emotion but also the within-group consensus around the emotional intensity. We will now test whether this result is replicated in another study that used small group discussions. Smith and Postmes (2011, Study 2) made British participants discuss different topics in each of three conditions. In the stereotype condition, participants discussed the stereotypes but also a plan to combat immigration problems. In the action condition, they discussed types of immigrants. In the action condition, they discussed stereotypes but also a plan to combat immigration problems. In the irrelevant condition, participants discussed the British monarchy. The dependent variable was participants’ action intentions.

In these data, the heterogeneous variance can be estimated using SPSS. This is achieved by making SPSS believe that condition is a within-subjects factor, as explained earlier. The model with homogeneous variance is as follows:

```
mixed action by condition groupno
/print = solution testcov g r
/fixed = condition |sstype(3)
/method = reml
/random = groupno
/repeated = steronly |covtype(diag) subject (pptno).
```

where “action” is the measure of action intentions, “groupno” indicates the nesting in discussion groups, “pptno” is the participant number, “condition” is the between-subjects manipulation, and “steronly” is a variable that contrasts the stereotype condition with the other conditions (the exact coding does not matter). For this model, omitting the Repeated statement would make no difference. The model with heterogeneous variance can be estimated with the following syntax:

```
mixed action by condition groupno
/print = solution testcov g r
/fixed = condition |sstype(3)
/method = reml
```

The only thing that has changed is the “covtype (diag),” indicating that this time, we want a diagonal rather than an identity matrix to be estimated. This represents a change from homogeneous to heterogeneous variance according to “steronly.” The deviance was 63.23 and 61.79 for the first and second models, respectively. The difference is 1.44, and a deviance test with one degree of freedom gives a p-value of .23. Based on these data, we would not reject the null hypothesis that variances are equal, although we have to note that the residual variance was smaller in the stereotype condition (.07) than in the other conditions (.11), which is consistent with our own research (Kuppens et al., 2013).

In MLwiN, the model with homogeneous variance is fitted with this syntax:

```
iden 1 "pptno"
iden 2 "groupno"
resp "actionint"
expl1 "cons"
setv 1 "cons"
setv 2 "cons"
expl1 "condition"
```

where “actionint” is the measure of action intentions, “groupno” indicates the nesting in discussion groups, “pptno” is the participant number, “condition” is the “condition” is a categorical variable indicating the between-subjects manipulation, and “cons” is a constant variable consisting entirely of 1s. Adding heterogeneous variance is achieved with the following command:

```
sete 1 "cons" "condition_3"
```

The deviance of these models is 54.148 and 52.658, respectively. The p-value for the deviance test is .22. The difference with SPSS is due to the fact that MLwiN only gives the full maximum likelihood deviance, whereas we requested restricted maximum likelihood deviance in SPSS. Both are valid for a deviance test of differences in variance, but there is a preference to use restricted maximum likelihood when possible.

Using SAS PROC MIXED, the syntax to fit the model with heterogeneous variance is as follows:

```
proc mixed method=REML noclprint covtest;
class condition steronly groupno;
model action = condition / solution ddfm=kr;
random groupno;
Repeated / group=steronly;
```

For the model with homogeneous variance, simply omit the Repeated statement. Even if both models were set up in a different way, the results for these models are identical to the results for the SPSS models.

Table 3 summarizes in which cases heterogeneity of variance can be modelled by the different software programmes in the way that we have discussed.
Table 3. Summary of functionality of SAS, MLwiN, and SPSS software using the methods discussed in this article

<table>
<thead>
<tr>
<th>Variance depending on between-subjects factor, without within-subjects design in data</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance depending on within-subjects factor, also with within-subjects design present in the data</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Variance depending on within-subjects factor, only two measures</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Variance depending on between-subjects factor, with also a within-subjects design in data</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Variance depending on between-subjects factor, without within-subjects design in data</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Variance depending on continuous variable</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Variance depending on interaction</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Conclusion:** “If all you have is a hammer, everything looks like a nail”

Even though it is easy to think of examples in which differential variability of the data is of substantive interest, such substantive ideas are only rarely formally tested. This might be an example of how the available tools in the data analytic toolbox influence the approach taken in the data analysis. Researchers might simply not have tested ideas about differential variance because the tool for doing so was unavailable or simply unknown to them. Similarly, the famous quote “you once know that hierarchies exist, you see them everywhere” (Kreft & De Leeuw, 1998, p. 1) refers to the fact that after learning to analyse hierarchical data, one starts seeing hierarchies everywhere. We hope that our explanation of how to model residual variance as a function of independent variables will be such an extra tool, one that will enable researchers to look at their data slightly differently and to test hypotheses about differential residual variance when these follow from their theoretical framework.

**REFERENCES**


**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article at the publisher’s web-site.