Leppink, J., & van Merriënboer, J. J. G. (2015). The Beast of Aggregating Cognitive Load Measures in Technology-Based Learning. *Educational Technology & Society, 18* (4), 230–245.

The Beast of Aggregating Cognitive Load Measures in Technology-Based Learning

Jimmie Leppink[*] and Jeroen J. G. van Merriënboer
School of Health Professions Education, Maastricht University, The Netherlands //
jimmie.leppink@maastrichtuniversity.nl // j.vanmerrienboer@maastrichtuniversity.nl
[*]Corresponding author

ABSTRACT

An increasing part of cognitive load research in technology-based learning includes a component of repeated measurements, that is: participants are measured two or more times on the same performance, mental effort or other variable of interest. In many cases, researchers aggregate scores obtained from repeated measurements to one single sum or average score per participant and use these aggregated scores in subsequent analysis. This paper demonstrates some dangers of this commonly encountered aggregation approach and presents two comprehensive alternatives: Split-plot analysis of variance (ANOVA) and more flexible two-level regression analysis. The core message of this paper is that the application of the aggregation approach can seriously distort our view of effects and relations of interest and should therefore not be used in cognitive load research. Multilevel analysis of repeated measurements data can account for various features of the data and constitutes a best practice.

Keywords

Technology-based learning, Medical education, Repeated measurements, Split-plot design, Multilevel analysis

Introduction

An increasing part of cognitive load research in technology-based learning, and cognitive load research in general, includes a component of repeated measurements. Then, participants – frequently students, employees, managers or clients – and in statistical language more commonly referred to as subjects are measured repeatedly (i.e., at least two times) on the same variable(s) of interest. Be it performance, mental effort invested in a cognitive activity or some other variable, repeated measurements data have one nice feature: they enable researchers to separate so-called within-subjects variance from between-subjects variance. When we have a sample of individual students perform an exam only once, we can distinguish between participants, for some students perform better than others do. In such a context, however, we cannot account for variation within participants, since we have only one measurement per student. In educational research, we are usually not interested exclusively in distinguishing between participants; we are interested in within-participants changes related to learning and development as well, and for the latter we need studies that include a component of repeated measurements.

Recent calls for repeated measurements of cognitive load

Recently, two series of well-designed randomized controlled experiments (Schmeck, Opfermann, Van Gog, Paas, & Leutner, 2015; Van Gog, Kirschner, Kester, & Paas, 2012) provided evidence for the statement that studies where learners have to perform a series of tasks, it is better to measure a characteristic of interest – for instance mental effort invested in a cognitive activity (Paas, 1992) – after each task (i.e., repeatedly) than once respectively. A main finding of both series of experiments was that single retrospective (i.e., delayed) ratings were significantly higher than the average of ratings obtained after each task. This probably happens because delayed ratings are mainly influenced by the relatively more complex problems (Schmeck, et al., 2015).

In another experiment (Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013), students solved problems on conditional and joint probabilities in two modes: in an explanation of six lines of text, and in formula notation. In one condition, students first studied the textual explanation and then the formula explanation, while in the other condition the order was reversed. Various cognitive load measures were included in this experiment (Ayres, 2006; Cierniak, Scheiter, & Gerjets, 2009; Leppink, et al., 2013; Paas, 1992; Salomon, 1984). However, instead of administering these measures at the end of the full order of two formats, the measures were administered after each format. This enabled researchers to not only test for differences between formats but also test for format order effects. One finding that has implications for instructional design is that students who are confronted with the

formula format before the textual explanation tend to experience higher extraneous cognitive load from the formula format than their peers who study exactly the same materials on joint and conditional probabilities in reversed order. The latter implication could not have resulted from aggregation of extraneous cognitive load scores across formats.

Repeated measurements data enable researchers to separate between-participants and within-participants variance. Unfortunately, many researchers have failed to appreciate this nice feature of repeated measurements data and have aggregated scores from repeated measurements to one sum or average score per participant (e.g., Ayres, 2006; Corbalan, Kester, & Van Merriënboer, 2008; Hoffman, 2012; Koriat, Nussinson, & Ackerman, 2014; Kostons, Van Gog, & Paas, 2012; Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013). Aggregating repeated measurements, we wipe out all within-participants variance and this can result in serious distortions of our views of effects and relations of interest (Leppink, 2015). Figure 1 provides an example of what can happen when studying the correlation between two quantitative variables that have been measured repeatedly but for which repeated measurements have been aggregated.

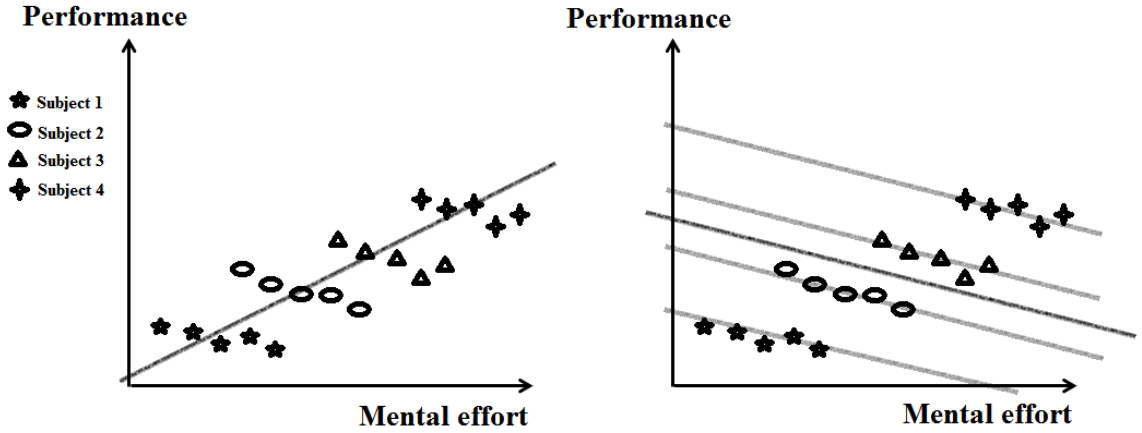

Figure 1. Ignoring repeated measurements (left) vs. accounting for repeated measurements (right)

In the example depicted in Figure 1, four students (small number for the sake of illustration) perform five tasks in counterbalanced order and rate the invested mental effort after each of five tasks. Each task yields one mental effort rating and one quantitative performance score. When we aggregate the scores for both measures – or compute a Pearson's correlation coefficient for the twenty (i.e., not aggregated) observations in the scatterplot – we find a positive correlation between mental effort and performance (left part of Figure 1). Thus, we would conclude – in this specific context – that students who report higher mental effort tend to display better performance as well. However, when taking the repeated measurements structure of the data into account, we find a negative correlation between mental effort and performance: for a given student, we would conclude that a task that requires more mental effort appears to yield somewhat lower performance (right part of Figure 1). The latter message has much more meaning in many educational settings, and provides a better interpretation of the available data, than the former message. Note that the example in Figure 1 is hypothetical; it could well be that an aggregation approach yields a negative correlation between mental effort and performance while taking the repeated measurements structure of the data into account results in a positive or no (linear) correlation at all. The point is that the conclusions might become different when the repeated measurements structure is taken into account when analyzing the data.

Monitoring of learning and mental effort

One context in which the comparison between aggregating scores vs. treating repeated measurements as such is important is the context of self-regulated learning. Self-regulated learning can be defined as an active, constructive, metacognitive process (i.e., metacognition is cognition about cognition, in this case: learning; Flavell, 1979). Important metacognitive processes within self-regulated learning are monitoring and control (Zimmerman & Schunk, 2001). The term monitoring is used to refer to the thoughts learners have about their cognition, and learners respond to the environment (e.g., adapt their behavior) based on these metacognitive thoughts. The latter is termed control (Nelson & Narens, 1990).

Self-regulated learning can be studied at three levels: at item level, at task or topic level, and at task-sequence level. Self-regulated learning at item level is about learners' monitoring of how well they have learned smaller pieces of information (e.g., word pairs, small passages of text), which affects how long they engage in (re)studying it (Metcalfe, 2009; Thiede & Dunlosky, 1999). Self-regulated learning at task or topic level involves learners' monitoring of their understanding of information on a particular topic, which determines whether they continue studying one piece of information or move on to another (Azevedo & Cromley, 2004; Azevedo, Moos, Greene, Winters, Cromley, 2008). Finally, self-regulated learning at the level of task sequence is about learners' monitoring of how well they performed a learning task after completing it. This is also referred to as "self-assessment" to distinguish it from monitoring during task performance and is used by learners to select next suitable tasks from an environment comprising tasks at different levels of complexity and with different levels of instructional guidance (Corbalan, et al., 2008; Kostons, et al., 2012).

Effectiveness of self-regulated learning in which learners can choose their own learning tasks greatly depends on the accuracy of students' self-assessment and task selection (Kostons, et al., 2012). Unfortunately, accurate self-assessment is very difficult for learners, and especially novices, as it appears to require some domain expertise and knowledge of assessment criteria (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Inaccuracies in task selection can arise when self-assessment is inaccurate, but even when self-assessment is accurate students frequently do not know how to select suitable tasks. When task selection is inaccurate, this will cause learners to work on tasks that do not match their knowledge level – meaning selected tasks are either too complex or too easy or tasks that have an inappropriate amount of instructional guidance and therefore impose a suboptimal cognitive load on students' minds – and consequently, students will not learn from engaging in self-regulated learning.

In a recent study (Kostons, et al., 2012), students worked on a series of genetics tasks, self-rated task performance on a scale from zero (minimum) to five (maximum) and mental effort on a scale from one (minimum) to nine (maximum) after each task, to use these ratings for selecting the next task (which out of five complexity levels and how much instructional support). Data obtained from such a study hold very interesting information about experienced mental effort from self-selected tasks, can as such provide feedback to learners and teachers about the appropriateness of a given task in terms of complexity and instructional support for a given student at a given stage, and provides a wealth of information for researchers interested in ways to improve the accuracy of such ratings to eventually improve the accuracy of task selection. Unfortunately, instead of approaching the data with flexible multilevel analysis, all ratings and performance measures were aggregated across all tasks performed by students, thereby not differentiating between ratings and performance on easier vs. more complex tasks or between ratings and performance on tasks with more vs. less instructional support. The same has been done in other studies in this context (e.g., Corbalan, et al., 2008; Kostons, Van Gog, & Paas, 2010; Kostons, et al., 2012; Salden, Paas, & Van Merriënboer, 2006).

Supporting the acquisition of self-regulated learning skills is of crucial importance in contemporary education for two reasons. Firstly, such skills are critical to effective learning in schools and other environments. Secondly, a rapidly changing society in which learners must be ready to develop new knowledge and skills autonomously and continuously does a great appeal on self-regulated learning skills. Continuing developments in technology-based education increasingly create room for both the application and research of self-regulated learning.

In fact, all three levels at which self-regulated learning takes place involve repeated measurements with smaller or larger intervals of time in between. Unfortunately, in virtually all studies published in this context, scores obtained from these repeated measurements are aggregated to sum or average scores. Yet, especially in a context of self-regulated learning, using repeated measurements as such appears to make much more sense than using aggregated scores. Suppose, for instance, that a student reports to have invested a rather high mental effort on a task on which this student performs rather poorly. It makes more sense to have the student (choose to) perform a somewhat easier task, and have the student base subsequent task selection on changes in reported mental effort and performance from task to task.

Scales for assessing cognitive load

Subjective rating scales have been used intensively to measure or estimate cognitive load experienced by learners, since the introduction of a nine-point one-dimensional mental effort rating scale by Paas (1992). Mental effort has

been defined as the cognitive capacity allocated to deal with the demands imposed by a cognitive task and as such reflects the actual and overall cognitive load (Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Paas & Van Merriënboer, 1994a). The overall cognitive load is assumed to be the sum of intrinsic cognitive load and extraneous cognitive load (Kalyuga, 2011; Sweller, 2010; Sweller, Ayres, & Kalyuga, 2011). Intrinsic cognitive load is a direct function of the complexity of a task on the one hand and the student's prior knowledge on the other hand (Sweller, 1994). More advanced students have more prior knowledge and can therefore be expected to experience a lower intrinsic cognitive load than less advanced students confronted with the same task. Extraneous cognitive load is due to features in the way in which information (e.g., instruction, explanation) is presented as a result of which the student has to invest mental effort that does not contribute to learning or task performance and is therefore extraneous to learning (Sweller & Chandler, 1994; Sweller, Chandler, Tierney, & Cooper, 1990). An optimal intrinsic cognitive load can be achieved through a proper match between the complexity of tasks and the learner's prior knowledge, while extraneous cognitive load should be minimized. Doing so, we can stimulate students to allocate still available working memory resources to deal with the intrinsic cognitive load and, as such, engage in learning. In a traditional version of cognitive load theory, the latter was referred to as a third category of cognitive load called germane cognitive load (Sweller, Van Merriënboer, & Paas, 1998). However, a recent reconceptualization of the categories of cognitive load resulted in a two-factor intrinsic/extraneous framework in which germane cognitive load is viewed as working memory resources allocated to dealing with intrinsic cognitive load (Kalyuga, 2011; Sweller, 2010; Sweller, et al., 2011). A series of recent studies have provided support for the possibility to distinguish between intrinsic and extraneous cognitive load using a questionnaire comprising three to four statement items with subjective rating scales for each category of load (Leppink, et al., 2013; Leppink, Paas, Van Gog, Van der Vleuten, & Van Merriënboer, 2014).

An advantage of the multi-item questionnaire developed by Leppink et al. (2013, 2014) to the aforementioned single-item mental effort rating scale introduced by Paas (1992) is that it enables researchers to distinguish between two fundamentally different categories of cognitive load; a single-item score simply cannot do that. It is statistically speaking impossible to measure or estimate to what extent a difference on the single-item mental effort rating scale can be attributed to differences in intrinsic or extraneous cognitive load without being absolute sure through solid experimental design that one type of cognitive load is kept constant across conditions. Only in the case that an experimental study fully succeeded in keeping intrinsic cognitive load constant across conditions one might want to interpret changes in mental effort as changes in extraneous cognitive load. However, even in that case one might find researchers debate how one can be sure that intrinsic cognitive load is constant across conditions without having any measure available as a check. Using a multi-item instrument, we can study whether we can use that multitude of items (i.e., directly observed or manifest variables) to measure or estimate one or more latent variables (i.e., variables that cannot be observed directly), in this context types of cognitive load (e.g., Leppink, et al., 2013, 2014). With a single item, that exercise is impossible.

The distinction between intrinsic and extraneous cognitive load is relevant, because an increase in cognitive load does not always hamper learning. When extraneous cognitive load is low, slight increases in intrinsic cognitive load may in fact stimulate learning in that there is more to learn from a task of some complexity than from a task that is very easy given a student's prior knowledge and therefore imposes a very low intrinsic cognitive load on the student's mind. In the latter case, there is virtually no need to allocate working memory resources to deal with intrinsic cognitive load, because there is hardly any intrinsic cognitive load to deal with in the first place.

The most recent version of the questionnaire developed by Leppink et al. (2014) comprises four items for measuring intrinsic cognitive load and four items for measuring extraneous cognitive load. Evidence from the same study suggests, however, that three items for each of the two categories – as reported by Leppink et al. (2013) – can be sufficient. That makes six items for intrinsic and extraneous cognitive load together. A questionnaire of such a size is easy to administer in a single measurement or in a limited number of, for instance, five repeated measurements. A disadvantage of a multi-item questionnaire like this one, however, is that it is more difficult to administer in a study that for instance has twenty repeated measurements. If students (or other individuals) have to learn or perform five tasks, rating six questions after each task may still not be asked too much from the students. However, having to rate six questions after each of twenty tasks is likely to be perceived as a tedious exercise perhaps even before the students have completed ten tasks. This may result in over-consistent rating or other undesirable responding that of course has little to do with actual intrinsic or extraneous cognitive load.

To unite the best of the six-item intrinsic/extraneous cognitive load questionnaire and the single-item mental effort rating scale, a solution in the case of as many as twenty tasks could be to ask students to provide a mental effort rating after each task and only complete the intrinsic/extraneous cognitive load questionnaire after a block of four or five tasks. This way, repeated measurements are available for both mental effort and the two categories of cognitive load, while the risk of undesirable responding is minimized.

Randomized controlled experiments

A major strength of cognitive load theory is that the guidelines and implications coming forth from this theory are generally based on randomized controlled experiments. One might argue that the term "randomized" is redundant here, given that the nature of experiments is that participants are randomly assigned to the categories of an independent variable; without such randomization, we would be dealing with a quasi-experiment. We nevertheless add the term "randomized" because the term "experiment" is sometimes encountered in situations when a study reported was actually a quasi-experiment. Further, the term "controlled" is used because in an experiment, an independent variable is systematically manipulated and, thus, there is a kind of control. Well-designed randomized controlled experiments are of great importance in educational research, because they allow causal conclusions with regard to effectiveness of instruction or some other treatment on learning or performance. Randomized controlled experiments enable us to study the effectiveness of instructional methods before they are implemented in education rather than afterwards.

Example

Let us consider the following example. The advent of technology has contributed enormously to possibilities to study the human brain. The human brain is a very complex structure and studying it is a challenge for many people, including students. Suppose that some researchers have developed a new supportive tool to actually help medical students in their study of the anatomy of the human brain. The human brain is a large three-dimensional structure, which is very difficult to represent in two-dimensional pictures. Two-dimensional pictures of (parts of) the brain are already complex study objects, but having to mentally build a three-dimensional picture of the brain or of a brain region is something that can require students to invest a high mental effort. The researchers have developed the new supportive tool to provide specific guidance to students while learning about the anatomy of the human brain. It is therefore expected to reduce the mental effort that must be invested by the students.

To study the effectiveness of this new tool, the researchers assign 100 students to either the experimental treatment condition (n = 50) or control condition (n = 50). Students in the control condition perform five learning tasks in logical order in a traditional way, while students in the experimental treatment condition perform the same learning tasks in the same order with help of the new supportive tool. In both conditions, students rate their mental effort on a visual analog scale from 0 (minimum) to 100 (maximum) after each of five tasks. The researchers then aggregate these five repeated measurements to one average rating per student and compare the two groups in terms of average rating. Since the new tool, used in the experimental treatment condition, is expected to reduce mental effort invested by the students, the researchers expect the average rating to be significantly lower in the experimental treatment condition than in the control condition.

Split plots

The type of study design in the aforementioned example is also known as split-plot design. This term stems from agricultural experiments in which split plots of land received different treatments and were monitored or measured across time (Fisher, 1925). Likewise, in the current example study, students are allocated to different treatment conditions and are measured repeatedly on the same variable of interest (Howell, 2012), more specifically, five times on mental effort.

The approach chosen by the hypothetical researchers in the example study constrains them to be satisfied with a comparison of the difference between the two conditions in rating averaged over the five tasks. However, what if:

- use of the supportive tool helps to reduce mental effort only after a number of tasks, meaning that it does not affect mental effort in the first couple of tasks;
- use of the supportive tool helps to reduce mental effort in the first couple of tasks (i.e., an initial stage) but its effect decreases and eventually vanishes;
- use of the supportive tool leads to an elevated mental effort in the first couple of tasks but decreases mental effort in subsequent tasks; or
- use of the supportive tool results in a decreased mental effort in the first couple of tasks but elevates mental effort in subsequent tasks?

These four scenarios point at a so-called interaction between condition and task: the effect of condition (i.e., supportive tool) depends on (i.e., is moderated by) task or, in other words, the effect of condition is not constant across tasks. Each of these four scenarios can have important implications for self-regulated learning and/or guidance needed in (self-regulated) learning, yet none of these four interaction scenarios can be studied by the aggregation approach. In fact, the aggregation approach ignores all possible interactions between condition and task (i.e., measurement occasion) and assumes (among others) that the effect of condition is the same across tasks, we may find no (significant) differences between conditions, while a repeated measurements approach may indicate how the effect of condition depends on task.

A not rarely heard argument from experimental researchers is that we do not need to test for the aforementioned interaction if we do not have a hypothesis about it beforehand or that there is only a need for that if a hypothesis about a particular main effect (in this context: a difference between conditions across tasks or time) is not supported by the data. However, interpretations of main effects (and interpretations of interaction effects likewise) are meaningful only if crucial assumptions underlying the tools used for analysis are met.

Non-interaction is such an assumption, even in for instance analysis of covariance (ANCOVA) or factorial analysis of variance (ANOVA) on single-measurement data when reporting main effects (e.g., Field, 2013; Howell, 2012; Huitema, 2011). This and other assumptions (e.g., normality, homogeneity of variance, independence of observations) should be examined not only if study results seem to not provide support for hypotheses but also if the results at first appear to provide such support. One of the potential outcomes of such a critical assessment of assumptions in the latter may be that a powerful conclusion about support for a hypothesis cannot be made given that non-interaction (or another important assumption) is not met and must therefore be accounted for through some change(s) in the statistical model. One of the problems with the aggregation approach is that non-interaction is automatically assumed and cannot be tested. This is a problem because condition by time or task interaction is quite common (Howell, 2012; Leppink, 2015; Tan, 2008), also in cognitive load research (Leppink, et al., 2013), and implies that what we have thought of as a main effect (i.e., the difference between conditions being constant across time or tasks) may vary quite a bit across time or tasks.

In the case of a so-called “ordinal” interaction (i.e., non-parallel lines that do not cross), main effects can strictly speaking be interpreted. If for instance an experimental treatment group and a control group are measured in terms of task performance three times, the average change across measurement occasions is different for the two groups but the average task performance is higher in one and the same group at each of the three occasions, it is fair to speak of a main effect of group even if the interaction (i.e., difference in change between groups) is statistically significant. However, even in such a case, a significant interaction term comprises information that is not present in main effects. If for instance the experimental treatment administered prior to the three measurements needs some time to have an effect on task performance, we may well see that the groups do not yet differ significantly on the first occasion but that group differences are larger and statistically significant at subsequent measurement occasions. Main effects do not comprise that information. Yet, when aggregating repeated measurements to one single score, the only thing we can interpret is a main effect. Finally, in the case of disordinal interactions, that is: when lines do cross, one should not interpret main effects at all (Howell, 2012), because which group performs better depends on another variable, here measurement occasion. This paper discusses an example of a disordinal interaction: the experimental treatment group starts off with slightly higher average mental effort (albeit it non-significantly) but at some point the difference reverses to have a significantly lower mental effort in the experimental group on the final two occasions.

Internal consistency vs. test-retest reliability

When interpreting aggregated scores, researchers sometimes refer to a "satisfactory" or "high" internal consistency of the scores on the tasks aggregated (e.g., Ayres, 2006; Corbalan, et al., 2008; Hoffman, 2012; Paas, 1992; Paas et al., 2003; Paas & Van Merriënboer, 1994b). Internal consistency is largely about the extent to which items assumed to measure the same characteristic yield similar scores. Hordes of researchers compute and report Cronbach's alpha coefficient (Cortina, 1993; Cronbach, 1951) as an estimator of internal consistency, which is about the extent to which items assumed to measure the same characteristic yield similar scores. This approach of computing internal consistency (although Cronbach's alpha is not always the best choice; Peters, 2014) and using factor scores through aggregation or through factor analysis generally makes good sense for multiple items measuring the same characteristic at one single point of time (e.g., three items for intrinsic or extraneous cognitive load, Leppink et al., 2013; 2014). However, applying that approach to single-item repeated measurements is a tricky enterprise. The reason for the latter is that, in a context like the one in the example study, two sources of variance resonate and are perfectly confounded in any estimate of internal consistency: variance due to error in the (here: mental effort) rating and variance due to task differences. In other words, we cannot know to what extent a Cronbach's alpha value reflects (a lack of) reliability in measurement in this specific context or variation due to differences between tasks. Further, in the following, it becomes clear that even a high Cronbach's alpha value does not imply absence of any of the aforementioned types of interaction.

A more appropriate perspective on reliability in studies of this kind is that of test-retest reliability, as estimates of the latter provide an indication of the extent to which a measure provides consistent results for the same participants at different measurement occasions. One would expect that participants who experience above average mental effort at the first measurement occasion are to some extent also more likely to experience above average mental effort at subsequent measurement occasions, especially in studies where the kind of treatment (either experimental or control group) is the same across a range of tasks of the same kind.

Method

The remainder of this paper demonstrates the aforementioned problem of ignoring potential condition by task interaction by aggregating repeated measurements in a randomized experiment and presents two comprehensive alternatives accounting for the repeated measurements. For educational purposes, data from this example study were simulated, and a detailed overview of the simulation procedure is available from the authors.

The advantage of a simulation study is that the outcomes of the study are known and as such it enables a comparison of strengths and weaknesses of various methods of analysis. The data were simulated such that, at the level of population of (possible) students, the supportive tool has no effect on mental effort for the first three tasks but results in a substantially lower mental effort compared to the control condition at subsequent tasks (in this case: the fourth and fifth task). Software programs such as SAS, SPSS, STATA, Mplus, and R provide facilities for dealing with repeated measurements data as done in the remainder of this paper. Since among educational researchers SPSS appears to be used much more than any of the other programs mentioned, it is perhaps interesting to mention that SPSS v21 was used for analysis in this case.

Approach (1): Aggregation

As mentioned previously, students in both conditions rate their mental effort on a visual analog scale from 0 to 100 after each of five tasks. The researchers aggregate these five repeated measurements to one average rating per student, and compare the two conditions in terms of average rating, expecting the average rating to be significantly lower in the experimental treatment condition. This comes down to a two-sample t-test or one-way *ANOVA* on the aggregated scores.

Approach (2): Split-plot ANOVA

Split-plot ANOVA – which is in fact a repeated measures ANOVA that includes at least one between-participants factor (e.g., treatment in this example study) – is a well-known tool among experimental psychologists (Howell, 2012), who for instance compared two or more conditions in a pretest posttest (and follow-up) control-group design in one study or another, the difference between conditions being a different treatment (or, in the case of the control condition, no treatment at all) between pretest and posttest. Split-plot ANOVA provides a valid tool for analysis if there is no missing data due to some students omitting one rating or another and certain assumptions with regard to variances of scores on tasks and correlations between task scores are realistic. If there is no missing data, split-plot ANOVA works fine if the variance of scores is not too different across tasks and not too different between conditions and if the correlation between scores is not that much different across pairs of tasks. Otherwise, the third approach – which is more flexible – may provide a slightly better alternative.

Approach (3): Two-level regression

In fact, split-plot ANOVA is a special type of a two-level regression model or, also called, mixed-effects model. The term mixed effects is used whenever an analysis involves one or more fixed effects and one or more random effects. The purpose of a study like this one is to generalize findings to a larger population of (possible) students, and we assume that the students in our study form a random sample from a population that has a particular and preferably Normal (i.e., bell-shaped) distribution. In other words, we treat "student" as a random effect. We can estimate student-specific random effects because we have repeated measurements from the same students. Treatment, however, is a fixed effect; we are interested in the specific comparison of experimental treatment and control condition, and we do not consider these two treatments as a random sample of a universe of possible treatments to which we generalize. Likewise, the repeated measurements are in this context treated as fixed effect; the interest lies in differences between conditions on these specific tasks and, to some extent, in differences between tasks as well. In the type of two-level regression we are discussing here, student and measurement occasion (1, 2, 3, 4, 5) are treated as hierarchical levels with student being the upper level. Split-plot ANOVA is a special type of two-level regression model in that it is somewhat restrictive in assumptions with regard to variances and correlations as mentioned before, assumptions that can be relaxed in a more flexible two-level regression approach as is demonstrated in the last paragraph of the Results section.

Results

Table 1 presents descriptive statistics per task per condition, and Figure 2 provides a graphical representation of the mean rating per task per condition.

Table 1. Mean (and *SD*) per rating per condition (based on simulated data)

Rating (0-100)	Experimental treatment	Control
First	60.80 (5.00)	59.72 (5.28)
Second	59.38 (6.36)	57.74 (6.65)
Third	55.78 (6.91)	55.38 (7.71)
Fourth	46.36 (8.50)	52.82 (8.53)
Fifth	40.12 (10.94)	48.62 (9.95)

Table 1 and Figure 2 indicate that the average mental effort is slightly higher in the experimental treatment condition in the initial stage (i.e., the first three tasks) but then drops and is lower than in the control condition after the fourth and fifth task.

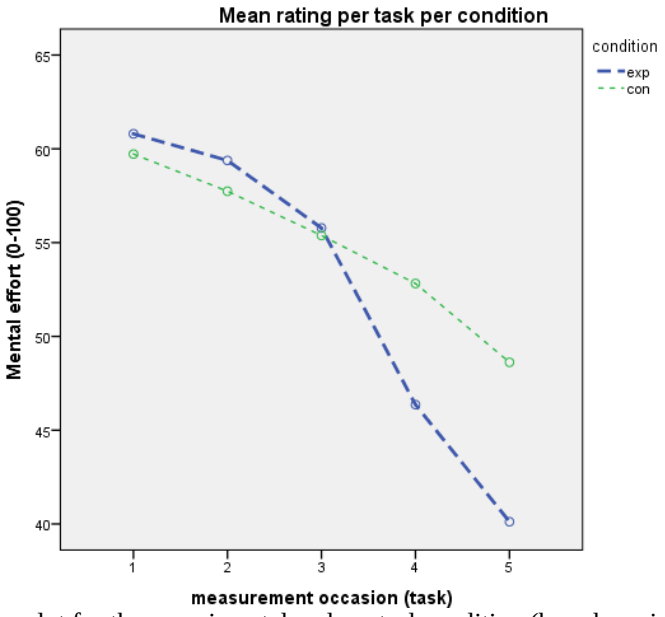

Figure 2. Means plot for the experimental and control condition (based on simulated data)

Approach (1): A**ggregation**

Researchers following the commonly encountered aggregation approach typically omit the step as done in Table 1 and Figure 2. Instead, they report that the five ratings yield a Cronbach's alpha of 0.85 for the two conditions together and even slightly higher when computed for experimental treatment condition (0.87) and control condition (0.87) separately. These high Cronbach's alpha values are then taken as justification for aggregating ratings obtained after different tasks to one average rating to then proceed as follows: the average rating in the experimental treatment condition of 52.51 (*SD* = 6.35) is slightly lower than the average rating of 54.88 (*SD* = 6.34) in the control condition. However, the difference is rather small, η^2 = 0.034 (values around 0.01, 0.06, and 0.14 indicate small, medium, and large effects, respectively), and not statistically significant at the conventional 0.05 significance level, $F(1, 98)$ = 3.493, p = 0.065. What follows in the case of such an outcome is an attempt to find an explanation for the unexpected result and/or a plea to replicate the study in some slightly different context, without considering the possibility of treatment-by-task interaction.

Approach (2): S**plit-plot ANOVA**

Table 1 indicates that for a given task, the two conditions have similar standard deviations. Levene's test per task reveals that the standard deviations of the two conditions are not statistically different in any of the tasks (with *p*-values ranging from 0.377 for the third task to 0.907 for the fourth task). Further, Box's test reveals that the covariance matrix for the five repeated measurements is not significantly different between the two conditions (p = 0.863). Finally, Mauchly's test indicates that the standard error for pairwise comparisons differs significantly across tasks ($p < 0.001$), which is not a surprise because Table 1 indicates that in both conditions the standard deviation around the mean rating is about twice as large in the fifth task than in the first task. We can to some extent correct for the latter by using an adjusted test, here Huynh-Feldt (Field, 2013), which in this specific case adjusts all degrees of freedom (*df*) by a factor 0.500 meaning that all degrees of freedom become twice as small. While this adjustment is somewhat conservative in that it comes at the cost of loss of some statistical power for detecting effects of interest (i.e., somewhat increased probability of Type II error), the adjustment provides some protection against an inflation of the probability of finding significant differences while in fact there is no effect (i.e., Type I error probability). An alternative is to use one of the multivariate tests (e.g., Pillai's Trace), which do not require (to adjust for departure from) equal standard errors.

The first thing we test is the treatment-by-task or split-plot interaction: $F(2.001, 196.142) = 21.889$, $p < 0.001$, $\eta^2 = 0.183$. It turns out that the non-parallel lines (i.e., interaction) displayed by (Table 1 and) Figure 1 is statistically significant and quite large. What we can do now is so-called simple effects analysis, that is: we perform a two-sample *t*-test or one-way ANOVA for the difference in average rating between conditions per task. If the interaction is not statistically significant, we do not need to perform simple effects analysis for then the difference in average rating between conditions does not differ significantly across tasks. Split-plot ANOVA provides the same test outcome as found in the aggregation approach for the difference in average rating over all five tasks together, and it is – provided that other assumptions are met, more about this follows when presenting the third approach – then interpretable. However, in the case of significant interaction, as is the case right now, it makes much more sense to inspect the difference between conditions per task.

Simple effects analysis reveals no significant difference between conditions on the first task, $F(1, 98) = 1.103$, $p = 0.296$ ($\eta^2 = 0.011$), no significant difference on the second task, $F(1, 98) = 1.587$, $p = 0.211$ ($\eta^2 = 0.016$), no significant difference on the third task, $F(1, 98) = 0.075$, $p = 0.785$ ($\eta^2 = 0.001$), a statistically significant difference on the fourth task, $F(1, 98) = 14.382$, $p < 0.001$ ($\eta^2 = 0.128$), and a statistically significant difference on the fifth task, $F(1, 98) = 16.532$, $p < 0.001$ ($\eta^2 = 0.144$). In other words, differences between conditions appear small and rather meaningless on the first three tasks, whereas after the fourth and fifth task the average mental effort is quite a bit lower in the experimental treatment condition. Finally, one note of caution on the reported η^2-values: even though in articles and in some statistical software output the term eta-squared or η^2 is used, the η^2-values reported are in fact partial eta-squared values (Field, 2013), because they reflect the proportion of variance explained by the effect under consideration that is not explained by other effects in the model (i.e., the unique contribution of the effect under consideration).

Approach (3): Two-level regression

It is clear that the split-plot ANOVA approach is more appropriate than the aggregation approach. However, split-plot ANOVA is quite restrictive in terms of correlations between repeated measurements: they should be more or less equal. In randomized experiments in which tasks are counterbalanced such that the tasks are provided to different participants in different order, that assumption may be feasible. In the current experiment, one and the same logical order is used. Further, split-plot ANOVA tends to be somewhat less feasible if there are considerable differences in standard deviation across tasks. Table 1 indicates that this is the case; the standard deviation increases throughout and is about twice as large after the fifth task than after the first task. A more flexible two-level regression approach allows for taking eventual departures from equal correlations and eventual departures from equal standard deviations into account. Elaborating on all possible ways to do so could easily fill a full paper on itself (Tan, 2008; Verbeke & Molenberghs, 2000). In the remainder of this section, we focus on the most likely candidate models for this specific context.

Given that the tasks are provided in one specific order (because this specific order appears logical in this study context), it is to be expected that ratings on subsequent tasks are more correlated than ratings from tasks further away from each other. In other words, task pairs 1-2, 2-3, 3-4, and 4-5 (i.e., nearest task pairs) can be expected to have the highest correlation. The correlation for task pairs 1-3, 2-4, and 3-5 (i.e., second-nearest task pairs) could be somewhat lower than for the aforementioned four task pairs but still somewhat higher than the correlation for task pairs 1-4 and 2-5 (i.e., third-nearest task pairs). Finally, task pair 1-5 (i.e., fourth-nearest task pair) is to be expected to yield the lowest correlation, because of the largest distance between the first and fifth task. If the four nearest task pairs have (more or less) one common correlation, the three second-nearest task pairs have (more or less) one common correlation (that is different from that of the nearest task pairs), and the two third-nearest task pairs have (more or less) one common correlation (that is different from the correlation of nearest task pairs and different from the correlation of the second-nearest task pairs), we are dealing with a so-called Toeplitz structure (Bareiss, 1969; Leppink, et al., 2013). A more restrictive version of this structure applies when the nearest task pairs have correlation *r*, the second-nearest task pairs have correlation r^2 (i.e., *r* times *r*) the third-nearest task pairs have correlation r^3 (i.e., *r* times *r* times *r*), and the final task pair (i.e., 1-5) has a correlation of r^4 (i.e., *r* times *r* times *r* times *r*). This is a so-called first-order autoregressive (AR1) structure. Given that the correlation between any pair of tasks (including nearest tasks) is usually smaller than one, AR1 implies that there is a predictable decline in correlations with increasing space between tasks. AR1 is a likely candidate in for instance growth studies (Tan, 2008). If AR1 fits the data well, its advantage over Toeplitz is that we have a more parsimonious model, that is: we use fewer degrees of

freedom for accounting for this correlation structure and keep these degrees of freedom for a more powerful test on effects of interest (here: treatment). However, if AR1 turns out too restrictive, Toeplitz is a common alternative candidate.

Commonly used criteria for comparing the fit of AR1, Toeplitz and other structures are Akaike's information criterion (AIC) (Akaike, 1973), Schwarz's Bayesian information criterion (BIC) (Schwarz, 1978) and the deviance information criterion (DIC) (Spiegelhalter, Best, Carlin, & Van der Linde, 2002). All these criteria have in common that they indicate the extent to which the model deviates from the data, and smaller values mean less deviance or, in other words, better fit. The (relative) parsimony of models under consideration is taken into account, since all these criteria penalize for the number of degrees of freedom consumed by the models under consideration. Further, the BIC has a penalty for the sample size (*N*). In the context of repeated measurements data, a danger of the latter is that *N* can be defined in different ways since it can refer to observations at different levels (Fitzmaurice, Laird, & Ware, 2004; Hedeker & Gibbons, 2006; Weakliem, 1999). Since this distinction is not made in the BIC, the BIC may sometimes favor (too) simple models. Finally, although there is some indication that the DIC can yield good performance in the context of multilevel analysis (Gelman & Hill, 2007), this criterion still needs more research and for now it (perhaps for that reason) is not included in some standard statistical packages like SPSS.

Let us compare six models in terms of the AIC:

- Model 1: all task pairs have the same correlation (i.e., compound symmetry) and the standard deviation is equal across tasks. This is more or less conform traditional split-plot ANOVA;
- Model 2: all task pairs have the same correlation (i.e., compound symmetry) but the standard deviation is not equal across tasks;
- Model 3: AR1 structure and the standard deviation is equal across tasks;
- Model 4: AR1 structure and the standard deviation is not equal across tasks;
- Model 5: Toeplitz structure and the standard deviation is equal across tasks; and
- Model 6: Toeplitz structure and the standard deviation is not equal across tasks.

Table 2. Outcomes of Model 6 (Toeplitz and unequal standard deviations) (based on simulated data)

Fixed effect		*B* (*SE*)	*df*	*t*-value	*p*-value	Confidence interval (95%)	
						Lower	Upper
Task 1		59.72 (0.83)	96.589	71.923	< 0.001	58.07	61.37
Task 2		57.74 (1.03)	103.263	56.010	< 0.001	55.70	59.78
Task 3		55.38 (1.03)	111.801	53.778	< 0.001	53.34	57.42
Task 4		52.82 (1.09)	128.729	48.245	< 0.001	50.65	54.99
Task 5		48.62 (1.33)	134.968	36.438	< 0.001	45.98	51.26
Task 1 *by* Treat [1]		1.08 (1.17)	96.589	0.920	0.360	-1.25	3.41
Task 2 *by* Treat		1.64 (1.46)	103.263	1.125	0.263	-1.25	4.53
Task 3 *by* Treat		0.40 (1.46)	111.801	0.275	0.784	-2.49	3.29
Task 4 *by* Treat		-6.46 (1.55)	128.729	-4.172	< 0.001	-9.52	-3.40
Task 5 *by* Treat		-8.50 (1.89)	134.968	-4.504	< 0.001	-12.23	-4.77
Random effect			Variance (*SE*)	Wald *z*	*p*-value	Confidence interval (95%)	
						Lower	Upper
Variance [2]	Task 1		34.47 (4.96)	6.949	< 0.001	26.00	45.71
	Task 2		53.14 (7.39)	7.185	< 0.001	40.45	69.80
	Task 3		53.02 (7.09)	7.477	< 0.001	40.80	68.92
	Task 4		59.93 (7.47)	8.023	< 0.001	46.94	76.52
	Task 5		89.02 (10.84)	8.215	< 0.001	70.13	113.01
Correlation	Nearest		0.79 (0.02)	34.861	< 0.001	0.74	0.83
	Second-nearest		0.58 (0.05)	12.882	< 0.001	0.49	0.66
	Third-nearest		0.38 (0.07)	5.522	< 0.001	0.24	0.51
	Task pair 1-5		0.20 (0.09)	2.149	0.032	0.01	0.38

Note. [1] Treatment is coded as follows: 0 = control condition, 1 = experimental treatment condition; [2] The variance is the standard deviation squared.

The AIC values are 3226.018 (Model 1), 3167.441 (Model 2), 3054.532 (Model 3), 3022.713 (Model 4), 3049.593 (Model 5), and 3021.908 (Model 6). Generally, the models that account for the fact that (as indicated in Table 1) the standard deviations vary quite a bit across tasks (i.e., Model 2, Model 4, and Model 6) perform better than the models that assume the standard deviation to be equal across tasks (i.e., Model 1, Model 3, and Model 5). Model 4 and Model 6 perform best, indicating that the assumption of the correlation being (more or less) equal across tasks is not realistic, and there appears to be a slight preference for the latter. Table 2 therefore presents the outcomes of Model 6.

Given the coding (i.e., 0 = control condition, 1 = experimental treatment condition), Task 1 represents the average rating after the first task in the control condition, and Task 1 by Treat represents the difference in average rating after the first task between the two conditions. In other words, the average rating is slightly higher in the experimental treatment condition after the first task, as can be seen in Table 1 and Figure 1. Task 2 represents average rating after the second task in the control condition, and Task 2 by Treat represents the difference in average rating after the second task between the two conditions. The same interpretation holds for the other three tasks, so Task 5 represents the average rating after the fifth task in the control condition while Task 5 by Treat represents the difference in average rating after the fifth task between the two conditions. In other words, we see that the two conditions do not really diverge in the first three tasks but on the fourth and fifth task, you see a clear difference between the two conditions. Since we view a decrease in mental effort as something positive in this context, the difference between the two conditions after the fourth and fifth task is in favor of the experimental treatment condition and thus in favor of the supportive tool.

Discussion

Some readers, especially the ones rooted in split-plot ANOVA, may wonder why go for a more complex two-level model if split-plot ANOVA leads to the same conclusion with regard to the treatment of interest. Indeed, both approaches beat the aggregation approach for an obvious reason: instead of wiping out all within-student variance, we account for it and conclude that the treatment has the expected effect only after a number of tasks. While the aggregation approach results in the mere conclusion that the supportive tool has a rather small and not statistically significant effect, both alternatives indicate that there is little to no reason to assume any treatment effect for the first three tasks but after that the expected positive effect of the supportive tool becomes evident.

If there is no missing data and departures from equal correlations and equal standard deviations are not that large, split-plot ANOVA generally provides a quick and efficient way to draw valid conclusions and present the findings. However, in the case of more severe departures from equal correlations and equal standard deviations, split-plot ANOVA tends to use inappropriate standard errors for particular hypothesis tests. Moreover, in the case of missing data, split-plot ANOVA tends to produce biased estimates, because participants who accidentally omit one rating or another are deleted completely in the split-plot ANOVA approach even if four of the five ratings are available.

Repeated measurements and required sample size

Treating repeated measurements data as such also has benefits for the statistical power or probability of detecting an effect in a random sample if that effect exists at population level. For instance, using G*Power (Buchner, Erdfelder, Faul, & Lang, 2009) – a program that enables researchers to calculate among others statistical power given statistical significance level, sample size, expected effect size, and for more complex studies some other factors or the required sample size given a desired statistical power, statistical significance level, sample size, and expected effect size – reminds us that, given a medium size effect and $\alpha = 0.05$ in a standard two-tailed test, to obtain a statistical power of 0.80 in a two-group comparison of means from a single measurement occasion, we need a sample size of $n = 64$ per group. Making the same kind of main effects comparison when treating two repeated measurements with a correlation of 0.50 (which is often reasonable: Hedeker & Gibbons, 2006) as such instead of aggregating these to a single mean score per group, we need a total sample size of $n = 49$ per group for that same statistical power. Further, when dealing with five measurements with an average between-measurements correlation of 0.50 (as in our example study), the sample size required for a power of 0.80 is only about $n = 39$ per group. To obtain a power of 0.80 for the

group-by-measurement interaction in such a study including five measurements, we need only about $n = 11$ per group.

Learning groups

Another type of study in which the validity of split-plot ANOVA is limited, is when individual participants are nested within, for instance, learning groups in which participants interact and cooperate for a particular period of time. Just like repeated measurements typically induce a within-participant between-measurements correlation, participants interacting and cooperating in learning groups tends to induce a within-group between-participants correlation. In the example study in this paper, students receive treatments individually. If treatment is applied at the level of small (e.g., problem-based) learning groups, it is recommendable to account for that in the analysis stage. With ANOVA we cannot do that, with multilevel analysis we can (Snijders & Bosker, 2011). For power or required sample size calculations it is always better to consult a statistician in such a case, as sample sizes at both the level of learning groups and the level of participants within learning groups depend on the magnitude of both fixed and learning-group-level random effects and it is therefore hard to give general guidelines that hold for most cases, except for the following.

The number of learning groups is more important than large numbers of individuals per learning group (Hox, 2010). Current-day software can provide reasonably accurate estimates for learning-group level variances when including as few as six to twelve learning groups (Browne & Draper, 2000), but of course estimates and their standard errors become more accurate with increasing sample sizes at both levels (i.e., learning groups and individuals). When interested mainly in experimental treatment effects, dealing with learning groups of size fifteen or smaller may be no problem (e.g., thirty learning groups of fifteen students each; Leppink, 2015), and when learning group effects are on the large side, multilevel analysis may be feasible even if learning groups comprise as few as two learners only (i.e., study in pairs; Leppink, Broers, Imbos, Van der Vleuten, & Berger, 2012). However, when dealing with effects of smaller size or interested in cross-level interactions, one may want to consider a 30/30 or 50/20 rule of thumb (Hox, 2010), where the first number refers to the number of groups and the latter to the number of individuals per group.

Departures from assumptions

Of course, any statistical model is based on assumptions. Some assumptions may be more realistic than others, and – in terms of impact of assumption departures on model outcomes – some assumptions are more important than others. For instance, most tests involving comparisons of means are fairly robust against some skewness. The assumption of homogeneity of variances is an important one for ANOVA models but can be relaxed in more flexible regression models. Further, in studies in which participants are nested within learning groups and also in quite some studies involving repeated measurements, random intercepts and/or random slopes are specified. These random intercepts and slopes are learning-group-level effects in the case participants are nested within learning groups, participant-level effects if individual participants are measured repeatedly, and in studies where individual participants nested within learning groups are measured repeatedly we expectedly have a combination of both (Leppink, 2015). The assumption typically made with regard to these random intercepts and slopes is that they are normally distributed around the fixed intercept and slope, respectively. For the kind of piecewise linear models used in this paper, some departure from these normality assumptions is generally not a serious problem, but for non-linear models and categorical response variable models things are trickier (McCulloch & Neuhaus, 2011). This leads to the final assumption, namely that – in the models discussed in this paper – we assume the response variables of interest to be of interval or ratio level of measurement. When dealing with learning-group nesting or repeated-measurements studies in which response variables of interest are binary or ordinal, generalized linear mixed models or generalized estimating equations (Hedeker & Gibbons, 2006; Molenberghs & Verbeke, 2005) should be used, and the latter also offer an alternative if interval/ratio level response variables show more severe departures from normality or homogeneity of variances. Finally, when dealing with a response variable that is a count (e.g., the number of typing errors in a task), a special case of the generalized linear mixed model, the Poisson model, can be used (Hox, 2010).

Conclusion

The message that we need to treat repeated measurement scores as they are, namely as scores obtained from the same participants measured repeatedly, applies to any study in educational research that includes repeated measurements data. No matter what level of self-regulated learning we are talking about – item level, task / topic level or task sequence level – we run a risk of drawing inappropriate conclusions whenever we aggregate repeated measurements data. This holds not only for mental effort or for the specific context of self-regulated learning. The message applies to any study that deals with the repeated measurement of performance, motivation, satisfaction, pain or other variable of interest. As Figure 1 indicates, caution with aggregation is also needed when examining relations between quantitative variables. This distinction also has relevance in experiments in the context of self-regulated learning, in which students select and perform a series of for instance eight tasks of either of five possible difficulty levels (1, 2, 3, 4, 5), rate their mental effort after each task, and receive instructional support (i.e., experimental treatment condition) or do not receive instructional support (i.e., control condition) in this process of task selection. The instructional support may then affect not only performance and mental effort rating, but even on which tasks the performance and mental ratings are based. We may see considerable differences between as well as within conditions (between and within students) in terms of performance, mental effort rating, and task selection pattern. The current practice is to aggregate performance across tasks, aggregate mental effort ratings across tasks, and aggregate difficulty level across tasks to some kind of "average" difficulty, perform ANOVA or regression on aggregated scores while ignoring all meaningful information in the repeated measurements. To estimate the effects of treatment, task difficulty level, and perhaps other variables of interest on performance, mental effort, intrinsic or extraneous cognitive load, we need multilevel analysis.

The core message of this paper is that the approach of aggregating repeated measurements data to single (sum or) average scores can distort our view of effects and relations of interest to a substantial extent and should therefore not be used in cognitive load research. Multilevel analysis of repeated measurements data that can account for various features of the data constitutes a best practice.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Academiai Kiado.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic load within problems. *Learning and Instruction*, *16*, 389-400. doi:10.1016/j.learninstruc.2006.09.001

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, *96*, 523-535. doi:10.1037/0022-0663.96.3.523

Azevedo, R., Moos, D., Greene, J., Winters, F., & Cromley, J. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with multimedia?. *Educational Technology, Research, and Development*, *56*, 45-72. doi:10.1007/s11423-007-9067-0

Bareiss, E. H. (1969). Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices. *Numerische Mathematik*, *13*, 404-424.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Buchner, A., Erdfelder, E., Faul, F., & Lang, A. G. (2009). G*Power: Statistical Power Analyses for Windows and Mac, G*Power version 3.1.2 [Software]. Retrieved from http://www.gpower.hhu.de/

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load?. *Computers in Human Behavior*, *25*, 315-324. doi:10.1016/j.chb.2008.12.020

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2008). Selecting learning tasks: effects of adaptation and shared control. *Contemporary Educational Psychology*, *33*, 733-756. doi:10.1016/j.cedpsych.2008.02.003

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, *78*, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83-87. doi:10.1111/1467-8721.01235

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, UK: Sage.

Fisher, R. A. (1925). *Statistical methods for research and workers*. Edinburgh, Scotland: Oliver and Boyd.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York, NY: Wiley.

Flavell, J. (1979). Metacognition and cognitive monitoring: A New area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906-911. doi:10.1037/0003-066X.34.10.906

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York, NY: John Wiley & Sons.

Hoffman, B. (2012). Cognitive efficiency: A Conceptual and methodological comparison. *Learning and Instruction*, *22*, 133-144. doi:10.1016/j.learninstruc.2011.09.001

Howell, D. C. (2012). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Learning.

Hox, J. (2010). *Multilevel analysis: techniques and applications* (2nd ed.). New York, NY: Taylor & Francis Group.

Huitema, B. E. (2011). *The Analysis of covariance and alternatives: statistical methods for experiments, quasi-experiments, and single-case studies* (2nd ed.). New York, NY: Wiley.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational Psychology Review*, *23*, 1-19. doi:10.1007/s10648-010-9150-7

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1624-1637. doi:10.1037/xlm0000009

Kostons, D., Van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education, 54*, 932-940. doi:10.1016/j.compedu.2009.09.025

Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A Cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*, 121-132. doi:10.1016/j.learninstruc.2011.08.004

Leppink, J. (2015). Data analysis in medical education research: A Multilevel perspective. *Perspectives on Medical Education*, *4*, 14-24. doi:10.1007/s40037-015-0160-5

Leppink, J., Broers, N. J., Imbos, T., Van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Prior knowledge moderates instructional effects on conceptual understanding of statistics. *Educational Research and Evaluation*, *18*, 37-51. doi:10.1080/13803611.2011.640873

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*, 1058-1072. doi:10.3758/s13428-013-0334-1

Leppink, J., Paas, F., Van Gog, T., Van der Vleuten, C. P. M., & Van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32-42. doi:10.1016/j.learninstruc.2013.12.001

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical science*, *26*, 388-402. doi:10.1214/11-STS361

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*, 159-163. doi:10.1111/j.1467-8721.2009.01628.x

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York, NY: Springer.

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 125-173). San Diego, CA: Academic Press.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A Cognitive load approach. *Journal of Educational Psychology*, *84*, 429-434. doi:10.1037/0022-0663.84.4.429

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*, 63-71. doi:10.1207/S15326985EP3801_8

Paas, F., & Van Merriënboer, J. J. G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*, 351-371. doi:10.1007/BF02213420

Paas, F., & Van Merriënboer, J. J. G. (1994b). Variability of worked examples and transfer of geometrical problem solving skills: A Cognitive-load approach. *Journal of Educational Psychology*, *86*, 122–133. doi:10.1037/0022-0663.86.1.122

Peters, G. J. Y. (2014). The Alpha and the omega of scale reliability and validity. *European Health Psychologist*, *16*, 56-69.

Salden, R. J. C. M., Paas, F., & Van Merriënboer, J. J. G. (2006). Personalised adaptive task selection in air traffic control: Effects on training efficiency and transfer. *Learning and Instruction*, *16*, 350-362. doi:10.1016/j.learninstruc.2006.07.007

Salomon, G. (1984). Television is "easy" and print is "tough": The Differential investment of mental effort in learning as a function of perceptions and attributes. *Journal of Educational Psychology*, *78*, 647-658.

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*, 93-114. doi:10.1007/s11251-014-9328-3

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, *6*, 461-464.

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An Introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583-639. doi:10.1111/1467-9868.00353

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295-312. doi:10.1016/0959-4752(94)90003-5

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*, 122-138. doi:10.1007/s10648-010-9128-5

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*, 185-223. doi:10.1207/s1532690xci1203_1

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology*, *119*, 176-192. doi:10.1037/0096-3445.119.2.176

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296. doi:10.1023/A:1022193728205

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer.

Tan, F. E. S. (2008). Best practices in analysis of longitudinal data: A Multilevel approach. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 451-470). London, UK: Sage.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study. *Journal of Experimental Psychology*, *25*, 1024-1037. doi:10.1037/0278-7393.25.4.1024

Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*, 833-839. doi:10.1002/acp.2883

Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., & Van Merriënboer, J. J. G. (2013). The Effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning*, *8*, 173-191. doi:10.1007/s11409-013-9100-0

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Weakliem, D. L. (1999). A Critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, *27*, 359-397. doi:10.1177/0049124199027003002

Zimmerman, B. J., & Schunk, D. H. (Eds.) (2001). *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed.). Mahwah, NJ: Erlbaum.