Reducing Task-related Variance in Performance Assessment Using Concept Maps

Jay Parkes\textsuperscript{1}, Dawn M. Zimmaro\textsuperscript{2}, Stephen M. Zappe\textsuperscript{2}, and Hoi K. Suen\textsuperscript{2}
\textsuperscript{1}University of New Mexico and \textsuperscript{2}The Pennsylvania State University

ABSTRACT

Task-related variance causes scores from performance assessments not to be generalizable and thus inappropriate for high stakes use. It is possible that task-related variance is due, in part, to students’ inability to transfer their knowledge consistently from one assessment task to another. Therefore, concept-mapping, a cognitive tool, might be useful to aid this transfer.

This study examines the effects of concept maps on the task-related variance components of Political Science performance assessments. On three quizzes, some students used concept maps while writing two essays, while other students did not. The task variance components remained unchanged across groups, but the person main effect components increased and the task-by-person interaction components decreased for those using concept maps. Also, the scores from the concept mapping groups had higher generalizability coefficients than for those who did not use a concept map.

A near consensus of opinion holds that reliable scores are difficult to obtain from performance assessments. Low reliability coefficients are seen as one of the major roadblocks, if not the pre-eminent roadblock, to the implementation of large-scale, high stakes performance assessments (Burger & Burger, 1994; Dunbar, Koretz, & Hoover, 1991; Linn, 1993, 1994; Mehrens, 1992; Messick, 1994). Numerous studies have demonstrated low reliability coefficients for performance assessment in diverse subject matters and populations such as behavioral observation in pre-school (McWilliam & Ware, 1994); middle school science (e.g., Shavelson, Baxter, & Gao, 1993); elementary and middle school writing and mathematics (e.g., Koretz, Stecher, Klein, & McCaffrey, 1994) and college writing (e.g., Nystrand, Cohen, & Dowling, 1993).

Address correspondence to: Jay Parkes, Educational Psychology Program, 128 Simpson Hall, University of New Mexico, Albuquerque, NM 87131-1246, USA. E-mail: parkes@unm.edu

Manuscript submitted: October 26, 1999
Accepted for publication: June 20, 2000
The most problematic source of error variance is task-related variance (Brennan, 1996; Brennan & Johnson, 1995; Green, 1995; Linn, 1993, 1994). Dozens of studies have documented the lack of agreement between scores on different tasks in performance assessment (e.g., Baker, 1992; Breland, Camp, Jones, Morris, & Rock, 1987; Brennan, Gao, & Colton, 1995; Gao, Shavelson, & Baxter, 1994; Koretz et al., 1994; Lane, Liu, Ankenmann, & Stone, 1996; Miller & Legg, 1993; Shavelson, Baxter, & Gao, 1993; Swanson, Norcini, & Grosso, 1987). Linn and Burton (1994) provided a possible path toward a solution to this problem: find ways to reduce task-related variance components.

Task-related variance components here include both the main effect variance component for task and the task-by-person interaction variance component. The former represents variation across tasks for all subjects. That is, the mean scores on several tasks are not the same. The latter represents an interaction between persons and tasks. This can occur when different persons are ranked differently by different tasks.

The problems of task and task-by-person variance could well be related to the problem of transfer of learning. As Linn, Baker, and Dunbar (1991) pointed out: “The limited generalizability from task to task is consistent with research in learning and cognition that emphasizes the situation and context-specific nature of thinking (p. 19).” Therefore, the problems of task and task-by-person variance may be found in the literature regarding knowledge transfer (cf. Lave, 1988; Perkins & Salomon, 1988).

From the literature on transfer, researchers have found that students often can solve a set of problems and seemingly understand a principle, yet they fail to employ that principle or solution in a novel situation. It seems reasonable to expect students to be able to do that, especially since transferring knowledge is an everyday part of life (Perkins & Salomon, 1988). These transfer problems sound very similar to the problem of task-related variance: differential performance on tasks where none is anticipated.

In the literature on transfer, researchers (e.g., Perkins & Salomon, 1989; Phye, 1989) have suggested that effective transfer or analogical reasoning is composed of both general and context-specific elements. In the literature on performance assessment, on the other hand, Green (1995) commented that performance tasks have little common variance and much specific variance. Given the similarity between these two descriptions, perhaps the solutions to transfer problems could apply to problems of task-related variance.

Many different solutions to transfer problems have been attempted. One particular type of solution focuses on inducing a schema, that is, activating
a generalizable knowledge base (Cooper & Sweller, 1987; Gick & Holyoak, 1983; Jelsma, Van Merriënboer, & Bijlstra, 1990; Novick, 1988; Phye, 1989; Sweller, 1989). For performance assessment, Suen, Sonak, Zimmaro, and Roberts (1997) have proposed a schema induction solution to task variance by hypothesizing that concept maps can be used to bridge the gap between tasks. The purpose of this study was to test this hypothesis.

This present investigation brought these lines of reasoning together empirically to address the issue of task-related variance. This review of literature suggests that task-related variance can be seen as a transfer problem and that solutions to transfer problems might also address the task variance problem. This present investigation is meant as an empirical check on the Suen et al. (1997) proposal: to use a concept map as a bridge between tasks on a performance assessment.

Concept maps, when based on associationist theory, are graphical representations of an individual’s knowledge framework, and they consist of nodes and labeled lines. The nodes correspond to relevant concepts in a domain and the lines express a relationship between a pair of concepts. The label on the line expresses how the two concepts are related (Shavelson, Lang, & Lewin, 1993). The lines should be labeled so that the meaning between the two concepts is explicitly expressed (Novak, Gowin, & Johansen, 1983). The combination of two nodes and a labeled line is referred to as a proposition. The combination of interconnected propositions is known as a network concept map. This was selected in place of other concept mapping techniques, such as hierarchical mapping. A sample map is included in Appendix A.

Concept mapping should provide an effective bridge across tasks for a number of reasons. First, research has shown that students with better schema are better able to transfer effectively (Catrambone & Holyoak, 1989; Cooper & Sweller, 1987; Gick & Holyoak, 1983). More broadly, visual representations also aid transfer (Beveridge & Parkins, 1987; Gick, 1985; Gick & Holyoak, 1980). Finally, techniques which help students focus on the conceptual structure of knowledge are better at enhancing transfer than techniques that focus on surface features (Bransford, Sherwood, Vye, & Rieser, 1986; Gentner, 1989; Paas, 1992). The present study is designed to employ these findings.

Therefore, using a concept map that focuses on the underlying theme or framework to be assessed while completing tasks on a performance assessment should help examinees perform more consistently. In addition, once a concept map has been used as a bridge between tasks on one performance assessment, task-related variance should continue to be small-
er even when examinees no longer have the concept map in front of them. This expectation is supported by considerable evidence that the most effective transfer occurs when many examples are provided and feedback is given to students (Cooper & Sweller, 1987; Gick & Holyoak, 1983; Lewis, 1989; Phye, 1989; Robins & Mayer, 1993). Thus, as students continue to see example solutions to similar problems and continue to gain knowledge, their performances should become more consistent and their transfer better. Similarly, research has shown that familiarity with the problem space also aids transfer (Pierce, Duncan, Gholson, Ray, & Kamhi, 1993; Weaver & Kintsch, 1992).

Two specific main hypotheses were tested in this study. First, concept mapping used as a bridge across tasks will reduce variance components associated with tasks. Second, task-related variance components should remain smaller over time, even after the bridge is removed. In addition to those predictions, other issues were also examined in this study. This study also explored the effects of the use of concept maps on the person main effect variance components, on the random error variance components as well as the effects on the generalizability coefficients.

**METHOD**

This study took place in an introductory Political Science course in American National Government at a large, northeastern university. Data from 187 students were used in these analyses. The sample consisted primarily of freshmen (37%) and sophomores (32%) and were 51% male, 38% female and 11% of an unidentified gender. The professor of the course proposes that politics can be thought of as a game with participants, rules, outcomes, resources, etc. It is from this underlying framework (the Play of Power) that the entire course is taught (Eisenstein, Kessler, Williams, & Switzer, 1996).

The primary assessment tool in the course was essay writing. Specifically, students were asked to take a newspaper article and understand and explain the situation in the article in terms of the metaphor. Two of these essays were completed as part of a quiz. In this study three of these quizzes are of interest. See Appendix B for an example quiz. Although it is true that essay tests are often claimed as performance assessments when, indeed, they are just as much a proxy for performance as a multiple choice item would be, that is not the case here. This assessment is indeed a performance of students applying their conceptual understandings of politics to a specific political occurrence. Writing about political
events is one of the most authentic ways in which consumers/observers of politics engage in politics, a level appropriate to an introductory course. If this were an upper level course for political science majors, then participating in a campaign or seeking office or running the student government would indeed be more authentic than an essay would. But that is not the case here.

Three different treatment conditions were employed in this study. In the concept mapping (CM) condition, students were trained during the first recitation section of the semester in the use of concept maps as a way to organize the Play of Power metaphor. Students were first shown a concept map and the process of concept mapping was explained to them. This was followed by a discussion of how concept mapping could help them study. Finally, they were asked to practice concept mapping on material unrelated to political science. This training lasted for approximately 30 minutes. They were then asked to create a concept map as a homework assignment and were allowed access to that concept map when working on the quiz. The mapping task was designed so it would contain the general Play of Power framework, but not specific information from the text chapters. During the quiz, students who completed the map and brought it with them were told to refer to the map when formulating their responses. The second treatment condition is related to the first. It came into existence after quiz 1 and was the previous concept map condition (PCM). These were students who had constructed and used a concept map on a previous quiz but were not using one on the present quiz. The third condition was the “no intervention,” or control, condition. Here, concept maps were neither required nor available to the students during the quiz.

Although all groups participated in both the concept mapping and control or previous concept mapping conditions, only a subset of those completed was chosen to be scored due to cost limitations. The subset was chosen to allow maximum comparisons between treatment groups while minimizing the amount of rating to be done. Those cells in the design that were not scored are visible in Table 1 because there are no estimates for them. Therefore, the following treatment assignments are relevant here. At the first quiz, group 1 was in the concept mapping condition and group 3 was in the control condition. At the second quiz, group 1 was in the previous concept mapping condition, group 2 was in the concept mapping condition, and group 4 was in the control condition. At the third quiz, group 2 was in the previous concept mapping condition and group 3 was in the concept mapping condition. Not all students in each condition completed the homework assignment given to them, so the treatment groups were defined based on what was completed.
The four groups were not created in random fashion. Rather, since there were four teaching assistants each assigned three recitation sections, the groups were defined by combining sections in a way as to minimize confusion for the teaching assistants. Therefore, students were assigned by recitation section to a treatment condition, but not in any way that should bias the treatments. This was confirmed using chi-square tests of independence, which revealed no differences in the proportion of males to females across the four intact groups \( \chi^2 (6, N = 187) = 4.748, p > .05 \) nor in the proportions of class (e.g., freshmen, sophomores, juniors, or seniors) \( \chi^2 (12, N = 187) = 11.71, p > .05 \) across the treatment groups. Additionally, the final course grades were available for the participants, and an analysis of variance across the three treatment conditions revealed no statistically significant differences \( F (3, 184) = 0.229, p > .05 \).

**Scoring the Quizzes**

The scoring rubric was developed for this particular course by a team of assessment consultants working with the course instructor. The full rubric measured four aspects of performance on a quiz: mobilizing general concepts, mobilizing specific concepts, applying concepts, and structuring an essay. Of interest in this study is the applying concepts subscale. This scale addressed the ability of students to apply the concepts from the Play of Power to the article at hand. Specifically, raters judged the degree to which the student connected, integrated, elaborated and contextualized the appropriate concepts on a 5-point likert scale. An example rubric is given in Appendix C.

Raters were recruited in Educational Psychology graduate courses and paid for their participation in the study. Three separate training and scoring sessions were conducted, one for each quiz. Three hours were devoted to training. Raters were first introduced to the course itself. Then they were introduced to the Play of Power metaphor and how it can be useful to understand politics. Once they had that basic understanding, they moved to a description of the task and the conditions under which the students were asked to write so that they understood fully the nature and constraints of the task.

The rubric was then discussed in detail and any questions or concerns addressed. In order to get acquainted with the specific tasks they would score, the raters did a quiz exercise as a group. At that point, they turned to calibrating their scoring. They read one sample essay and discussed it in general terms based on the rubric. Then they read a second sample essay and assigned scores to it. Those scores were then discussed and any points of discrepancy debated. Once they felt comfortable with their understand-
ing of that rubric, the process was repeated with the second article. The goal of the training sessions was to establish a “group think” among the raters so that they were interpreting the language of the rubric in the same way.

A 6-hour scoring session was held the next day. Pairs of raters were nested within treatment condition. The raters were not permitted to discuss their scores with one another. Each pair scored approximately 50 participants’ quizzes or 100 essays.

**Analyses**

Seven separate generalizability studies were conducted, one for each treatment condition at each quiz, using this model:

\[
\hat{\sigma}^2_{\chi} = \hat{\sigma}^2_p + \hat{\sigma}^2_t + \hat{\sigma}^2_r + \hat{\sigma}^2_{pt} + \hat{\sigma}^2_{pr} + \hat{\sigma}^2_{prt},
\]

where \( p \) indexes persons, \( t \) indexes tasks, and \( r \) indexes raters. The GENOVA software was employed to estimate the variance components and their associated standard errors. These estimates are given in Table 1. The task main effect variance components and the task-by-person variance components were all compared across the different treatments, sometimes in within-subjects comparisons where possible, but usually between-subjects. Second, the person main effect component and the random error component were also compared. Finally, the generalizability coefficients for each cell were calculated and compared.

For each comparison of interest, the following statistical approach was employed. First, a reference group was defined as the group to which the other estimate would be compared. For example, when a concept mapping group is compared to a control group, the control group becomes the reference group. That is, any CM effect exists in comparison to the control group or no treatment effect. Since the null hypothesis, then, is that both estimates were sampled from the same population, the two relevant estimated standard errors were pooled through

\[
\text{se}_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)(se_1)^2 + (n_2 - 1)(se_2)^2}{n_1 + n_2 - 2}}.
\]

Then a 95% confidence interval, built through \( CI \pm 1.96\text{se}_{\text{pooled}} \), was employed. These confidence intervals then were used to determine statistically significant differences. The pooled standard errors of all analyses are given in Table 2.
Table 1. Estimated Variance Components and Standard Errors.

<table>
<thead>
<tr>
<th>Group</th>
<th>Variance Component</th>
<th>QUIZ 1</th>
<th>QUIZ 2</th>
<th>QUIZ 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CM (N = 47)</td>
<td>PCM (N = 47)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Person main effect (p)</td>
<td>0.147 (0.074)</td>
<td>0.078 (0.056)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Task main effect (t)</td>
<td>0.023 (0.025)</td>
<td>0 (0.016)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater main effect (r)</td>
<td>0.012 (0.018)</td>
<td>0 (0.21)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Task-by-person interaction effect (pt)</td>
<td>0 (0.065)</td>
<td>0.028 (0.049)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater-by-person interaction effect (pr)</td>
<td>0.073 (0.08)</td>
<td>0.089 (0.059)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater-by-task interaction effect (rt)</td>
<td>0.003 (0.011)</td>
<td>0.033 (0.032)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random error effect (prt)</td>
<td>0.481 (0.098)</td>
<td>0.316 (0.064)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalizability Coefficient</td>
<td>0.48</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>p</td>
<td>0.133 (0.059)</td>
<td>0.069 (0.065)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0.043 (0.042)</td>
<td>0 (0.008)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>0.049 (0.048)</td>
<td>0.005 (0.015)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt</td>
<td>0 (0.051)</td>
<td>0.141 (0.06)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pr</td>
<td>0.036 (0.056)</td>
<td>0.107 (0.054)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rt</td>
<td>0.009 (0.013)</td>
<td>0.012 (0.015)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prt</td>
<td>0.353 (0.072)</td>
<td>0.243 (0.051)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalizability Coefficient</td>
<td>0.56</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>p</td>
<td>0.073 (0.059)</td>
<td>0.233 (0.089)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0.007 (0.008)</td>
<td>0 (0.019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>0 (0.001)</td>
<td>0 (0.16)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt</td>
<td>0.177 (0.053)</td>
<td>0.018 (0.071)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pr</td>
<td>0.093 (0.038)</td>
<td>0 (0.038)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rt</td>
<td>0 (0.001)</td>
<td>0.026 (0.03)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prt</td>
<td>0.156 (0.031)</td>
<td>0.429 (0.096)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalizability Coefficient</td>
<td>0.33</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>p</td>
<td>0.141 (0.073)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0 (0.008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>0 (0.007)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt</td>
<td>0.072 (0.059)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pr</td>
<td>0.078 (0.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rt</td>
<td>0.008 (0.013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>prt</td>
<td>0.315 (0.066)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalizability Coefficient</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Specific Hypotheses
A series of relationships should exist in the data if the theoretical underpinnings of the treatment is sound. First, for both the task and task-by-person variance components, the components for the CM group should be smaller than those for the control group. Second, again for both the task and task-by-person variance components, the components for the PCM group should be smaller than those for the control group. Third, although the theory is not strong enough to predict a direction, the comparison between PCM and CM on the task-related variance components will also be examined for differences. Fourth, the effects of treatment condition on both the person main effect variance component and the random error variance component will be examined. Fifth and finally, differences in
generalizability coefficients across treatment conditions will be examined. All of the relevant estimated variance components with their respective standard errors as well as the relevant generalizability coefficients are provided in Table 1.

**Hypothesis 1: Task-related variance components will be smaller for concept mappers than for control groups**

Hypothesis 1 was tested through several different comparisons. The task main effect components were compared between the concept map condition and the control condition at quiz 1 and quiz 2. Neither comparison was statistically significant. The within-subjects comparison of group 3 when they were in the control condition at quiz 1 and in the concept mapping condition at quiz 3 also showed no statistically significant difference for the task main effect component.

The same three comparisons were made for the estimated task-by-person interaction effects. At quiz 1, the concept map task-by-person estimated component was smaller than the control group’s. This same result was found in the within-subjects comparison for group 3 from quiz 3 to quiz 1. At quiz 2 the concept map estimated task-by-person interaction component was smaller than the control group estimate, though the difference was not statistically significant.

**Hypothesis 2: Task-related variance components will be smaller for previous concept mappers than for control groups**

Only at quiz 2 is it possible to compare the previous concept map condition and the control condition. For neither the task main effect nor for the task-by-person interaction effect were there statistically significant differences between the previous concept map group and the control group. The previous concept map group did have a smaller task-by-person component, though not statistically significantly so.

**Hypothesis 3: Comparing CM and PCM on task-related variance components**

Two sets of comparisons are available for hypothesis 3: quiz 2 and quiz 3. The concept mapping group served as the reference group for these comparisons. Again, no statistically significant differences were observed for either the task main effect or the task-by-person interaction effect variance components. In absolute terms, the CM condition had smaller task-by-person variance estimates than did the PCM condition. Also in absolute terms, the PCM at quiz 2 had a smaller task main effect component than did the CM condition, though at quiz 3 there were no absolute differences.
In addition to those analyses, several other relationships were observed that support the underlying assumptions of this study. Although reducing task-by-person variance through the use of concept maps is an important theoretical development, there needs to be practical significance to redistributing the variance within assessment scores. If, by reducing the task-by-person variance components, other components get larger or otherwise adversely affect the overall generalizability of the assessment scores, then there is little practical benefit. In the set of analyses to follow, the effects of treatment on the person main effect and the three way interaction (random error) component are examined as well as the effect of treatment on the generalizability coefficients for the scores.

**Hypothesis 4: Comparing the conditions on person and random error variance components**

Six possible comparisons are available for each component. The concept map condition was compared to the control condition at quiz 1 and quiz 2. Though there was no statistically significant difference for the person main effect component in either comparison, the random error component was statistically significantly larger for the concept map condition than the control condition for quiz 1.

The concept map condition (reference group) was compared to the previous concept map condition at quiz 2 and quiz 3. At quiz 2, neither the person components nor the random error components were statistically significantly different across the treatment conditions. At quiz 3, however, both components were statistically significantly larger for the concept map condition than for the previous concept map condition.

At quiz 2, no statistically significant differences were noted for either the person main effect component or the random error component when the previous concept map and the control conditions were compared. Finally, statistically significant differences were observed for both components in the group 3 within-subjects comparison of the quiz 1 control condition with the quiz 3 concept map condition. For both components, the concept map condition produced larger estimates than did the control condition.

**Hypothesis 5: Are the generalizability coefficients different across treatment conditions**

So the practical and immediate question thus becomes what effect does the use of concept mapping have on the generalizability coefficient? Looking at the variance components piecemeal has produced mixed results; an overall examination of the generalizability coefficients becomes the true
test. With this particular variance model, the g-coefficient would be given by:

\[
\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pt}^2}{n_t} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{ptr}^2}{n_t n_r}}
\]

where \(n_t\) is the number of raters reading each essay (here, \(n_t = 2\)); and \(n_t\) is the number of tasks, or essays, written by each student (here, \(n_t = 2\)).

In Figure 1 below, the g-coefficients for the three conditions are given for each quiz. Note that there is no PCM condition at quiz 1 and no control condition at quiz 3.

In all cases, the g-coefficient is larger for the concept mapping group than for the control or previous concept mapping group. And, at quiz 2, the CM is largest followed by the PCM condition and the control condition with the smallest coefficient. Thus the redistribution of variance due to concept mapping is having a positive effect. It is also noteworthy that the g-coefficients for the concept mapping condition continue to get larger.

Fig. 1. Generalizability coefficients for the three treatment conditions on each of the quizzes.
throughout time whereas the control condition g-coefficients remain relatively flat across time.

**Results Summary**

A number of findings can be synthesized from these analyses. First, no effects of any condition were observed for the estimated task main effect variance components. Second, statistically significant and/or differences in the predicted direction were observed on the task-by-person variance components when CM was compared to control. Third, mixed results were observed for effects of CM condition on both the estimated person main effect and random error variance components. Finally, the generalizability coefficients were larger in the CM condition for all three quizzes than either the PCM or control coefficients.

**DISCUSSION AND CONCLUSIONS**

This study provides empirical support for the notion that the problem of task-related variance in a performance assessment can be explained at least partially through the cognitive phenomenon of transfer, and, based on that explanation, a method for reducing it has been successful. But all of this is in need of some qualification and some context.

**Discussion**

The concept maps seemed to have the intended effect of reducing the task-by-person variance component but not the intended effect of reducing the task main effect. On the whole, this is not an entirely negative finding. The effect of the concept map is an individual difference, not a blanket group effect. That is to say that the quality and correctness of the maps and the ability of the student to employ the map while writing would both be different from student to student. Between treatment conditions, the task variance was not different, but a smaller task-by-person effect means that within each student, the difference between the two tasks was less. So each student’s own two performances were more consistent. Therefore, it seems quite reasonable to have seen task-by-person effects but not task main effects. Even with these conjectures, some drop in task main effect theoretically should have been observed and it was not.

The mixed results regarding the variance components from the Previous Concept Map condition inform the issue of exactly what the concept maps are doing. There are two potential effects of having constructed and used a concept map on a previous quiz. First, the theoretical argument behind
using concept maps was that the maps would strengthen and solidify the students’ internal cognitive network of political science concepts. The second effect was that having the concept map aided the transfer between the framework and the tasks but did not have a strong effect on the students’ internal cognitive structure. Since previous concept mappers did not perform as theory predicted when compared to control students, perhaps the second effect is the one observed here. Feedback from the teaching assistants with the course and other research conducted on the maps themselves (Zimmaro, 1998; Zimmaro, Zappe, Parkes, & Suen, 1999) seem to indicate that the students did not fully implement concept maps. More specifically, some of the maps were quite poor; and some students expressed confusion about how the maps should help them. This may also have been exacerbated by the fact that some students had as much as 2 months’ time pass between their brief training in concept mapping and their turn to use the maps for the quizzes. It seems more appropriate to make the argument that the maps were helpful “on the spot” but did not have lasting effect as they were implemented.

Another competing hypothesis is that the maps standardized the tasks by making the tasks more similar and thus limiting the generalizability of these performances to a broader domain. This also does not seem likely given that the task main effect was not different across groups. The maps brought consistency to the students’ performances, not to the tasks themselves. This should strengthen the validity of the scores and this technique rather than weaken it.

The effects of the concept maps on the person main effect and the random error term are also consistent with the theoretically assumed benefits of concept maps. Apparently, the maps introduced variation into the assessment situation in positive places. If the concept map introduces an individual difference, then larger person variance makes sense. Unfortunately, this variance also spills over into random error. In this particular case, the generalizability coefficients went up and this introduction of variance was not problematic.

The concept maps do not seem to have served as a “cheat sheet,” however. In a study of the validity of the concept maps, Zimmaro et al. (1999) report no mean differences across treatment conditions, with the exception that the control group scored lower on Quiz 2 than both the outline and concept map conditions.

Not only were the generalizability coefficients larger for concept mappers than for non-concept mappers but also, across time, the g-coefficients rose for each successive group of concept mappers. This might be the result of instruction in the course. As was mentioned previously, the course
focused on teaching this framework, so that all students - not just the concept mappers - were learning this framework. They also became more practiced at the tasks as time went on, a factor which Linn and Burton (1994) list as helping increase the generalizability of the scores. But if instruction and practice were the sole explanations, the same increase should have been seen in the control groups. It was not, which means that the redistribution of variance due to the concept maps actually bolsters the generalizability of the scores.

In the end, the generalizability coefficients for these scores are still quite low. Relying on only one quiz, these scores would not be sufficient to make high stakes decisions about these students. It seems, therefore, that working from the supposition that task-related variance involves transfer and using concept maps to reduce task-related variance is not yet robust enough to have tangible, practical benefits. That concept maps might affect the generalizability of performance assessment scores, though, is worth pursuing and additional work to strengthen the effects so that they would be practically important seems warranted.

The major limitation to this investigation is the relatively small sample size within each cell of the design. This issue has several ramifications. First, it probably contributed to the small and negative variance component estimates shown in Table 1. These estimates, in turn, complicated the procedure for doing statistical significance testing on the estimates because a number of techniques could not be computed on such small or negative estimates. Second, the small sample sizes means the results must be cautiously generalized.

The relatively weak effect that the concept maps apparently had on the cognitive structure of the students is due to two factors. First, students were all trained at the beginning of the semester for 25 minutes in concept mapping techniques. For some of them, it was 2 months before they got to use those techniques. It would have been much more beneficial to the students to be able to construct their maps immediately after training. Second, the maps could have been better embedded in the instruction. Students were not reminded often about what the maps should do for them. They were not helped to use them or encouraged to make them a part of their studying for the course. Had the maps been embedded in the course more, the stronger cognitive effects hypothesized might have been observed.

Conclusions
This study has taken an important step forward in addressing the issue of task-related variance in performance assessments. In order to minimize
task-related variance, some explanation for it was offered, and, from that explanation, some techniques developed and tried. This study posited that poor transfer could explain some of the task-related variance. This supposition seems to have been validated in that the technique of concept mapping did reduce task-by-person variance in some instances. That opens the door for more work on manipulating transfer on performance tasks.

Furthermore, concept maps seem to be effective tools for reducing task-by-person variance, but not by some method such as standardizing the tasks or weakening the complex nature of the tasks which would lead to under-specification of the domain to which the scores would generalize. Nor is it necessary to rely only on increasing the number of tasks to attain “acceptable” levels of generalizability, a potentially expensive and prohibitive approach.

The implications are not for reliability alone; there are validity implications as well. Concept maps seem to bring consistency to the student’s performance, which is to say, reduces construct-irrelevant variance while increasing construct-relevant variance. Concept maps seem able to minimize context specificity of tasks and help the student focus on the underlying conceptual framework. So scores from those who had used a concept map should actually be better indicators of what students can do and fewer such tasks would be needed to achieve that goal.

This approach to task-related variance is notable because it asks why the variance is present and then reduces that variance based on the explanation. This study suggests that task-related variance in performance assessment might be a transfer issue. Therefore, solutions to transfer issues might help reduce task-related variance components in performance assessment.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the cooperation of Professor James Eisenstein in whose class the study was conducted. Funding and in-kind resources for this research were provided by the Penn State College of Education Alumni Society and the Schreyer Institute for Innovation in Learning.

REFERENCES


APPENDIX A

Sample Student Concept Map of the Play of Power Framework.

Note: This is an electronic recreation of an actual student map. For the study, the students created the maps free-hand, but it has been electronically recreated here for publication.
APPENDIX B

The quiz essay prompt.

Please read the article reproduced below carefully. Then answer the following question, spending about 20 minutes total.

What general concepts and specific information about politics covered in class or in the textbook can you use to understand what is described in the article? For each concept or piece of information, discuss briefly how it applies to a specific paragraph in the article (refer to the paragraph number). Make sure you confine your answer to a discussion of how the concepts and information can be applied. Do not merely restate the facts and events presented in the article.
Sample Scoring Rubric

Application – The degree to which the writer connects, integrates, elaborates, and contextualizes concepts. That is, how well did the writer:
- define concepts?
- explain concepts?
- show the relationship between concepts?
- give details and examples from the article of a specific concept (citing paragraph numbers)?
- show how general concepts play out in the specifics of the article?

To what extent did the writer do these things?

1 2 3 4 5
never rarely sometimes frequently extensively