Tutorials in Quantitative Methods for Psychology
2008, vol. 4(1), p. 21-34.

DOI: 10.20982/tqmp.04.1.p021

Eliminating Aggregation Bias in Experimental Research: Random Coefficient Analysis as an Alternative to Performing a 'by-subjects' and/or 'by-items' ANOVA

Glenn L. Thompson
University of Ottawa

Experimental psychologists routinely simplify the structure of their data by computing means for each experimental condition so that the basic assumptions of regression/ANOVA are satisfied. Typically, these means represent the performance (e.g. reaction time or RT) of a participant over several items that share some target characteristic (e.g. Mean RT for high-frequency words). Regrettably, analyses based on such aggregated data are biased toward rejection of the null hypothesis, inflating Type-I error beyond the nominal level. A preferable strategy for analyzing such data is random coefficient analysis (RCA), which can be performed using a simple method proposed by Lorch & Myers (1990). An easy to use SPSS implementation of this method is presented using a concrete example. In addition, a technique for evaluating the magnitude of potential aggregation bias in a dataset is demonstrated. Finally, suggestions are offered concerning the reporting of RCA results in empirical articles.

Researchers routinely transform their data in order to satisfy the assumptions of statistical analyses (e.g. regression analysis). For example, log, reciprocal, and square-root transformations are all used to correct the shape of empirical distributions so that the assumption of normality (Gaussian distribution) is satisfied. Such considerations also guide how raw data is prepared for analysis. For instance, the regression/ANOVA approach that is taught in undergraduate and graduate-level statistics classes requires that each data point be independently collected or at least uncorrelated. In practical terms, this means that each participant must contribute a single data-point to an analysis. This assumption is violated in many situations such as when a repeated measures design is used. The solution in the special case of repeated measures is to extract the offending 'correlated' between-subject variance prior to analysis (e.g. repeated-measures ANOVA; correlated t-test). A limitation of this strategy is that it is only applicable to cases where researchers are interested in comparing the repeated observations or matched observations. This approach is not applicable to cases where participants generate a number of responses that is greater than the number of experimental conditions. Often, this is the case when the independent variable (IV) is the property of items or a collection of items, such as a comparison of reaction times (RTs) in response to high- and low-frequency words.

Glenn L. Thompson, School of Psychology, University of Ottawa, Ottawa, Ontario, Canada.

This work was supported by two Natural Sciences and Engineering Council of Canada (NSERC) scholarships. I would like to thank my doctoral supervisor Alain Desrochers for agreeing to let me use some of our data here in slightly modified form as an example. I would also like to thank two anonymous reviewers for their useful comments on a previous draft of this paper. Finally, I would like to thank Denis Cousineau for his helpful suggestions concerning the text and the macro code.

Correspondence concerning this article should be addressed to Glenn Thompson, School of Psychology, University of Ottawa, 145 Jean-Jacques Lussier, Box 450, Station A, Ottawa, Canada, K1N 6N5 or e-mail (GlennLThompson@gmail.com).

In such cases, the data points or observations contributed by each subject constitute clusters of inter-dependent scores within the dataset as a whole that are often summarized by a mean to satisfy the assumption of independence (e.g. mean RT in response to high- or low-frequency words). Such clustering is a consequence of a dataset structure that is said to be complex, hierarchical, or multi-level.

The purpose of this paper is to present an SPSS macro for analyzing multi-level data using the Random Coefficient method proposed by Lorch & Myers (1990). First, the case is made for abandoning a commonly used strategy for accommodating multi-level data, namely aggregation (i.e. computing subject- or item-type means for each experimental condition), in favour of Random Coefficient Analysis (RCA). This discussion is limited to a relatively non-technical summary of previous work. Readers interested in more technical details, such as mathematical proofs, may consult Lorch & Myers (1990) and the other relevant sources cited below. The development of the justification behind using RCA is followed by (a) instructions for using the SPSS syntax provided here to perform RCA, (b) a description of how the results of such analyses are interpreted using a concrete example, (c) demonstration of a method for evaluating the magnitude of potential bias in a dataset, and (d) suggestions on how to report a RCA in an empirical article.

Strategies for Analyzing Hierarchically Structured Data

It is easiest to introduce the concept of hierarchical structure with a concrete example. As discussed above, hierarchically structured datasets contain clusters of inter-dependent observations that are caused by the presence of multiple levels of analysis (e.g. participant, item). For instance, a researcher may examine whether a group of participants reads frequently used words aloud more rapidly than words that are used infrequently in print. A sample of high- and low-frequency words (e.g. 20 each) is collected and each participant (e.g. N = 40) generates a single response for each word. The comparison of interest is high- vs low-frequency words on RT. Each item is associated with 40 responses and each participant produced 20 responses per condition. Clearly, whether the dataset is examined from the item or the subject perspective, it contains correlated observations that must be accommodated somehow.

Cohen, Cohen, West, & Aiken (2003) identify three possible ways of dealing with this situation, which they refer to as the "clustering" problem (p. 539). The simplest strategy is called disaggregation, which amounts to ignoring the correlated responses in the dataset despite the fact it violates a fundamental assumption of regression (i.e. independent observation). The problems with this strategy are so obvious that they will not be considered further here. A much more common strategy involves replacing correlated observations with an estimate of central tendency like the mean so that the dataset satisfies the assumption of independent observation or, stated differently, the assumption of uncorrelated observation (Hox, 2002). This strategy is called *aggregation* and it yields results that are subject to important, but oft-ignored, conceptual limitations (e.g. ecological fallacy; Robinson, 1950) and a host of concomitant statistical problems that are discussed below. A third strategy, sometimes called *Random Coefficient Analysis* (i.e. RCA), involves first analyzing the data within individual participants, and then determining whether the magnitude of this within-subject effect differs significantly from zero on average for the sample of participants. In what follows, the drawbacks of the aggregation strategy are presented and then the superior alternative RCA is discussed.

The Aggregation Strategy

The habitual way of dealing with the type of hierarchically structured data discussed here is aggregation (Hox, 2002). Aggregation involves computing an estimate of central tendency to summarize multiple scores at one level of analysis (e.g. items) with a single observation at another level of analysis (e.g. participants). For example, the responses of each participant might be averaged over items within cells of the experimental design (averaging over subjects is also an option, but bias remains an issue).[1] For the scenario developed earlier, this procedure would result in two observations per participant: mean RT for high- and low-frequency words. The 'by-subjects' solution yields data that can be submitted to a repeated-measures analysis because the structure of the dataset has been simplified to reflect the comparison of interest, which is high vs low frequency words. However, while accurately estimating cell means, aggregation decreases the complexity of datasets at the expense of (i) forcing researchers to choose between performing a by-subjects analysis or a by-items analysis, (ii) decreasing the accuracy of population variability estimates, and (iii) inflating the probability of spuriously rejecting the null hypothesis (Lorch & Myers, 1990; Raudenbush & Bryk, 2002; Raaijmakers, 2003).

Type-I error inflation arises in the by-subjects frequency example because the treatment effect is confounded with the

[1] Raaijmakers (2003) recommends analyzing aggregated data by-subjects or by-items, but not both unless there is a special reason for doing so. Statistics that are available for combining the F-tests produced by both analyses are too conservative in most cases (i.e. they inflate Type-II error rates). The RCA method presented here renders the point moot, however, as it effectively combines both by-subjects and by-items analyses.

degree to which the treatment effect varies across participants. Thus, statistically significant effects that are observed with ANOVA using the aggregate strategy may be due to a participant by experimental-effect interaction rather than the experimental effect per se (Lorch & Myers, 1990). To understand why this is so, the logic of the F-test (ANOVA, Regression) needs to be understood.

The reasoning behind the F-ratio test is simple. If an estimate of overall variance is large enough relative to an estimate of error variance, then the null hypothesis of no experimental effect is rejected (Howell, 2002, pp. 324-325). The estimate of overall variance is called the mean square treatment (MS_{effect}) and the estimate of error variance is called the mean square error (MS_{error}). The MS_{effect} estimate comprises two general components: (a) variance caused by the experimental effect and (b) error variance. In contrast, the MS_{error} variance estimate is comprised solely of error variance. If MS_{effect} and MS_{error} are equal, then F is equal to 1, and there is obviously no treatment effect. If MS_{effect} is greater than MS_{error}, then the probability of obtaining the resulting F value – assuming the null hypothesis of no effect is true – is determined using its degrees of freedom and the known F-ratio probability distribution. If this probability is small enough (say .05), we reject the null hypothesis in favor of the alternative: the treatment has an effect. The logic of the F-ratio test, of course, only holds if the error variance component within MS_{effect} is comparable to that represented by MS_{error}.

Regrettably, the aggregation strategy produces MS_{effect} and MS_{error} terms that contain qualitatively different error estimates. The violation of the F-ratio's logic is apparent when one examines the MS_{effect} (1) and the MS_{error} (2) terms that result from aggregation:

$$MS_{effect} = [\text{Treatment effect}] + [\text{(Subject} \cdot \text{Treatment Interaction error)} + \text{(Residual error)}] \quad (1)$$

$$MS_{error} = [\text{Residual error}] \quad (2)$$

An F-ratio based on such a mean square error does not disentangle the contribution of the experimental effect (i.e. the linear component) and the degree to which the treatment effect varies across participants (i.e. the non-linear participant by experimental-effect interaction). Thus, a statistically significant effect could be due to one of three things: (a) a significant experimental effect, (b) significant variation in the treatment effect across participants, or (c) both of these things. The ambiguity is caused by the absence of the participant by treatment interaction in the MS_{error} term. This source of error must be present in the denominator of the F-ratio to statistically control for its presence within the numerator.

An obvious solution to this problem would be to generate an F-ratio based on an MSerror term that includes both the 'residual error' and the 'participant by treatment interaction' components. This more appropriate error term can be obtained by analyzing the effect of item variables within each subject separately and then testing whether estimates for these effects differ reliably from zero, on average, across participants. The MSerror that is produced using such a strategy effectively isolates the treatment effect when the F-ratio is computed (3).

$$MS_{error} = [\text{(Subject} \cdot \text{Treatment error)} + \text{(Residual error)}] \quad (3)$$

Among other things, this strategy avoids the biases inherent in the aggregation strategy, and it is less awkward to apply because the statistical analysis is tailored to the dataset rather than vice-versa.

Random Coefficient Analysis (RCA)

Lorch & Myers (1990) recommended the use of a simple RCA procedure for analyzing the type of multi-level data that is common in experimental research.[2] Within this repeated-measures analysis. To control for the error variance associated with aggregation bias, the product of each dummy variable with the predictor variable(s) is entered into the regression equation (i.e. the participant by treatment interaction).

While this strategy and the RCA method presented in the body of this paper accomplish the same goal, there are reasons for choosing one over the other as circumstances dictate (Cohen et al, 2003, pp. 565-566). The fixed-effects approach is appropriate when the cluster has a substantive meaning, but less useful for cases where the grouping is simply a random sample across which a researcher wishes to generalize experimental effects. For example, a social-psychology project where participants are clustered within ethnic neighborhood may be an appropriate case. In the case of item-attribute effects, the participants are not meaningful

[2] An analytical strategy that accomplishes the same thing as RCA is called the *fixed-effects approach to clustering* (Snijders & Boskers, 1999, as cited in Cohen et al, 2003, p. 541; Presented as an alternative way of implementing RCA in Lorch & Myers, 1990). This strategy requires that the raw data file (as defined in the body of the text) be analyzed with a single regression equation. The experimental effect and error variance (e.g. participant by treatment interactions) are disentangled by entering a series of dummy variables in the regression equation along with the predictor variables. The simple between-participant differences are statistically controlled by coding the identity of each participant using N-1 dummy variables. These variables remove the same variance that is associated with the main effect of participants (i.e. the individual differences) in a standard

context, the term random refers to the fact that RCA examines the effect of IVs on the dependent variable (DV) indirectly via the values of unstandardized beta coefficients that are sampled 'at random' from a probability distribution for each participant (for elaboration on this use of the term 'random', see Cohen et al, 2003, p. 544).

RCA is a two-step procedure for evaluating the reliability of effects in hierarchically structured designs that has been used only sporadically by cognitive scientists (e.g., Borowsky, Owen, & Masson, 2002; Chateau & Jared, 2003) despite the fact that it estimates the statistical significance of experimental effects more accurately than does aggregation. At step one, the analysis begins with the assessment of item-level effects within each participant (e.g. item-characteristics predicting subject responses). For step two, the statistical significance of the item-level effects across participants is assessed using standard tests like the single-sample t-test and possibly ANOVA/regression. In other words, the influence of subject-level variables can be evaluated at this point (e.g. individual differences like gender). The combination of steps 1 and 2 are essentially equivalent to a least-squares estimated hierarchical linear or multi-level model (Hox, 2002; Raudenbush & Bryk, 2002). Such analyses can be applied to many types of hierarchically structured data (e.g. students nested within schools; children nested within families), but only the special case of items nested within participants will be considered here. In what follows, steps 1 and 2 of the Lorch & Myers (1990) method for RCA are each described in turn.

The first step in performing an RCA is to run an ordinary regression within each participant. This level of analysis can be considered the item-level or level-1. Each regression at the item-level involves the prediction of a DV (e.g. RT) on the basis of a set of predictors that can be item attributes or other types of variables (see Hox, 2002). These predictors can be both main effects and interactions involving categorical (e.g. ANOVA design; Cohen et al., 2003, pp.302-308) and/or continuous variables (pp. 255-300). The N regressions performed during the first step of a RCA yield N unstandardized beta coefficients for each item-level IV. These beta coefficients serve as data at the level of participants in the second step of RCA (level-2).

For the second step, the beta coefficients from step one can be used to answer at least two kinds of level-2 statistical questions (i.e. participant-level questions). First, a researcher might be interested in determining whether estimates of an item-level effect, which are represented by beta coefficients computed for each participant, are significantly different from zero for the sample on average. Alternatively, a researcher might be more interested in evaluating the effect of individual differences by, for example, comparing groups of participants to each other. Both types of comparisons are possible so long as the same type of regression analysis is performed within each participant (i.e. same predictor variables).

The first type of participant-level test is performed by comparing a collection of beta coefficients to the value of zero using a single-sample t-test. A statistically significant result indicates that, providing that the null hypothesis is true, the probability of observing an average beta coefficient as big as this or bigger for the sample of participants is less than the nominal alpha level. The second kind of participant-level test evaluates the relationship between participant variables (e.g. gender) and item-level parameters. To test whether a participant-level variable has a direct effect on the DV (main effect), its association with the item-level intercepts is evaluated. The intercept is useful here as it is a baseline value on the DV for each participant with which a predictor may or may not be associated. Of some use is the fact that, when predictors are centered (see below), regression equation intercepts can be interpreted as an unweighted mean for the participant on the DV. To test whether a participant-level variable interacts with an item-level predictor by modulating its effect on the DV, a regression predicting the item-level unstandardized beta coefficients is performed. The nature of the observed relationship depends on whether the participant-level variable is associated with an increase or a decrease in the absolute magnitude of the item-level beta coefficients (increasing or decreasing the strength of the effect) and whether the direction of the effect is reversed.

In summary, RCA represents the complexity of hierarchical datasets in a single, if multi-step, procedure without introducing the bias associated with aggregation. The advantages of RCA are many and they include (a) unbiased estimation of error variance, (b) synthesis of 'by-subject' and 'by-item' analyses within a single procedure without artificially inflating Type-II error (for a discussion of the limitations associated with other strategies for combining subject and item analyses, see Note 2; for a more detailed discussion, see Raaijmakers, 2003), (c) outputs that facilitate use of alternative forms of data presentation (e.g. confidence intervals for main effect and interaction slopes; for recommendations, see Loftus, 1996; for formula and other details, see Loftus & Masson, 1994; Masson & Loftus,

---

per se except in so far as they are associated with participant-attributes (e.g. gender). Therefore, the type of research discussed here (item-attribute effects within and across participants) is more appropriately submitted to an RCA. The only exception is the case where there are fewer than 10 participants, in which case the fixed-effects method is recommended.

2003), and (d) the possibility of using continuous predictor variables so as to avoid the loss of power associated with imposing an artificial dichotomy on the predictor to satisfy the requirements of a by-subjects ANOVA with aggregated data (on the cost of dichotomization, see Cohen, 1983; Donner & Eliasziw, 1994; Hunter & Schmidt, 2004, p. 210). Finally, the procedure does not require major leaps in conceptual and mathematical understanding for the typical researcher because it is based on a straight-forward combination of statistical techniques that are covered in undergraduate-level statistics courses (regression, t-test).

The simplicity of RCA is advantageous as the strengths and limitations of the statistical procedures on which it is based are well-studied and familiar to most researchers. The responsible application of RCA involves, among other things, ensuring that the assumptions of regression and of t-test analyses are satisfied for the data to which they are applied. For example, the assumptions of regression must be verified within each participant to ensure the validity all inferences (for a detailed treatment see Cohen et al., 2003; for an introductory treatment see Tabachnick & Fidell, 2001). Similarly, care should be taken to plan studies that are likely to have sufficient power for detecting meaningful effects at each step in the analysis. Formulas for power calculation are widely available for regression and t-tests. [3]

When should RCA be applied? RCA is appropriate when data are hierarchically structured and the fundamental assumptions of regression are satisfied (for its application to the case of a binary DV, see Myers & Broyles, 2000). Many, if not most, experiments reported in the cognition literature that examine the effect of item-attribute variables across participants meet these criteria. Under certain circumstances, RCA may prove to be useful as a tool for verifying results obtained using an aggregation strategy (e.g. it is reported alongside conventional analyses by Chateau & Jared, 2003). In principle though, RCA can and should replace more commonly-used techniques like aggregation when it is appropriate (see above), unless the total number of participants is very small (e.g. $N < 10$), in which case a procedure known as the *fixed effects approach to clustering* should be employed (see Note 2).

More advanced procedures for estimating parameters (e.g. maximum likelihood) and adjusting parameters (e.g. Empirical Bayes estimation) within a random coefficient framework are available with sophisticated specialized programs like HLM (e.g. Raudenbush, Bryk, & Congdon, 2004) or MLwin (e.g. Rasbash et al., 2000), and also in SPSS. However, the simple RCA method described here and developed by Lorch & Myers (1990) is a viable alternative to aggregation for researchers without the background necessary for using more advanced techniques effectively (for readable introductions to such analyses see, Hox, 2002; Raudenbush & Bryk, 2002).

A Macro for RCA

It is possible to perform the RCA described above using the SPSS drop-down menus (i.e. with a mouse), but this procedure is time-consuming and prone to errors. A more efficient and reliable strategy is to run RCA analyses using a user-supplied program called a macro. In the appendix, the macro syntax for performing RCA as well as some syntax for executing the macros is presented. The functions performed by this syntax are described in what follows. In the final section, this macro is applied to some realistic data in the hopes of facilitating understanding of RCA in general and the macro in particular.

RCA in SPSS

To perform a RCA, the relationship between the DV (e.g. RT) and the IVs must be summarized by beta-weights for each participant. In the appendix, the macro named 'RCAsetup' performs this function. To use this macro, the raw data file must contain a variable that identifies each participant uniquely (i.e. an ID variable), one variable or column for each item-level independent variable (IV), and a variable/column for the dependent variable (DV). Essentially, the data file should be structured so that each row represents a single experimental trial. The datum (DV

[3]. Green (1991) proposed formula for estimating power and sample size for regression analyses. To ensure adequate power for the item-level coefficients (e.g. minimum of .80), the following formula can be applied for estimating N: $N \geq (8 / f^2) + (m - 1)$, where N is the number of observations required, m is the number of beta coefficients to be estimated within each participant, f^2 is equal to $r^2 / (1 - r^2)$, and r^2 is the expected effect size (ω^2 or another adjusted "percent explained" effect-size estimator may be used in place of r^2). For t-tests, Campbell and Thompson (2002) present a simple technique for computing effect sizes, power, and required N to achieve a given level of power for tests with 1 degree of freedom (for other cases, see Levine, 1997). First, the expected effect size (Cohen's d) is estimated based on past research (See Thalheimer & Cook, 2002, for many simple formulas for calculating observed d in published research using commonly reported statistics like MS_{error}) or the values proposed by Cohen (1988) for small (.20), medium (.50), and large (.80) effects. For both regression and t-tests, special circumstances that would reduce power like non-normal distributions within participants or unreliable measures would require that the sample size be increased to maintain the stated level of power (Tabachnick & Fidell, 2001, p. 117).

value) for a given trial is generated by a participant (Subject ID) in response to an item with particular properties (IV 1, IV 2, etc…). For example, the first trial in a data set may contain an RT of 566, a subject ID of 1, and values for the IVs of .5 (e.g. for 'high frequency') and -.5 (e.g. for 'low imageability'). If necessary, an IV representing the interaction between two IVs can be created by computing a variable representing the product of the two variables/vectors (for important cautions when using continuous IVs, such as the need to center predictors, see, Cohen et al., 2003, pp. 255-300).

If the data file is structured appropriately, executing the macro should generate a new data file containing one row with an intercept and a series of unstandardized beta-weights for each participant. The number of beta-weight variables in the new file should be equal to the number of first-order predictors (e.g. item-level variables like frequency or imageability). In the new file, the initial ID field is preserved but all other variables in the original file will be absent.[4]

The second macro tests the mean value of each beta-weight variable against 0 (i.e. a one-sample t-test). This macro generates an output containing descriptive statistics for the beta-weight variables (Mean, Standard Deviation, Standard Error), 95 percent confidence intervals for the average beta-weights, and summary statistics for a single-group t-tests, which is equivalent to a repeated-measures test of the differences between conditions with unbiased error terms. If an ANOVA or a Regression is desired using participant-level IVs, then it can be performed through the drop-down menus in the usual manner using the intercepts (CONST_) as a DV to test for a main effect, or using the beta-weights as DV to test for interactions between IVs (item-level by participant-level interaction).

To use these macros effectively, they should be executed in isolation. This can be accomplished by selecting the relevant syntax, and selecting 'run current' from the right-click menu. Executing this syntax loads the macros into SPSS memory. Once the macros are in memory, separate syntax must be provided for analyzing data with the macros (i.e. macro calls). This syntax must begin with the macro name and be followed by a list of variables to be included in the analysis. For the first macro, this list of variables must be provided in a specific order: ID variable name in rounded brackets, followed by the DV name in rounded brackets, and finally a list of item-level IVs in rounded brackets. An example of syntax for executing the macros is provided in the Appendix, but the variable names that correspond to those in your database must replace the default names. A description of how to analyze the dataset that is available for download with this article using the supplied macro syntax is presented in the following section.

An Example to Try with the RCA Macro

This section begins with a description of how to use the RCA syntax to analyze the data provided. A detailed description of how to interpret the results generated by this analysis is then undertaken, which is followed by suggestions for effect size estimation in RCA, and a comparison of the RCA results with the results of an analysis based on aggregated data.

Replication of the example reported here requires the use of two files that are available for download with this article. The first is an SPSS data file labeled *Thompson.sav*. The second is an SPSS syntax file labeled *Thompson.sps*. The data file contains four variables (columns): participant ID number, the DV (RT in milliseconds, ms, for a specific item) and the IVs Imageability (coded as - .5 = low-imageability, + .5 = high-imageability), Frequency (coded as - .5 = low-frequency, + .5 = high-frequency), and the interaction between the two, which was obtained by the following SPSS command:

```
COMPUTE fxi = freq*imag.
EXECUTE.
```

The data are structured so as to allow the regression equivalent to ANOVA to be performed within each participant. The logic behind using the values .5 and -.5 to denote membership within levels of the IVs is explained in the section below entitled 'Coding issues'.

To perform Step one of RCA on these data, both the data file and the syntax file identified above must be opened in SPSS using File/Open/Data and File/Open/Syntax from the drop-down menu. Begin by examining the syntax file. It contains two types of lines: those that begin with the character * are dedicated to comments explaining the syntax, which are ignored by SPSS; those that do not begin with this character contain active syntax that is interpreted by SPSS when executed. Initially, the RCA macros must be loaded into memory. To do this, select the block of text containing the macro syntax, right-click the mouse, and then click the 'run current' option. The macro is now available to be called

[4] If applicable, participant-level variables may be recovered from the original file by using the Data/Merge Files/Add Variables from the drop-down windows. From the resulting window, simply select the original data file and choose to import the participant-level variable of interest. The two files can be matched using the participant ID field. Once participant-level variables are in the same file as the unstandardized beta coefficients, level-2 hypotheses can be tested in the usual way using intercepts as DV to test main effects and the average beta coefficients as DV to test for interactions (see text).

upon like any other syntax command. To call the first macro into action, select the syntax beginning with the word 'RCAsetup'. Executing this syntax causes SPSS to perform a standard regression analysis within each participant individually, and then to create and open a data file named betas.sav containing a row of unstandardized beta coefficients for each participant. Examine the output file that is generated for error messages and then close the output file without saving. If there are no problems, execute the next line of syntax to call the second macro into action and test whether the item-level effects are significantly different from zero for the sample of participants (e.g. for frequency, imageability, and frequency by imageability).

Coding the predictors.

In this example, we are analyzing ANOVA type data (categorical predictors, continuous DV) using a regression approach. In order to produce meaningful unstandardized beta coefficients, the predictors, in this case frequency and imageability, must be given appropriate values. Cohen and colleagues (2003) suggest a number of methods for coding categorical variables that produce equivalent overall regression equations, but different unstandardized beta coefficients. Arguably the simplest of these strategies is to "dummy code" the IVs assigning the value of 0 to one group and the value of 1 to the other.[5] However, in most cases it is desirable to center predictors so that 0 represents the average value of each predictor. Centering predictors prior to running regression produces beta coefficients that represent the effect of a predictor averaged over levels of the other predictors included in the analysis, which is useful since that is precisely the type of effect that an F-test in an factorial ANOVA table evaluates. Similarly, centering predictors causes the intercept to be equal to the value of the DV when all predictors are average (i.e. 0), which makes it the unweighted participant mean on the DV across all predictors.

In the present case, the item-level IVs are centered round the value zero because the low-frequency and high-imageability items are coded as -.5 while the high-frequency and high-imageability items are coded as +.5. We use the absolute value of .5 so that the difference between groups is equal to 1 [.5 – (-.5) = 1], which is important to ensure that the unstandardized beta coefficients is easy to interpret. If the difference between codes was a value other than 1, the beta coefficient would not represent the average difference between levels of a main effect. This is true because beta coefficients represent the average increase in DV associated with a 1-unit increase in the predictor.

Interpreting the results.

The Thompson.sav file contains data from 64 participants (mixed condition; Thompson & Desrochers, 2003) that were modified slightly by adding a small non-zero value that was sampled from a normal distribution to each observation. The original data were taken from an experiment examining the influence on lexical decision performance of lexical frequency (i.e. the frequency of occurrence of a word in a corpus of text) and imageability (i.e. the ease with which participants evoke a mental image in response to a word). Visual lexical decision is a task that requires participants to discriminate between real words and nonsense words that are presented one at a time on a computer screen by button press. The DV in the data file is reaction time (RT) in milliseconds (ms). Each participant made twenty-five responses to words per experimental condition (e.g. 25 highly imageable words of low-frequency, 25 highly imageable words of high-frequency, etc…) for a total of 100 observations per participants minus the data for incorrect responses, which were discarded prior to analysis. Participants made an equal number of responses to nonsense words and these were also discarded.

As noted above, regression analyses were performed separately for each participant (i.e. step one). Interpretation of an average beta coefficient for a sample of participants is similar to the interpretation of beta coefficients generated by a more conventional analysis. If a beta coefficient is interpreted as the average x-unit increase in the DV associated with a 1-unit increase in the IV, then the average beta for a sample of participants is interpreted as the mean average x-unit increase in the DV associated with a 1-unit increased in the IV for the sample. From an ANOVA perspective, an average beta is simply the average mean difference between conditions for the sample.

For the present example, interpretation of the average beta coefficients is relatively easy. Because the main effects discussed here only have two levels (coded -.5, +.5), an average beta coefficient is equal to the average difference in ms across participants between the two conditions and also the average effect of the IV in ms. If the procedure described above was executed correctly, the results should indicate that the average beta coefficients for the Frequency effect, the Imageability effect, and their interaction are -129.13, -53.33, and 84.43 respectively. The single-sample t-tests indicate that all three effects are significantly different from zero. Thus, we have observed statistically significant main

[5] If there are more that two levels (g) per predictor, the categorical variable is represented by g-1 dummy variables (e.g. 0 vs 1) or contrast code variables (.5 vs -.5). For details on how to devise coding schemes for regressions with categorical independent variables (especially orthogonal coding schemes), see Cohen et al (2003), Tabachnick & Fidell (2001, pp. 149-150), or another good statistics textbook.

Table 1. Magnitude of Aggregation Bias: Comparing ANOVA Statistics and Effect Size Estimates for the Aggregation and RCA Strategies

Strategy		MS_{effect}	SS_{error}	MS_{error}	F-ratio	d_{Cohen}	Partial η^2
Aggregation (A)							
	Frequency (F)	1067243.74	346104.96	5493.73	194.27	-1.76	0.76
	Imageability (I)	182013.45	205091.60	3255.42	55.91	-0.94	0.47
	F x I	114060.36	207184.95	3288.65	34.68	1.48	0.36
RCA							
	F	1067242.72	351601.56	5580.98	191.23	-1.74	0.75
	I	182013.50	208300.96	3306.36	55.05	-0.94	0.47
	F x I	114060.47	210460.74	3340.65	34.14	1.47	0.35
Bias (A – RCA)							
	F	1.02	-5496.60	-87.25	3.04	-0.02	0.01
	I	-0.05	-3209.36	-50.94	0.86	0.00	0.00
	F x I	-0.11	-3275.78	-52.00	0.54	0.01	0.01

Note. The observed (retrospective) power for all tests is effectively 1. MS_{effect} is equal to SS_{effect} because all treatment effects have a single degree of freedom. Partial η^2 reflects the proportion of variance explained by a predictor after between-subject variance and the variance attributable to the other predictors has been removed. The symbol d_{Cohen} is an estimate of effect size that expresses the mean difference between conditions in standard deviation units. All A and RCA effects in the example data are statistically significant for degrees of freedom (1, 63) at $p < .01$.

effects for frequency and imageability, and a significant interaction between the two. These significant effects are interpreted as follows. The direction of the frequency effect (negative) indicates that high frequency words (coded +.5) are read aloud 129.13 ms (Standard Error = 9.27) more rapidly than low-frequency words (coded -.5), t (63) = 13.93, p < .001. The direction of the Imageability effect (negative) indicates that high-imageability words (coded +.5) are recognized 53.33 ms (Standard Error = 7.13) faster than low-imageability words (coded -.5), t (63) = -7.48, $p < .001$. These two main effects are qualified by a statistically significant interaction, average unstandardized beta = 84.43 (Standard Error = 14.34), t (63) = 5.89, $p < .001$. The signs of the main effects (both negative) and the interaction (positive) indicate that the effect of one IV is reduced as the value of the other increases. Decomposing the interaction so that it can be fully interpreted requires a bit more work.

Simple effects testing.

Statistically significant interactions are the justification for examining the statistical significance of one IV within levels of another. The reason for this is clear in RCA as the unstandardized beta coefficient for the interaction is literally the average difference between the simple effects (or simple slopes) of one IV across levels of the other for the sample. Because the difference between simple effects is significant here, we know that the effect of imageability is statistically different depending on the frequency of the associated words. We now might want to determine more precisely what the nature of the imageability effect is within levels of frequency. For example, is the effect of imageability reversed as frequency-level changes? It is possible to answer this question using the simple slope tests that are available in the literature (Aiken & West, 1991; Cohen et al, 2003). However, an easier way to perform simple effects testing with categorical predictors is to re-run the analysis that tests the effect of only one of the IVs (e.g. imageability) twice: once using only low-frequency words and once using only high-frequency words.

To perform simple effects testing with the example data, open the original data file (Thompson.sav) and then execute the first block of simple effects syntax. Then, open the original data file again (without saving the version that is already open) and execute the second block of simple effects syntax (see Thompson.sps). This procedure will re-execute the original analysis twice, once with high-frequency words

and once with low-frequency words. If performed correctly, the results should indicate that the effect of imageability is statistically significant for low-frequency words only. The average beta value for the imageability effect within the low-frequency condition is -95.54 ms (Standard Error = 11.72), $t(63) = 8.15$, $p < .001$. In contrast, the average unstandardized beta coefficient for high-frequency words is only -11.11 ms (Standard Error = 8.19), $t(63) = 1.36$, $p = .18$. Examination of the 95 % (within-subject) confidence interval for this non-significant difference, which is provided in the output, indicates that the data are consistent with both a relatively large high-imageability advantage over the low-imageability condition (lower-bound for the difference between conditions: -32.88 ms) and a small effect that is of about the same magnitude as the observed difference in the opposite direction (upper-bound for the difference between conditions: + 10.65 ms). In other words, the evidence for an effect of imageability with high-frequency words, in either direction, is weak to say the least. Note that the difference between the simple slopes is equal to the average beta coefficient for the interaction that was obtained in the overall analysis, – (-11.11) – (-95.54) = 84.43.

Calculating effect-size in RCA.

Lorch & Myers (1990) did not recommend an estimate of effect size for the RCA technique described here. Reaction time (e.g. in ms) is a DV that has an intuitive meaning and therefore standardized measures of effect size are less relevant than they otherwise might be. However, if standardized estimates of effect size are desired, it is possible to compute an appropriate within-subjects Cohen's d (d_{Cohen}) by dividing the average beta coefficient value for the sample of participants, which can be considered an average difference between conditions, by its standard deviation, which can be considered the standard deviation for the difference between conditions (for more details, see Cohen, 1988; Howell, 2002, pp. 235-236). This estimate of effect-size can be conceived as an expression of the experimental effect (absolute difference between conditions) in standard deviation units (i.e. a standardized effect or standardized difference). For more information on d_{Cohen} consult the sources noted above or a good textbook. To explore the implications of different effect sizes on things like distribution overlap try the program g*power 3, which is freely available for download (Faul, Erdfelder, & Buchner, in press).

Readers that prefer thinking about effect size in terms of percent of variance explained can calculate partial-η^2, which is the estimate of observed effect size produced by SPSS for repeated-measure designs, for RCA using sum of squares that are calculated in the manner described below. For any given IV, partial-η^2 is the proportion of variance in the DV left unexplained by the other IVs in the analysis that is accounted for by that variable (analogous to a partial correlation, or more accurately partial-r^2). Using within-subject SS, the formula for this calculation is: partial-η^2 = SS_{effect} / (SS_{effect} + SS_{error}). The d_{Cohen} and partial-η^2 values for the RCA along with those obtained by analyzing the aggregated data are reported in Table 1.

Practical Importance of Aggregation Bias

At this point, one might wonder what the importance of aggregation bias is in practice. Lorch & Myers (1990) examined this issue through simulations, but empirical study of this issue in experimental psychology (esp. cognitive psychology) is hard to find. The purpose of this section is to provide two types of evidence designed to convince skeptical readers of the practical importance of aggregation bias. First, the results of the Lorch & Myers simulations are briefly reviewed. These results demonstrate that in principle the magnitude of aggregation bias can be quite large. Then, researchers are provided with a method to estimate the potential bias in their own data. By providing researchers with this method, it is hoped that the problem presented by aggregation bias will be more difficult to ignore. The method is demonstrated using the data described above, but this demonstration is not intended as a general test of the practical importance of aggregation bias nor should it be interpreted as such.

Review of Simulation Results

Lorch & Myers (1990) demonstrated that the magnitude of aggregation bias, which they operationalized as Type-I error inflation in their simulations, depends on at least three factors that seem to interact synergistically. In other words, the effect of one of these factors is magnified as the magnitude of the others increases. The first and most obvious factor is the variance of between-condition differences in the population (i.e. the population variance for the average beta coefficients computed in RCA). The larger the variance is in the population, the greater the potential magnitude of aggregation bias. We can extend the implications of this finding a bit by calling upon what is known about sampling error, which causes over- or under-estimation of the population variance from sample to sample. Because the magnitude of aggregation bias depends on this variance, it should also vary from sample to sample even if all other factors are kept constant. These fluctuations will be especially important when sample sizes (number of participants) are small (see the law of large numbers; central limit theorem). The two other factors examined by Lorch & Myers (1990) were (a) the number of items per participant and (b) the inter-item correlation within experimental conditions. Both factors are positively associated with

aggregation bias, which is a somewhat counter-intuitive finding in that both factors are positively associated with a score's — in this case a mean's — reliability. In fact, 'improvement' on these two parameters can actually degrade the quality of the analysis in situations where RCA is appropriate and aggregation is used instead. According to the results reported by Lorch & Myers (1990), even with a conservative estimate of variability (.25), inter-item correlation (.20) and the use of only 10 items per participant, aggregation bias can inflate the nominal alpha level from .05 to .79 (their Table 2)! Again, sample to sample fluctuations in reliability (for a discussion of this issue, see Thompson & Vacha-Haase, 2000) could cause additional variation in the extent of aggregation bias across experiments, even those with identical methodologies.

In addition to the factors examined by Lorch & Myers, many other methodological factors (e.g. the magnitude of real experimental effects, the number of participants, design complexity) could directly or indirectly determine the extent of bias. Though we know some things about what tends to increase or decrease the amount of potential bias in a given dataset, it is currently impossible to know what the practical importance of aggregation bias will be for any given experiment. Ultimately, it is best to avoid aggregation bias altogether by using a bias-free technique like RCA.

Assessing the Magnitude of Aggregation Bias with Real Data

The demonstration provided by Lorch & Myers (1990) should be sufficient to convince researchers of the practical importance of aggregation bias. Nevertheless, researchers may be more motivated to change their ways if they are able to evaluate the extent of potential aggregation bias within their own data. Further, in the absence of formal meta-analyses examining the issue, informal tests of aggregation bias that are conducted by researchers could increase awareness of the problem and its potential magnitude for specific types of research. For these reasons, a method is demonstrated for comparing RCA results to those produced by analyzing aggregated data. A secondary benefit of providing this demonstration is that it requires the conversion of RCA statistics into a form that is familiar to many researchers: the ANOVA summary table. Among other things, this transformation allows the easy computation of effect-size estimates like partial-η^2.

To perform the comparison, two types of values were obtained: (a) ANOVA and effect size estimates from an analysis using the aggregation method and (b) ANOVA and effect size estimates computed based on the RCA statistics from the example reported above. The results for the analysis of aggregated data were obtained by running the syntax in the file Thompson_aggr.sps with the following data file open: Thompson_aggr.sav. This data file was produced by transforming the original (Thompson.sav) using the aggregation and restructure functions in SPSS. In contrast, the standard ANOVA counterparts to the RCA statistics reported above were obtained through calculations that were made without the use of SPSS. This conversion process must be explained in detail to be easily understood and replicated. First, the sum of squares effect (SS_{effect}) and the sum of squares error (SS_{error}) are computed based on the unstandardized beta values and their respective standard deviations.

To compute the SS_{effect} values, the unstandardized beta coefficients are squared and then multiplied by N. The exception is the interaction coefficient, which is divided by two before squaring.[6] To obtain the SS_{error} for a given effect, the standard deviation of its coefficient is squared and then multiplied by N. Again, the exception is the interaction SS_{error}, which requires that the standard deviation be divided by two before squaring. The other F-test statistics are derived from the resulting sums of squares values in the usual manner. First, each sum of squares value is divided by its degree(s) of freedom to produce corresponding mean-square values. Second, F-ratios are computed by dividing each MS_{effect} estimate by its associated MS_{error} term. The ANOVA statistics for the RCA and Aggregation analysis are reported for easy comparison in Table 1.

The first thing to note about Table 1 is that the MS_{effect} is identical for both analytical strategies. The hand calculation of the RCA values introduced a little rounding error, but otherwise the comparison is consistent with formal demonstrations that RCA and aggregation produce equivalent estimates of the absolute magnitude of treatment effects (i.e. estimates of means and therefore differences between means; Lorch & Myers, 1990). The second thing to note about the table is the difference between the MS_{error} values that are produced by the two strategies. For the main effects and the interaction, the error term is larger in the RCA than it is in the aggregation analysis. This result was expected because RCA produces relatively unbiased error

[6] Without getting into too many details, we have to divide the interaction beta coefficient by two because the difference between .025 and -.025 is .5 rather than 1. To be able to compute a SS_{effect} that is comparable to (a) the two main effects and (b) the estimate produced by an ANOVA using aggregated data, the 'squared deviations' must be put on the same basis. Since beta coefficients reflect the average increase in the DV associated with a 1-unit increase in x, the scale of the interaction coefficient is effectively double that of the effect used to calculate the SS_{effect} for the aggregation ANOVA. Dividing the interaction coefficient by two puts all effects on the same basis again.

estimates (Lorch & Myers, 1990), which are larger and therefore result in standardized effect-size measures that are smaller than those obtained from aggregated data (e.g. d_{Cohen}).

It is clear that aggregation bias is present in the data. The magnitude of this bias can be examined through standardized statistics like the F values (same df throughout) and the effect size estimates. Note that in Table 1 the differences between the F-ratios across aggregation and RCA strategies range from .54 to 3.04. Given the overall size of the experimental effects, these differences might be considered small. The effect size estimates would seem to support that interpretation because the observed aggregation bias is limited to about 1 percent of partial variance explained (frequency effect, the interaction). Nevertheless, the aggregation bias observed here could mean the difference between a significant result and a null result with a smaller sample size or when examining less powerful effects, which is important if only because statistical significance plays a role in determining whether a study is published.

Reporting RCA Results

At present, there are no official norms for reporting the results of RCA. In principle, the results can be reported as a regression analysis or as an ANOVA analysis as long as use of the technique is acknowledged. A regression analysis can vary in complexity depending on the number of predictors involved. When the number of item-level predictors is large, the results of the analysis can be reported in a table as with any other regression analysis, with proper acknowledgement that the results were obtained using Lorch & Myer's method (for an example, see Chateau & Jared, 2003). If the regression analysis is simply an implementation of an ANOVA design (3 or fewer IVs), then its results can be reported within the body of the text in a manner similar to that used to report ANOVA analyses in the literature (see the example below).

Interestingly, the output generated by the RCA macro presented here lends itself well to reporting within-subject confidence intervals of different types as well as within-subject d_{Cohen} effect sizes (for a critique of the dominant hypothesis testing philosophy, see Loftus, 1996). Within-subject confidence intervals (95%) for main effects and interactions are provided automatically in the output (i.e. for the average beta coefficients). Further, within-subject confidence intervals can be computed for individual cell means by first transforming the standard error for the average beta coefficients (i.e. difference between means) into the standard errors for the means themselves (Note: the value will be the same for both means) by the following formula: $SE_{mean} = SE_{difference} / \sqrt{2}$ (for additional formula and recommendations for calculating confidence intervals for various designs and comparisons, see Loftus & Masson, 1994; Masson & Loftus, 2003).

The italicized paragraphs that follow demonstrate how the RCA results that are discussed above might be reported in-text within the results section of an empirical article. Estimates of effect size are not reported directly, but d_{Cohen}, for example, can be computed using the reported information in the manner described above (simply convert the standard error into a standard deviation, $SD = SE \cdot \sqrt{n}$. Similarly, the average beta coefficients are not reported because they can be derived from the reported t-values and their standard errors ($b = t \cdot SE$) or from the difference between the reported means. The example is intended to reflect a style that is typical of articles reporting repeated-measures ANOVA results in the field of cognitive psychology.

The frequency by imageability (2 x 2) design was analyzed using random coefficient analysis (Lorch & Myers, 1990). Random coefficient analysis (RCA) is a multi-level regression technique that produces unbiased error term estimates, unlike ANOVAs based on aggregated data. Within RCA, the magnitude of an experimental effect is first estimated within each participant and then the hypothesis that these within-subject effects are significantly different from zero for the sample is tested. The tests of main effects and interaction are reported as one-sample t-tests, which in this case are equivalent to correlated t-tests of the difference between conditions, because that is how such effects are evaluated in RCA.

The results indicate that the latency advantage for high-frequency words over low-frequency words (578.42 vs 707.55 ms) is significant for the sample of participants, t (63) = 13.93, SE difference = 9.27, p < .001. Similarly, the advantage of high-imageability words over low-imageability words (616.32 vs 669.65 ms) is significant across participants, t (63) = -7.48, SE difference = 7.13, p < .001. These two main effects are qualified by a statistically significant interaction, t (63) = 5.89, SE difference = 14.34, p < .001. Decomposition of the interaction confirmed that high-imageability words were associated with faster responses than the low-imageability words (659.78 vs 755.33 ms), t (63) = 8.15, SE difference = 11.72, p < .001. In contrast, the effect of imageability fell short of significance for high-frequency words (572.86 vs 583.78 ms), t (63) = 1.36, SE difference = 8.19, p = .18. The 95 percent confidence interval for this non-significant difference indicates that the data are consistent with both a relatively large high-imageability word advantage over the low-imageability condition (lower-bound difference: -32.88 ms) and a small effect that is of about the same magnitude as the observed difference (i.e. -11.72), but in the opposite direction (upper-bound difference: + 10.65 ms). In other words, the results do not support the idea that imageability exerts a meaningful effect on reaction times when words are also high-frequency.

Conclusion

The numerous flaws of the aggregation strategy that is widely applied by cognitive psychologists and experimental psychologists more generally were reviewed. In its stead, it has been proposed that a procedure sometimes referred to as random coefficient analysis should be used to test the effect of item-attribute variables (Lorch & Myers, 1990). The simple RCA procedure proposed by Lorch & Myers was described in general terms, and then an easy to use program for performing random coefficient analysis in SPSS (version 11 or better) was presented. The operation of this program, called a macro, was explained in terms of a concrete example using data that is available for download with this article. The results obtained from analysis of this data were interpreted in detail and suggestions were offered for reporting such results in empirical articles. Finally, a method for evaluating the potential magnitude of aggregation bias for any dataset was presented so that researchers can better appreciate the consequences of choosing to report analyses based on aggregated data.

In closing, RCA can and should be applied in most cases when a repeated measures design is used to examine item-attribute effects. More generally, RCA is preferable to aggregation whenever multi-level data are involved (Hox, 2002). Whether a traditional hypothesis testing approach is adopted or more informative confidence intervals are used, it is important that error variance be accurately estimated because otherwise the validity of effect size estimation (e.g. power- and meta-analyses) and hypothesis testing are negatively affected. In a discipline where the difference between $p = .03$ and $p = .06$ can mean the difference between a manuscript's publication and its rejection, the use of strategies like aggregation that are known to bias error estimation and therefore p-values is hard to justify, no matter what the practical importance of such bias may be.

Program Availability

A syntax file containing both macros is available directly from the author (GlennLThompson@gmail.com) or online (http://www.geocities.com/glennleothompson /OriginalCode.html). The code may also be typed into an SPSS syntax editor from the Appendix or downloaded from the journal's website, where the data reported in this article may also be found.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Borowsky, R., Owen, W., & Masson, M. (2002). Diagnostics of phonological lexical processing: Pseudohomophone naming advantages, disadvantages, and base-word frequency effects. *Memory & Cognition*, *30*, 969-987.

Campbell, J.D. & Thompson, V.A. (2002). More power to you: Simple power calculations for treatment effects with one degree of freedom. *Behavior Research Methods, Instruments, & Computers*, *34*, 332-337.

Chateau, D., & Jared, D. (2003). Spelling-sound consistency effects in disyllabic word naming. *Journal of Memory and Language*, *48*, 255-280.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlational analysis for the behavioral sciences.* (Third ed.) New Jersey: Lawrence Erlbaum Associates.

Donner, A. & Eliasziw, M. (1994). Statistical implications of the choice between a dichotomous or continuous trait in studies of inter-observer agreement. *Biometrics*, *50*, 550-555.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (in press). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods.*

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 449-510.

Howell, D.C. (2002). *Statistical Methods for Psychology, Fifth Edition.* Pacific Grove, CA: Wadsworth.

Hox, J. (2002). Multi-level analysis: Techniques and Applications. London: Lawrence Erlbaum.

Hunter, J.E. & Schmidt, F. L. (2004). *Methods of Meta-Analysis, Second Edition*. London: Sage Publications.

Levine, J. (1997). Overcoming feelings of powerlessness in "Aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84-106.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 161-171.

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

Lorch, R. & Myers, J. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 149-157.

Masson, M.E. & Loftus, G. R. (2003). Using confidence intervals for graphically-based data interpretation. *Canadian Journal of Experimental Psychology*, *57*, 203-220.

Myers, L. & Broyles, S. (2000). Regression coefficient analysis

for correlated binomial outcomes. *Journal of Applied Statistics, 27*, 217-234.

Raaijmakers, J. (2003). A further look at the "Language-as-Fixed-Effect Fallacy". *Canadian Journal of Experimental Psychology, 57*, 141-151.

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2000). MLwin. Multilevel Models Project Institute of Education.

Raudenbush, S. & Bryk, A. (2002). *Hierarchical linear models.* Thousand Oaks: Sage Publications.

Raudenbush, S., Bryk, A., & Congdon, R., (2004). HLM for Windows (6.00). HLM Software.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351-357.

Tabachnick, B., & Fidell, L. (2001). *Using Multivariate Statistics, Fourth Edition.* Boston: Allyn and Bacon.

Thalheimer, W., & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology.* Retrieved January 30th, 2007 from http://work-learning.com/effect_sizes.htm.

Thompson, G. & Desrochers, A. (2003, June). The role of word imageability in visual word recognition. Poster presented at the 13th Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (BBCS), Hamilton, Ontario, Canada.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is datametrics: the test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195.

Appendix

This appendix contains SPSS syntax (version 11 or later) for the two macros that were discussed within the main body of the text (see the downloadable syntax file named: *Thompson.sps*). The macro named 'RCAsetup' performs step one of RCA, which involves running regression analyses within each participant and saving the resulting intercepts and unstandardized beta coefficients to a new datafile. The macro named 'RCAtest' performs step two, which involves testing whether the unstandardized beta coefficients are significantly different from zero for the sample of participants. In what follows, each block of syntax (italicized) is presented within its own section and is preceded by a short description.

Macro for Step One: RCAsetup

The macro called RCAsetup accepts three variable-name parameters as input. The first is the name of the participant identification variable. The second is the name of the dependent variable. The third is a list of the names of the independent variables. If it is executed with an appropriately structured data file open, the program creates a data file called *betas.sav* containing the participant identification variable and an intercept as well as unstandardized beta coefficient(s) for each participant. A commented version of the syntax is available in the downloadable SPSS syntax file. Example syntax for the macro is provided in the final section of the appendix, but the default variable names must be replaced.

```
DEFINE RCAsetup (!positional !enclose ( '(', ')' )/
    !positional !enclose ( '(',')' )/
    !positional !enclose ( '(',')' )).

SORT CASES BY !1 .
SPLIT FILE BY !1 .
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /NOORIGIN
    /DEPENDENT !2
    /METHOD=ENTER !3
    /OUTFILE=COVB('C:\temp1.sav') .
SPLIT FILE OFF.

GET FILE = 'c:\temp1.sav'.
SELECT IF (rowtype_ = 'EST').
SAVE OUTFILE='C:\temp2.sav'
    /DROP=DEPVAR_ ROWTYPE_ VARNAME_
    /COMPRESSED.

GET FILE = 'C:\temp2.sav'.
EXECUTE.
SORT CASES BY !1 .
CASESTOVARS
    /ID = !1
    /GROUPBY = VARIABLE .
SAVE OUTFILE='C:\betas.sav'.
GET FILE = 'C:\betas.sav'.
EXECUTE.

ERASE FILE= 'c :\temp1.sav'.
ERASE FILE= 'c:\temp2.sav'.
EXECUTE.
!ENDDEFINE.
```

Macro for Step Two: RCAtest

The macro called RCAtest accepts a single variable-name input parameter: a list of the item-level independent variables. If it is executed with the file called *betas.sav* open, it produces as output a t-test for each variable that evaluates the hypothesis that the associated effect is statistically significant for the sample of participants. A commented version of the syntax is available in the downloadable SPSS syntax file. Example syntax for the macro is provided in the

following section, but the default variable names must be replaced with more appropriate ones.

```
DEFINE RCAtest (!positional !enclose ( '(', ')' )).
T-TEST
    /TESTVAL = 0
    /MISSING = ANALYSIS
    /VARIABLES = !1
    /CRITERIA = CI(.95) .
!ENDDEFINE.
```

Calling the Macros: Example syntax

What follows is example syntax for calling the macros reported above. In order for the syntax to work, the macros they refer to (RCAsetup, RCAtest) must have been previously loaded into memory. Macros are loaded into memory by selecting the associated syntax and executing it. To adapt the example below to a particular case, simply replace the default variable names (ID, DV, IV1, and IV2) with appropriate variables from the dataset to be analyzed.

```
RCAsetup (ID) (DV) (IV1 IV2).
RCAtest (IV1 IV2).
```

Manuscript received 1 May 2006
Manuscript accepted 4 June 2007.