An Introduction to Multilevel Modeling for Research on the Psychology of Art and Creativity

By: Paul J. Silvia

Silvia, P. J. (2007). An introduction to multilevel modeling for research on the psychology of art and creativity. *Empirical Studies of the Arts, 25,* 1-20.

TABLES AND FIGURES CAN BE FOUND AT THE END OF THE ARTICLE.

*****Note: Figures may be missing from this format of the document**

Abstract:
This article introduces some applications of multilevel modeling for research on art and creativity. Researchers often collect nested, hierarchical data—such as multiple assessments perperson—but they typically ignore the nested data structure by averaging across a level of data. Multilevel modeling, also known as hierarchical linear modeling and random coefficient modeling, enables researchers to test old hypotheses more powerfully and to ask new research questions. After describing the logic of multilevel analysis, the article illustrates three practical uses of multilevel modeling: (1) estimating within-person relationships, (2) examining between- person differences in within-person processes, and (3) comparing people's judgments to a criterion. The breadth, flexibility, and power of multilevel modeling make it a useful analytic tool for the multilevel data that researchers have been collecting all along.

Article:

Statistical methods can open new doors by enabling new kinds of hypotheses to be developed and tested. This article describes the usefulness of *multilevel modeling*—sometimes known as hierarchical linear modeling or random coefficient modeling—for empirical research on art and creativity (Hox, 2002; Luke, 2004). Although it sounds exotic, multilevel modeling is a straightforward extension of conventional regression analyses. Because it is more general, multilevel modeling enables researchers to test hypotheses that cannot be tested with conventional regression or ANOVA models.

Learning multilevel modeling is worth the effort. First, multilevel models test the predictions that researchers mistakenly think that they have been testing. Psychology makes within-person predictions—

when a person thinks *X*, that person is likely to do *Y*—but it conducts between-person tests of those predictions. Research typically measures *X* and *Y* once in a sample of people (or averages across several measures to get single scores) and then estimates the association between *X* and *Y*. This between-person analysis estimates correlations between *variables*—what happens to *Y* when *X* changes. It contains no information about within-person relationships. Testing within-person predictions requires (1) assessing people many times and (2) analyzing the multiple assessments instead of aggregating them.

Second, multilevel models harness the entire data that researchers are already collecting. In the psychology of art, for instance, a typical study presents people with a set of pictures. Each person gives several ratings for each picture (e.g., ratings of preference, complexity, novelty, disorganization). Thus, the ratings are nested within each participant. Instead of exploiting the statistical strength of the repeated assessments, researchers usually average across the pictures to gain single scores per person. After all the work to collect extensive data on each person, the researcher simply averages the scores. Multilevel modeling can identify interesting relationships that are obscured by averaging.

Third, and most importantly, multilevel modeling allows researchers to test interesting hypotheses that cannot be tested with conventional ANOVA and regression approaches. Within- person effects and between-person effects are conceptually distinct. As a result, between-person analyses needn't replicate within-person analyses. In fact, the two analyses can show opposite effects (see Nezlek, 2001). Presuming that between-person analyses of aggregates apply to individual people is the well-known *ecological fallacy* (Robinson, 1950). In addition to assessing within-person relationships, multilevel modeling allows researchers to explore variability in intrapersonal processes. People will differ in within-person relationships. For example, some people will prefer simple art, and other people will prefer complex art. With multilevel modeling, researchers can explain the variability in within-person relationships.

This article shows how multilevel modeling can be applied to common empirical questions in the psychology of art and creativity. This brief article can only outline the logic of multilevel analysis and highlight some of its uses. Like all statistical techniques, multilevel modeling has special assumptions and analytic considerations (e.g., power, centering, estimation methods). Before jumping into multilevel modeling, the reader should consult books and articles that examine multilevel modeling in more detail. Excellent introductions are given by Cohen, Cohen, West, and Aiken (2003, chap. 14), Luke (2004), and Nezlek (2001). Advanced treatments are provided by Hox (2002), Kreft and de Leeuw (1998), Raudenbush and Bryk (2002), and Snijders and Bosker (1999).[1]

Hierarchies Are Everywhere

Multilevel data are everywhere. Many researchers in the psychology of art and creativity have been collecting multilevel data without knowing it. As a result, they have been analyzing their data in ways that fail to exploit the power of the data's hierarchical structure. Nesting is the defining feature of multilevel data. Multilevel data have a hierarchical structure in which scores at one level are nested within another level. Any study that took several measurements from a person or group has a multilevel structure. The most common kind of multilevel data structure has two levels. The lower-level data, called Level 1 data, are nested within higher-level data, known as Level 2 data. Level 1 scores are within-person data; analyses of these scores provide estimates of intraperson relationships. Level 2 scores are between-person data; analyses of these scores provide estimates of relationships between variables.[2]

In empirical research on art and creativity, the most common multilevel structure has many responses or ratings (Level 1) nested within people or groups of people (Level 2). Consider an experiment in which 30 people read 10 poems and judge each poem's creativity and complexity. The researcher has information about the 30 people, such as scores on personality scales, demographic scores, or scores denoting membership in a group. These are Level 2 variables; they describe the higher-level unit (the person, in this case). If the researcher analyzed only the Level 2 data, he or she would have a conventional between-person analysis. The researcher also has information about the 10 poems. If each person rated each poem's level of creativity and complexity, then the researcher has 300 ratings of creativity (30 people giving 10 ratings) and 300 ratings of complexity. Analyzing only the Level 1 data would estimate within- person relationships between judgments of creativity and complexity.

Uncertainty about the correct sample size is a hallmark of multilevel data. In this experiment, does n = 30 (number of Level 2 units, i.e., people) or does n = 300 (number of Level 1 units, i.e., ratings)? The answer is “both”: the researcher has hundreds of ratings nested within 30 participants. Without multilevel modeling, researchers resort to ignoring a level of data. Researchers might ignore the Level 1 data by averaging across the 10 poems. This condenses the 300 creativity ratings into 30 scores, one for each person. The researcher can then conduct ordinary multiple regression analyses. The problem, however, is that the researcher has obliterated a lot of information—he or she went to the trouble of collecting more than 30 observations, but the analysis fails to exploit the full richness of the data set. Furthermore, the averaged scores will be lumpy unless all 10 poems were highly similar. Finally, the researcher is no longer examining within-person relationships. At the within-person level, the 30 people probably had different relationships between ratings of complexity and creativity. Some people may have found complex poems to be more creative; other people may have judged simple poems to be more creative. These patterns are obscured by averaging the within-person ratings.

A researcher determined to examine within-person effects might ignore the Level 2 scores. This analysis treats the 300 ratings as independent observations. An obvious problem, however, is that the researcher doesn’t have 300 independent observations: he or she has 30 sets of nested observations. The excessive degrees of freedom will cause the analysis to underestimate the true amount of error. Furthermore, the researcher is violating the assumption of independence that is central to conventional regression analyses. Nested data are interdependent: people differ in their averaged ratings, the range of values they use in their ratings, and the standard deviations of their ratings. The interdependence must be modeled statistically.

What Is Multilevel Modeling?

Multilevel modeling allows researchers to estimate the effects of Level 2 and Level 1 variables simultaneously. Like conventional regression models, multilevel models can be represented using equations. Consider an experiment in which 40 people viewed 8 pictures. For each picture, people gave ratings of *preference* and of *complexity* on 1–7 Likert scales. Before viewing the pictures, people completed a self-report scale measuring *sensation seeking*. The ratings of preference and complexity are the Level 1 data; the sensation seeking scores are the Level 2 data. Table 1 shows (fictional) data for the first 4 people in the sample.

Theories of aesthetic preference would predict that a person will prefer complex pictures more than simple pictures (Berlyne, 1971). This is a within-person prediction, and it can be assessed by the following equation:

$$\text{Level 1:} \qquad Y_{ij} = \beta_{0j} + \beta_{1j} + r_{ij}$$

The equation can be oversimplified for the sake of clarity:

$$\text{Level 1:} \qquad \text{Preference} = \text{Intercept} + \beta_1(\text{Complexity}) + \text{Level 1 Error}$$

Notice that the Level 1 equation resembles a conventional regression equation: the dependent variable *Y* is a linear function of an intercept, a slope, and an error term. Note, however, that the equation has additional subscripts. These subscripts reflect the nested character of the data. The preference rating *Y* that person *j* gave for picture *i* is a linear function of person *j*'s intercept (β_{0j}), the slope relating person *j*'s rating of complexity to preference (β_{0j}), and residual error for picture *i* for person *j* (β_{0j}).

Conventional regression provides a useful metaphor for understanding multilevel modeling. Forty people viewed 8 pictures, so imagine that the researcher conducted 40 within-person regression analyses. Each analysis provides an estimate of the single person's intercept (that person's preference when other variables are at 0) and an estimate of the single person's slope (how strongly complexity predicted preference). After doing all 40 analyses, the researcher has a distribution of 40 intercepts and a distribution of 40 slopes. Table 1 shows the descriptive statistics and the slopes for the first 4 people in the sample. First, notice that the four people have different average levels of preference and complexity and different standard deviations. These variations reflect the nested character of the ratings and motivate the use of multilevel modeling. Second, notice that the four people have different unstandardized regression weights that connect complexity to preference. These slopes differ in magnitude and in direction. Person 1 and Person 3 had higher preference scores for pictures higher in complexity; Person 2 and Person 4 had higher preference scores for pictures lower in complexity. Averaging the ratings would have obscured these differences.

Why did people differ in their intercepts and slopes? What explains why some people had positive slopes and other people had negative slopes? Recall that we have distributions of 40 intercepts and 40 slopes. We can use these within-person scores as dependent variables: the outcomes of the Level 1 equation become dependent variables in Level 2 equations. Specifically:

$$\text{Level 1:} \qquad Y_{ij} = \beta_{0j} + \beta_{1j} + r_{ij}$$

$$\text{Level 2:} \qquad \beta_{0j} = \gamma_{00} + \gamma_{01} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} + u_{1j}$$

To simplify the Level 2 equations:

Level 2: **Intercept = Group Average + γ(Sensation Seeking) + Level 2 Error**

Slope = Group Average + γ(Sensation Seeking) + Level 2 Error

Notice that $ß_{0j}$, the intercept in the Level 1 equation, is now the dependent variable in the first Level 2 equation. Likewise, the slope $ß_{1j}$ is now the dependent variable in the second Level 2 equation. The Level 2 equations model variance in the within-person intercepts and slopes. For the intercept equation, person ,'s $ß_0j$ score is modeled as a linear combination of the sample's average intercept γ_{00} plus an effect of a Level 2 predictor (sensation-seeking score) and residual error. If the weight for sensation seeking is significantly positive, then sensation seeking explains variance in the Level 1 intercepts. (Note that this analysis is akin to a conventional regression analysis, in which a between-person variable predicts an average of within-person scores.)

For the slope equation, a person's *relationship* between complexity and preference is the dependent variable. It estimates the within-person slopes as the sample's average slope γ_{10} plus an effect of sensation seeking and residual error. The coefficient for sensation seeking, known as a *cross-level interaction,* has a meaning unique to multilevel modeling. This coefficient estimates whether sensation seeking scores explain variability in complexity–preference slopes. In Table 1, for example, two people had positive slopes and two people had negative slopes. Perhaps people high in sensation seeking had positive slopes (they preferred complex art) whereas people low in sensation seeking had negative slopes (they preferred simple art). If this coefficient is significant, then sensation seeking explains between-person differences in the within-person relationships.

It's important to realize that the conventional regression metaphor is only a metaphor. First, multilevel modeling estimates the coefficients simultaneously, not through a two-step process of feeding Level 1 estimates into Level 2 equations. Simultaneous estimation greatly reduces the number of tests that are conducted. Second, conventional regression uses ordinary least squares to estimate coefficients, whereas multilevel modeling uses maximum likelihood

estimation. Finally, multilevel modeling can model the coefficients as *random coefficients*, whereas conventional regression models the coefficients as *fixed* (i.e., each person has the same slope and intercept).

The rest of this article describes how multilevel modeling can be used in practice. These examples will illustrate issues and advantages unique to multilevel modeling. While working through the examples, we'll introduce some new concepts and issues, such as estimating nonlinear effects, exploring variance components, and eliminating "third variables."

Example 1: Estimating Within-Person Relationships

The simplest use of multilevel modeling is to estimate within-person relationships. If a researcher has repeated measures on a group of participants, the repeated measures (Level 1) are nested within features of the participants (Level 2). Sometimes the researcher is interested only in the Level 1 effects. Instead of testing hypotheses at the between-person level, the researcher may decide that a within-person test is more appropriate.

Within-Person Relationships Between Appraisals and Emotions

A recent model of emotional responses to art (Silvia, 2005-b) contends that cognitive appraisals of art predict aesthetic emotions. Like most of psychology's predictions, this is a within-person prediction: when a given person makes appraisal *X*, he or she is more likely to experience emotion *Y*. An experiment used multilevel modeling to explore how cognitive appraisals predicted the emotion of interest (Silvia, 2005a). Based on past research (Silvia, 2005c, 2006b), two appraisals were expected to predict feelings of interest: appraising something as new and complex, and appraising something as understandable.

In the experiment, 34 people viewed 29 pictures (Silvia, 2005a). For each picture, people gave ratings of *interest*, *complexity*, and *ability to understand* on 1–7 scales. The 34 people are the Level 2 units; ratings for the 29 pictures are the Level 1 units. The following equations model the effects of complexity and ability to understand on interest.

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} + \beta_{2j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10} + u_{1j}$

$\beta_{2j} = \gamma_{20} + u_{2j}$

Oversimplified:

Level 1: Interest = Intercept + β_1(Complexity) + β_2(Understand) + Level 1 Error

Level 2: Intercept = Overall Intercept + Level 2 Error

Complexity Slope = Overall Slope + Level 2 Error

Understand Slope = Overall Slope + Level 2 Error

Because the experiment was interested only in within-person relationships, no additional Level 2 predictors were entered. The analyses estimated γ_{00} (the overall intercept), γ_{10} (the overall effect of complexity on interest), and γ_{20} (the overall effect of ability to understand on interest). This example shows that the Level 1 coefficients are estimated via the Level 2 equations. We want to know the value of β_{1j} for example, but we estimate that by estimating γ_{10} and u_{1j}.

The estimate of $\gamma_{10,}$ the overall slope for complexity, was b = .347. This value is an unstandardized regression weight, and it has the same interpretation as in conventional regression analyses. With a standard error of .032, the regression weight was highly significant, $t(33) = 10.7$, $p < .001$. The estimate of $\gamma_{20,}$ the overall slope for ability to understand, was b = .509, SE = .057. This coefficient was significant as well, $t(33) = 8.9$, $p < .001$. Thus, at the within-person level, both appraisals significantly predicted interest.

Analyses of within-person effects will often show that people differed in these effects: people may have different intercepts or different slopes relating predictors to outcomes. The overall slope for complexity, for example, was b = .347, but that represents the average slope for the sample—some people had a bigger slope, and others had a smaller slope. Multilevel modeling expresses these deviations as *variance components*. The Level 2 u_{1j} values, for example, represent the between-person deviations from the overall complexity slope. Estimating the variance of these deviations tells a researcher how widely people differed in their slopes. If a variance component is large, then the within-person effects differed between people. Researchers can then try to model this variance. For example, what Level 2 predictors might explain why some people had larger slopes than others? If variance components are small, then there's little between- person variance to explain—the within-person effects didn't differ much between people. (We'll see how to add Level 2 predictors later in the paper.)

Multilevel modeling software can estimate each person's intercepts and slopes. Sometimes these are important in their own right. Appraisal theories of emotion assume that appraisals never negatively predict emotions (Kuppens, Van Mechelen, Smits, & De Boeck, 2003), so in this experiment it was worth examining whether anyone had negative slopes between appraisals and interest (Silvia, 2005a). Figure 1 shows the slopes relating appraisals to interest. Each panel has 34 lines, one for each participant. The lines differ in their intercept and in their slope, but all of the lines have a positive slope. Thus, the direction of the appraisal–emotion relationships was invariant in this sample.

But what about the third-variable problem? This is a correlational study, so perhaps some trait like openness to experience caused both the appraisals and the interest. Unlike conventional between-person analyses, within-person analyses avoid the third-variable problem. This fact is one of multilevel modeling's biggest advantages. In the study we just described, we estimated the relationship between 29 interest ratings and 29 complexity ratings. This relationship can't be explained by a between-person trait like openness, because each person has only one openness score. A single score has no variance, so it can't covary with the 29 ratings. As a result, within- person analyses estimate relationships that are unconfounded by individual differences. This feature makes multilevel modeling appealing to researchers who use nonexperimental designs.[3]

Effects of Objective Pictorial Balance on Preference

Within-person analyses aren't limited to correlating several ratings, such as the ratings of interest and complexity described earlier. Researchers can use within-person variables that are coded to reflect within-person factors, such as manipulated stimulus features. For example, Wilson and Chatterjee (2005) developed the Assessment of Preference for Balance test, which has 65 images of circles, 65 images of hexagons, and 65 images of squares. These images vary in pictorial balance (Locher, 1996, 2006), and each image has a score that captures the image's objective level of balance. To validate the test, Wilson and Chatterjee (2005, Experiment 2) had 30 participants rate each picture on a 1–5 preference scale. Thus, ratings of the pictures are nested within participants.

Wilson and Chatterjee analyzed the data with conventional between-person regression analyses. They averaged the preference scores across all 30 people and then correlated the objective balance scores with the average preference scores. For the circle and hexagon images, they found strong linear effects of balance on preference. In fact, the linear effect accounted for

73% (circles) and 78% (hexagons) of the between-person variance. For squares, the linear effect accounted for only 44% of the between-person variance. This disappointing result lead Wilson and Chatterjee to drop the square stimuli from the test, which left 130 images (65 circles, 65 hexagons).

Like many hypotheses, the hypothesis that balance affects preference is a within-person hypothesis. The presumed processes are intraindividual processes, so ideally research should show that a single person's preference scores are affected by the manipulation of balance. A within-person analysis of Wilson and Chatterjee's data is straightforward. The large number of within-person assessments—65 for each stimulus type—is ideal for within-person analyses. I reanalyzed their data using multilevel modeling. For simplicity, preferences for circles, hexagons, and squares were treated as separate dependent variables, but one can estimate multivariate multilevel models (see Hox, 2002). Preferences were estimated with the following equations:

Level 1: $$\text{Preference}_{ij} = \beta_{0j} + \beta_{1j}(\text{Balance Score}) + \beta_{2j}(\text{Balance Score}^2) + r_{ij}$$

Level 2: $$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

This equation estimates preference as a linear function of an intercept, a linear effect of balance, and a quadratic effect of balance. The predictor variables are simply centered versions of each picture's objective balance score.

The within-person analyses found primarily *quadratic* effects of balance on preference; Figure 2 depicts the relationships. For preference for circles, there was a nonsignificant linear effect, b = .0074, SE = .007, $t(29) = 1.02$, $p < .32$, but a significant quadratic effect, b = -.0004, SE

= .00009, $t(29) = 4.14$, $p < .001$. For preference for hexagons, there was a nonsignificant linear effect, $b = -.004$, $SE = .011$, $t(29) = .35$, $p < .73$, and a marginally significant quadratic effect, $b = -.0003$, $SE = .00016$, $t(29) = 1.82$, $p < .079$. For preference for squares, there was a significant linear effect, $b = .029$, $SE = .009$, $t(29) = 3.14$, $p < .004$, and a significant quadratic effect, $b = -.0008$, $SE = .00014$, $t(29) = 5.62$, $p < .001$.

Wilson and Chatterjee's (2005) experiment is a nice example of how between-person analyses and within-person analyses can show different effects. Unlike the between-person effects, the within-person effects were primarily quadratic. Imbalance reduced preference, and this negative effect accelerated as imbalance increased (see Figure 2). Furthermore, the images of squares had the weakest between-person relations but the strongest within-person relations. Finally, hexagons had the strongest between-person relations but the weakest within-person relations. It's tempting to ask "which analysis is right?", but neither analysis should be considered the correct analysis. Between-person and within-person analyses answer different questions. The between-person regressions estimate relations between *variables*: they show how balance affects preference. The within-person analyses estimate relationships within the typical person in the sample: they show how changes in balance affect a given person's changes in preferences.

Example 2: Do Within-Person Relationships Differ Between People?

Within-person analyses typically find that people have different within-person relationships. In Figure 1, for example, it's obvious that people had different intercepts and slopes. With multilevel modeling, a researcher can model variance in within-person intercepts and slopes. Conventional regression analyses cannot test such hypotheses.

Do Individual Differences Moderate Within-Person Relationships?

Many studies explore how individual differences predict emotions and perceptions related to art (e.g., Feist & Brady, 2004; Rawlings & Bastian, 2002). These studies are usually between-person studies: after presenting people with many paintings, the researchers average across the paintings. Individual differences are then used to predict these averages. Multilevel models can assess these between-person effects, but they can go a step further: how do individual differences predict within-person relationships?

Recent studies have found large individual differences in preferences for disturbing art (Rawlings, 2000, 2003). In a recent experiment, 83 people viewed 20 classical paintings (Silvia & Turner, 2006). The paintings varied along a soothing–disturbing dimension. After viewing the picture as long as they wished, people rated how disturbing they found the picture. The two Level 1 variables are thus rated disturbingness and viewing time. Before viewing the pictures, people completed measures of personality traits. Breadth of interest, a trait similar to openness to experience, represents a tendency to be interested in intellectual and abstract things (see Silvia, 2006b, chap. 4). Overall, people should spend more time viewing disturbing paintings. This within-person effect, however, should be moderated by breadth of interest. People high in breadth of interest should have particularly strong relations between viewing time and disturbingness.

The multilevel equations can clarify how these predictions are tested.

Level 1: $\text{Viewing Time}_{ij} = \beta_{0j} + \beta_{1j}(\text{Rating of Disturbingness}) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Breadth of Interest}) + u_{0j}$

$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Breadth of Interest}) + u_{1j}$

At Level 1, viewing time is a linear function of person *j*'s intercept, a slope relating the person's ratings of disturbingness to viewing time, and Level 1 error. The intercept and the slope are predictors at Level 1, but they become dependent variables at Level 2. At Level 2, the intercept $ß_{0j}$ is a linear function of the overall sample intercept (g_{00}), an effect of breadth of interest, and Level 2 error. Likewise, the slope $ß_{1j}$ is a function of the overall sample slope (g_{10}), an effect of breadth of interest on the slope, and Level 2 error. Breadth of interest can thus affect the level of viewing time, but it can also affect how disturbingness predicts viewing time.

The study found intriguing multilevel effects. The overall sample intercept (γ_{00}) was 11831 milliseconds (ms). The γ_{10} coefficient reflects the effect of breadth of interest on the intercepts. Breadth of interest had a modest and nonsignificant effect on the intercepts, $b = 5022$, $SE = 3331$, $t(81) = 1.51$, $p < .14$, so people high in breadth of interest didn't necessarily spend more time in general viewing the pictures. The γ_{10} coefficient reflects the overall slope relating ratings of disturbingness to viewing time. This coefficient was marginally significant: as expected, people spent more time viewing pictures that they found disturbing, $b = 179$, $SE = 97$, $t(81) = 1.84$, $p < .069$. This coefficient represents an average of all 83 within-person slopes, and there was a lot of variability in these slopes. The g_{11} coefficient reflects the *cross-level interaction*—it denotes how breadth of interest affects the strength of the relation between viewing time and ratings of disturbingness. This coefficient was significant, $b = 1182$, $SE = 518$, $t(81) = 2.28$, $p < .025$, indicating that the Level 2 variable moderated the Level 1 relationship.

Cross-level interactions are best grasped by graphing the relationships. Figure 3 shows how the disturbingness–viewing time slopes vary as a function of breadth of interest. The two lines represent low and high levels of breadth of interest. This grouping is for graphical convenience. Multilevel analyses don't artificially split continuous variables: they can estimate cross-level interactions with fully continuous variables. The significant interaction comes from the different slopes. For people high in breadth of interest, viewing time strongly predicts ratings of disturbingness. For people low in breadth of interest, the slope is essentially flat.

Do Groups Have Different Within-Person Relationships?

The Level 2 scores can be group memberships. An experiment might include existing groups (experts and novices; Locher, Smith, & Smith, 2001) or it might manipulate a between- subjects variable (color versus black-and-white pictures; Polzella, Hammar, & Hinkle, 2005). The logic of multilevel modeling is the same for Level 2 groups and Level 2 individual differences. For example, a recent experiment explored how training in art affects emotional responses to art (Silvia, 2006a). Past research found that appraisals of novelty–complexity and of one's ability of understand predict interest (Silvia, 2005a, 2005c; Turner & Silvia, in press). Does training in art influence how these appraisals affect interest? For instance, perhaps novices' feelings of interest are more affected by perceptions of whether they can understand the painting. Testing this prediction involves testing cross-level interactions: do Level 1 within-person relations between interest and appraisals vary as a function of Level 2 groups?

Three groups of participants—low, medium, and high training—viewed abstract pictures and rated each picture for interest, complexity, and ease of understanding. The multilevel equation that tested this hypothesis was:

$$\text{Level 1:} \quad \text{Interest}_{ij} = \beta_{0j} + \beta_{1j}(\text{Complexity}) + \beta_{2j}(\text{Understand}) + r_{ij}$$

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Training Group}) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Training Group}) + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(\text{Training Group}) + u_{2j}$$

At Level 1, interest is modeled as a function of an intercept and slopes for the two appraisals. At Level 2, the analyses estimate the overall sample intercept γ_{00} and the overall sample slopes for complexity γ_{10} and understanding (γ_{20}). For this study, the central coefficients are the cross- level interactions of the training group on the slopes relating appraisals and interest (γ_{11} and γ_{21}). The complexity–interest slope was significant, $b = .336$, SE _ .033, $t(66) = 10.3$, $p < .001$, but the cross-level interaction with training was only marginally significant, $b = .097$, SE _ .053, $t(66)$ _ 1.82, $p < .072$. This effect means training did not substantially explain the variance in within- person relationships. Figure 4 depicts this marginal interaction. Note that the three groups have different intercepts—the high training group found the paintings more interesting overall—but that the three slopes are essentially the same. Similarly, the understanding–interest slope was significant, $b = .257$, SE _ .038, $t(66) = 6.78$, $p < .001$, but the cross-level interaction with training was not, $b = -.05$, SE _ .076, $t(66)$ _ .67, $p < .51$. People in the different training groups had the same within-person effects of appraisals on emotions.

Cross-level interactions test intriguing predictions. We are not asking if groups differ in the *amount* or *level* of some variable. In this example, we aren't testing if art training affects the amount of interest or the amount of appraised complexity. Instead, we're asking if some intrapersonal process is the same for different groups. In this example, we're asking if the intrapersonal causes of interest are the same for groups that differ in their art training. Multilevel modeling can assess group differences in levels of variables and in *relationships* between variables. These process-oriented predictions cannot be posed or tested using conventional regression analyses.

Example 3: Comparing People's Judgments to Criterion Judgments

A third application of multilevel modeling is the comparison of people's judgments to criterion judgments. Consider, for instance, a study in which people view 10 movies over the course of a couple months. For each movie, people make many judgments about the movie's level of creativity, including an overall creativity judgment. Each movie has been evaluated by a group of experts that arrived at creativity score for each movie (cf. Simonton, 2005, 2006). It's interesting to assess whether a person's creativity judgments of the movies converge with the expert ratings of creativity. We would expect variability in the level of similarity: some people's judgments will highly correlate with the experts' judgments, but other people's judgments will be unrelated or even negatively related.

Comparing a person's judgment to a criterion judgment is a simple multilevel model:

Level 1: $\text{Person's Creativity Judgment}_{ij} = \beta_{0j} + \beta_{1j}(\text{Experts' Creativity Judgment}) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10} + u_{1j}$

In this equation, the expert ratings are used to predict the individual person's judgments. Each person will have an intercept (the person's average creativity judgment when the other predictors are at 0) and a slope that specifies the relation between the person's judgments and the experts' judgments. Note that a significant slope does not mean 1:1 agreement, but it indicates that the person's judgments follow the same direction as the experts' judgments.

Naturally, one can use Level 2 variables to model variability in the Level 1 relationships. Perhaps group memberships or individual differences can predict how closely a person's creativity judgments align with the criterion judgments. These equations explore how two Level

2 variables—training in cinema and personal interest in movies—predict the Level 1 coefficients:

Level 1: $\text{Person's Creativity Judgment}_{ij} = \beta_{0j} + \beta_{1j}(\text{Experts' Creativity Judgment}) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Cinema Training}) + \gamma_{02}(\text{Interest in Movies}) + u_{0j}$

$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Cinema Training}) + \gamma_{12}(\text{Interest in Movies}) + u_{1j}$

Thus, this model enables a researcher to compare people's judgments to criterion judgments (Level 1 model) and to explain why people differ in how closely their judgments match the criterion (Level 2 model).

Conclusion

Many theories in the psychology of art and creativity make predictions about intrapersonal processes, so researchers should conduct within-person tests of their within-person predictions. Instead of averaging across people, researchers can use multilevel modeling to estimate within-person and between-person relationships. Researchers can simultaneously test within-person predictions and between-person predictions, and they can propose interesting hypotheses that cannot be tested with conventional regression analyses. Because many researchers are already using multilevel designs, they should use multilevel modeling to harness the full information contained within their data.

References

Berlyne, D. E. (1971). *Aesthetics and psychobiology.* New York: Appleton-Century-Crofts.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Feist, G. J., & Brady, T. R. (2004). Openness to experience, non-conformity, and the preference for abstract art. *Empirical Studies of the Arts, 22,* 77–89.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage.

Kuppens, P., Van Mechelen, I., Smits, D. J. M., & De Boeck, P. (2003). The appraisal basis of anger: Specificity, necessity, and sufficiency of components. *Emotion, 3,* 254–269.

Locher, P. J. (1996). The contribution of eye-movement research to an understanding of the nature of pictorial balance perception: A review of the literature. *Empirical Studies of the Arts, 14,* 143–163.

Locher, P. J. (2006). Experimental scrutiny of the role of balance in the visual arts. In P. Locher, C. Martindale, & L. Dorfman (Eds.), *New directions in aesthetics, creativity, and the arts* (pp. 19–32). Amityville, NY: Baywood.

Locher, P. J., Smith, J. K., & Smith, L. F. (2001). The influence of presentation format and viewer training in the visual arts on the perception of pictorial and aesthetic qualities of paintings. *Perception, 30,* 449–465.

Luke, D. A. (2004). *Multilevel modeling.* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07–143). Newbury Park, CA: Sage.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin, 27,* 771–785.

Polzella, D. J., Hammar, S. H., & Hinkle, C. W. (2005). The effect of color on viewers' ratings of paintings. *Empirical Studies of the Arts, 23,* 153–163.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Rawlings, D. (2000). The interaction of openness to experience and schizotypy in predicting preference for abstract and violent paintings. *Empirical Studies of the Arts, 18,* 69–91.

Rawlings, D. (2003). Personality correlates of liking for "unpleasant" paintings and photographs. *Personality and Individual Differences, 23,* 395–410.

Rawlings, D., & Bastian, B. (2002). Painting preference and personality, with particular reference to Gray's behavioral inhibition and behavioral approach systems. *Empirical Studies of the Arts, 20,* 177–193.

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

Silvia, P. J. (2005a). Cognitive appraisals and interest in visual art: Exploring an appraisal theory of aesthetic emotions. *Empirical Studies of the Arts, 23,* 119–133.

Silvia, P. J. (2005b). Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology, 9,* 342–357.

Silvia, P. J. (2005c). What is interesting? Exploring the appraisal structure of interest. *Emotion, 5,* 89–102.

Silvia, P. J. (2006a). Artistic training and interest in visual art: Applying the appraisal model of aesthetic emotions. *Empirical Studies of the Arts, 24,* 139–161.

Silvia, P. J. (2006b). *Exploring the psychology of interest.* New York: Oxford University Press.

Silvia, P. J., & Turner, S. A., Jr. (2006). *Viewing disturbing pictures.* Unpublished raw data, University of North Carolina at Greensboro.

Simonton, D. K. (2005). Film as art versus film as business: Differential correlates of screenplay characteristics. *Empirical Studies of the Arts, 23,* 93–117.

Simonton, D. K. (2006). Cinematic creativity and aesthetics: Empirical analyses of movie awards. In P. Locher, C. Martindale, & L. Dorfman (Eds.), *New directions in aesthetics, creativity, and the arts* (pp. 123–136). Amityville, NY: Baywood.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24,* 322–354.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Turner, S. A., Jr., & Silvia, P. J. (in press). Must interesting things be pleasant? A test of competing appraisal structures. *Emotion*.

Wilson, A., & Chatterjee, A. (2005). The assessment of preference for balance: Introducing a new test. *Empirical Studies of the Arts, 23,* 165–180.

Author Note

Paul J. Silvia, Department of Psychology, University of North Carolina at Greensboro.

I am grateful to Anjan Chatterjee for providing the raw data from Wilson and Chatterjee (2005, Experiment 2) for reanalysis.

Correspondence should be addressed to Paul J. Silvia, Department of Psychology, P. O. Box 26170, University of North Carolina at Greensboro, Greensboro, NC, 27402-6170. E-mail: p_silvia@uncg.edu.

Footnotes

1. Many statistical packages can estimate multilevel models. Two popular options are HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), a program designed for multilevel modeling, and SAS's PROC MIXED functions. For more information, consult Raudenbush and Bryk (2002) and Singer (1998).

2. Multilevel modeling needn't have people as the Level 2 units and responses as the Level 1 units, although this will be the most common application in the psychology of art and creativity. Any nested data structure can be analyzed with multilevel modeling, such as children within classrooms, classrooms within school districts, and members of Congress within states. Furthermore, multilevel models can have more than two levels and more than one dependent variable.

3. Although between-person confounds are avoided, researchers should be alert for within-person confounds. An unmeasured within-person variable could act as a confounding third-variable for a within-person effect.

Table 1

An example of (fictional) multilevel data. Four people viewed 8 pictures and rated each picture for preference and for complexity (1–7 scale). Responses to the 8 pictures (Level 1) are thus nested within each person (Level 2).

	Person 1		Person 2		Person 3		Person 4	
	Preference	Complexity	Preference	Complexity	Preference	Complexity	Preference	Complexity
Picture 1	2	1	5	2	1	2	6	3
Picture 2	6	7	2	3	3	4	7	2
Picture 3	2	5	3	6	4	3	3	4
Picture 4	7	3	2	5	2	3	6	4
Picture 5	1	4	7	1	4	4	5	6
Picture 6	5	4	5	3	1	1	7	2
Picture 7	2	3	6	4	3	2	7	1
Picture 8	4	4	3	6	2	3	5	1
M	3.62	3.87	4.12	3.75	2.50	2.75	5.75	2.87
SD	2.19	1.73	1.88	1.83	1.19	1.03	1.34	1.72
b	.413		-.670		.800		-.395	

Note. M = the person's average rating; *SD* = the standard deviation for the person's ratings; *b* = the unstandardized regression weight that describes the complexity–preference slope for that person.

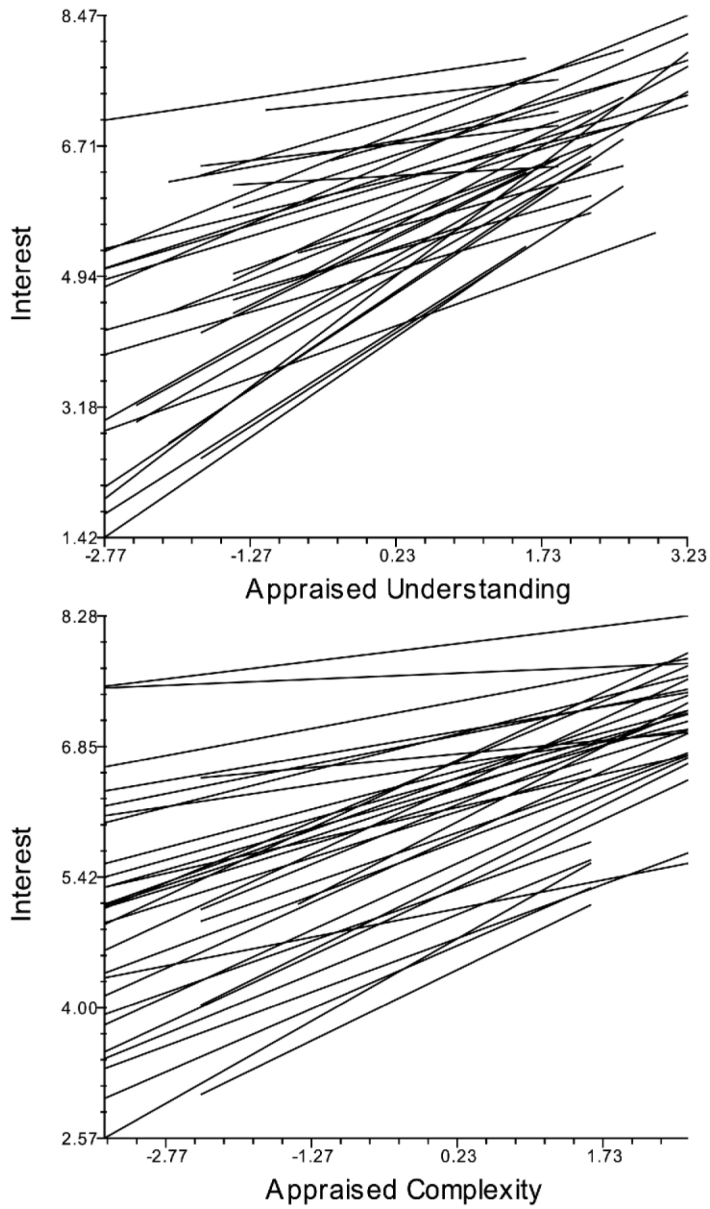

Figure 1. Distributions of within-person slopes (Silvia, 2005a).

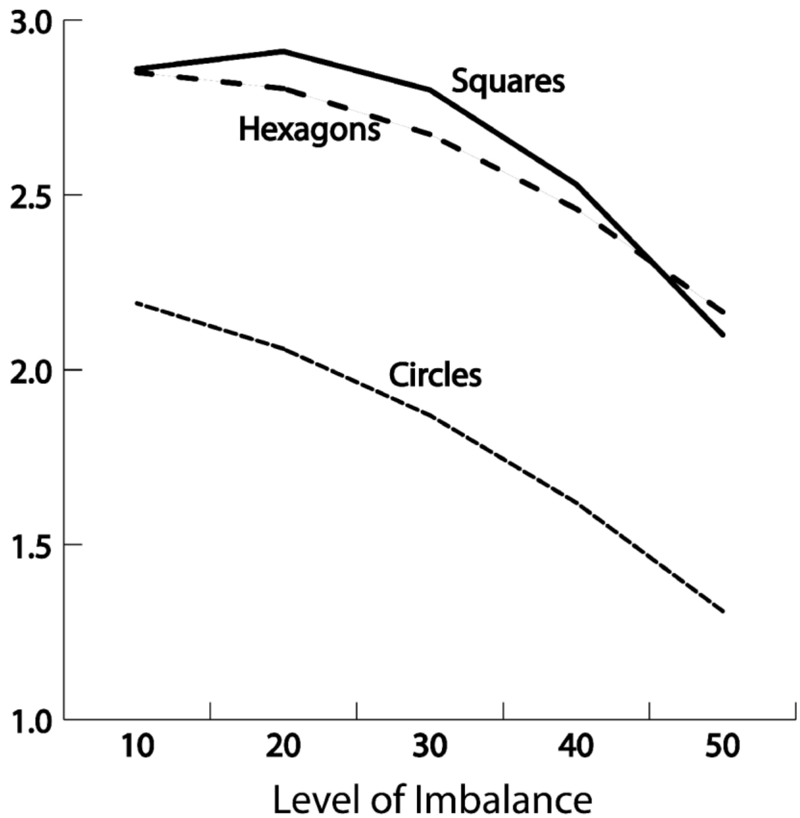

Figure 2. Within-person relationships between preference for a picture and the picture's level of imbalance. The data are from Wilson and Chatterjee (2005, Experiment 2).

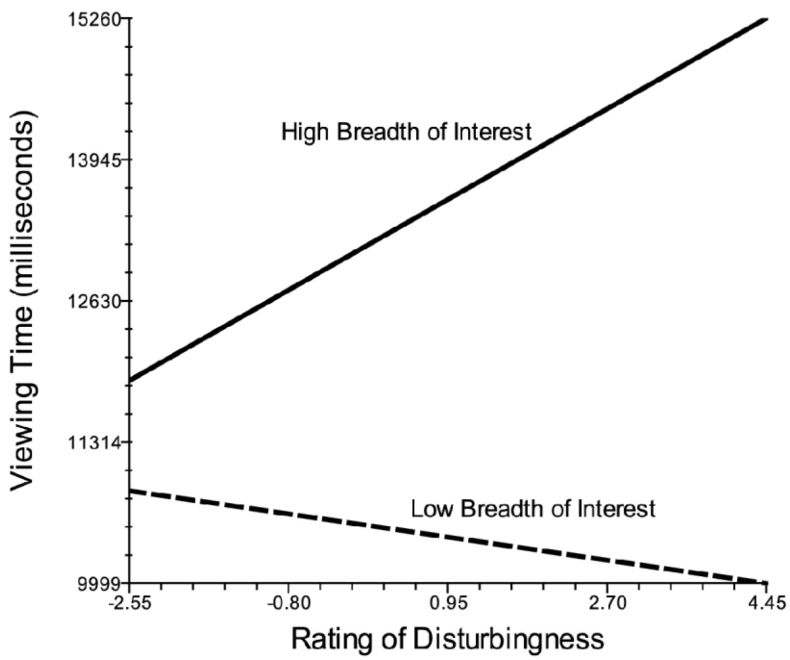

Figure 3. A cross-level interaction between breadth of interest (Level 2) and the Level 1 relation between viewing time and ratings of disturbingness. The Level 1 relationship becomes significantly stronger as breadth of interest increases (Silvia & Turner, 2006).

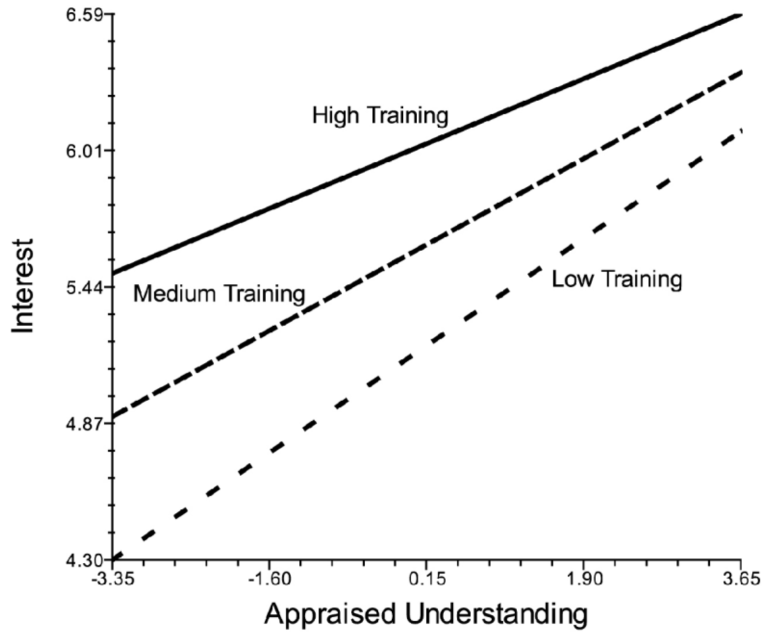

Figure 4. A cross-level interaction between training in art (Level 2) and the Level 1 relation between interest and complexity. The Level 1 relationship doesn't appreciably change across the three discrete training groups (Silvia, 2006a).