A Hierarchical Linear Model of Variability in Test Motivation Across Students and Within Students Across Tests

Abigail R. Lau
Dena A. Pastor

James Madison University
Center for Assessment and Research Studies

Paper to be presented at the October 2007 meeting of the Northeastern Educational Research Association.

A Hierarchical Linear Model of Variability in Test Motivation Across Students and Within Students Across Tests

Introduction

Examinees are less likely to give their best effort on the test when their performance on the test has little or no consequences for them personally (Wise & DeMars, 2005). This reduced motivation level is understandable given the low-stakes conditions for examinees. However, the behavior is problematic because it poses a threat to the validity of the data collected from the test. Thus, understanding examinee behavior in low-stakes conditions is critical for establishing the trustworthiness of the data collected in the low-stakes testing conditions that occur in program evaluation, instrument development, research participant-pool, and some federally mandated educational testing programs (e.g. NAEP).

The problem of unmotivated examinees is unlikely to be solved completely due to the inherent nature of individuals. However, measurement researchers could identify characteristics of examinees, items, and tests that engender motivation. Encouragingly, some variables have been found to inspire people to engage in more effort on low-stakes tests (Sundre & Moore, 2002, Sundre & Kitsantas, 2004). For example, in a low-stakes proctored testing session, undergraduates tend to exhibit more effort when they have attentive enthusiastic proctors and are in smaller classrooms. Ultimately, these findings and others like them, could be used to design testing programs which maximize motivation in addition to maximizing measurement precision. More research on factors related to effort on low-stakes tests is needed to approach this goal.

Between Examinee Variability in Test-taking Motivation

Existing studies on examinee motivation have documented a substantial presence of unmotivated examinees. In a low-stakes program assessment testing context, Swerdzewski, Dainis, Finney and Harmes (2007) found that only 40% of undergraduate students put forth effort for all seven tests administered in a testing session. In a similar testing context, Wise and DeMars (2006) found that approximately 5% of examinees did not put forth effort on 90% or more of the items on a test (Wise & DeMars, 2006).

Furthermore, Swerdzewski et al. (2007) found that examinees who are less motivated may represent a qualitatively different type of person than examinees who tend to give good effort. For example, in a higher-education assessment context, male business majors who self-report attending college to obtain a prestigious job are more likely to be unmotivated examinees. Thus, it seems that people may differ in their overall willingness to engage in low-stakes tests, and the differences may be related to individuals' demographic and background characteristics.

Within Examinee Variability in Test-taking Motivation

Regardless of an examinees characteristics, few examinees try on all items for all tests administered in a low-stakes testing session (Swerdzewski et al., 2007; Wise & DeMars, 2006). Examinees appear to adopt different profiles of motivation across the tests. Some examinees try on 90% or more of the items on all tests, while other examinees only provide sufficient effort on developmental tests but not achievement

tests. Finally, another set of examinees did not engage in effort on any test (Swerdzewski et al., 2007).

Furthermore, examinees are more likely to try on certain kinds of items than other types of items. For example, items with graphs are more likely to receive examinee effort (Wise, Kong & Pastor, 2007), where as items that require open responses are less likely to engender optimal motivation (Sundre & Anastasia, 2001). Thus, examinees appear to be discriminating in how they approach exerting effort on low-stakes exams. However, more research is needed to explore which testing characteristics are associated with increased motivation and which are associated with lower motivation.

Purpose of the Study

Research to date has shown that low-motivated examinees exist and that there is variability in motivation across and within examinees. Hierarchical Linear Models (HLM) can be used to model this variability and estimate the variance components at each of two levels, with level 1 being the test level and level 2 being the student level. For example, HLM can model how students differ in their test motivation across tests, and how students differ from one another in their overall motivation level. Furthermore, test-level and student-level characteristics can be incorporated into the model to try to explain the variance components at each level. One advantage of HLM models over regular multiple regression models is that the multi-level nature of the models accounts for dependency in the data. In this case, the dependency would be that of test administrations within students. More specifically, students' level of motivation on each test they receive is conceptualized as being an observation from a population of possible observations of their overall tendency to try on tests.

The present paper is a description of an HLM study of how examinee and test characteristics relate to motivation on a low-stakes test. Specifically, test-type (achievement versus developmental) was used to explain within-person variability in motivation, and two examinee-level variables were isolated to account for between-person variability in motivation. It was expected that lower motivation would be observed for achievement tests than developmental tests for two reasons: (a) less concentration are needed to respond to developmental items and (b) developmental items only require interest in and knowledge of one's self, not of the content area being assessed by the achievement tests. Examinee level variables were gender and Scholastic Aptitude Test (SAT) scores. Gender was examined because previous research has found gender differences in examinee motivation, with males having lower motivation than females (Eklöf, 2007; Wise, Kingsbury, Thomason, & Kong, 2004). Previous research has been inconclusive about the effects of SAT, with some studies finding no relationship between academic ability and motivation (Kong et al., 2006; Wise & DeMars, 2006; Wise & Kong, 2005) and others finding a significant positive relationship (Wise, Kong & Pastor, 2007). Because the relationship between academic ability and examinee motivation may be dependent on test-type, the interaction of SAT with test-type was of particular interest in the current study.

The specific research questions (arranged into three sets) and are described in more detail below.

Research Questions: Set 1. Do examinees differ from one another in effort? Does examinees' effort differ across tests?

An unconditional HLM model was used in the present study to quantify the extent to which examinees vary from one another in their tendency to engage in low-stakes tests. This model was also used to examine how consistent examinees are (within themselves) in their response time effort across different tests.

Research Questions: Set 2. Are there differences in the amount of effort examinees put forth on achievement versus developmental tests? Is there examinee to examinee variation in the differential amount of effort being put forth on these two types of tests?

The level-1 predictor of test-type was added to the unconditional model to determine whether there was a significant difference in the amount of effort put forth on achievement versus developmental instruments and if this relative difference in effort varied substantially across examinees

Research Questions: Set 3. Can differences among examinees in effort on achievement and developmental instruments be explained by gender or academic ability?

Predictors were then added at level-2 to investigate if gender or academic ability (as measured using SAT) could explain examinee differences in motivation on the achievement and developmental instruments.

Methods

Participants and Procedures

Examinees were 843 randomly selected students with 45-70 credit hours (typically second semester sophomores) at a midsized southeastern university who completed a battery of achievement and developmental tests during a required three hour computer-based testing session in the spring of 2006. The testing was considered low-stakes because although students were required to participate in the testing, their individual scores on the test were not reported, and there were no consequences or rewards for their performance on the tests. Examinees were approximately 64% female, 80% Caucasian, 25% third-year students, 75% second-year students.

Instruments

Examinees completed a battery of seven instruments, two of which were multiple-choice achievement tests and five of which were Likert-type developmental scales. One achievement test measured student's knowledge and ability in Fine Arts (114 items), and the other measured their scientific and quantitative reasoning skills (42 items). Developmental scales included a 60-item instrument measuring students' goal orientation, attitudes toward intelligence, and meta-cognitive awareness, a 45-item scale measuring attitudes toward diversity, a 30-item scale measuring student-worry, a 25-item scale measuring student social self-efficacy, and a 66-item scale assessing students' reasons for attending college. Scores on the specific tests were not computed for this study, since the study focused on student's motivation on the test rather than their performance on the test. Thus, the psychometric details of the individual instruments are not reported.

Measurement of Examinee Motivation

Examinee motivation was measured using the response time effort index (RTE, Wise & DeMars, 2005; Wise & DeMars, 2006). Test RTE is defined as the proportion of items on the test for which an examinee engaged in solution behavior. Solution behavior is defined as spending more than a minimum amount of time responding to an item. The minimum amount of time (i.e. the response-time threshold) was set for each item individually based on how quickly a person could possibly read and process the item. If an examinee took longer than the response time threshold to respond to an item, the examinee is given credit for solution behavior for the item (Kong, Bhola, & Wise, 2005). The frequency of solution behavior is summed across all items on a test, and divided by the total number of items to obtain the test RTE score. RTE scores have been shown to have high reliability and to correlate with performance on the test and other measures of examinee motivation (Wise & DeMars, 2006). RTE scores were computed for each of the seven tests in the battery of assessments. Thus, each examinee has seven RTE scores which can range from 0 to 1 and indicate examinee motivation on a given test.

Data analysis

All HLM models were analyzed using restricted Maximum Likelihood estimation (RML) in SAS 9.0. RML was chosen instead of full Maximum Likelihood because to avoid the negative bias of the variance components associated with the latter estimation procedure. However, because RML was used deviances of the models can only be compared when they differ in random effects, not when they differ in random and fixed effects.

HLM Models

The specific model equation used to answer each set of research questions are shown in the Appendix. The first set of research questions was examined using a random-effects ANOVA model (unconditional model), which estimated the amount of between (τ_{00}) and within (σ^2) person variance in RTE scores. These variance components were used to compute an intra-class correlation coefficient (ICC) which indicates the proportion of total variance in RTE scores that is between students. In other words, the ICC is the variance in RTE scores that is due to the fact that RTE scores are nested within students. The unconditional model also indicates the range of student mean RTE scores in the population and the overall grand mean of RTE scores across students.

To answer the second set of research questions, the Level 1 predictor of test-type was added to the model, with intercepts and slopes at level 2 free to vary and covary. If the variation or covariation of the level 2 random effects was not statistically significant, the model was re-estimated constraining the previously non-significant effects to zero. To determine which random-effect variance components (i.e. slope variation, interception variation, or slope-intercept covariation) was statistically significant, model comparisons were made in which one variance component was constrained to zero at a time.

To answer the third set of research questions, gender and SAT were added as predictors at level 2. Gender and SAT were only added as predictors to those level 2 effects associated with significant variance components in the final model used to answer the second set of research questions.

Results

Research Question: Set 1. The parameter estimates for the unconditional model are shown in Table 1. The intra-class correlation coefficient (ICC) computed from the variance components estimated in the unconditional model indicated that 61% of the variance in RTE scores is between students, which is an amount statistically greater than zero. Although proportionately there is more variability in motivation between students than within students, the extent to which motivation varies across tests within students is still a sizeable percentage (39%) of the total variability.

The unconditional model estimate of the grand mean motivation score (across all tests and all students) was .793. Given sampling variability, plausible values for the grand mean fall between .78 and .81. Thus, we are fairly confident that on average students engaged in solution behavior for approximately 80% of the items on a given test. However, individual students' tendency to engage in low-stakes tests, (i.e. their within student average motivation score across tests) varied wildly. Plausible values for examinees' motivation scores in the population range from .33 to the maximum of 1.0[1]. Students' average motivation scores indicate the proportion of solution behavior a student provides for low-stakes tests on average. Thus, students range from providing about 30% solution behavior on average, to providing 100% solution behavior on average for a test.

Table 1
Parameter estimates for Research Question: Set 1 model

Fixed Effect	Coefficient	*SE*	*t*	
Intercept, γ_{00}	0.793	0.008	94.07*	
Random Effect	**Variance Component**	***SE***	χ^2	***df***
Intercept, u_{0j}	0.055	0.003	3334.3*	1
Level 1 effect, r_{ij}	0.035	0.0007		

* $p < .0001$

Research Questions: Set 2. To account for within student variability in motivation (level 1 variance) across tests, test-type was added as a level-1 predictor variable. First, test-type was added as a level 1 predictor with slopes (the relationship between test-type and motivation) and intercepts (the average motivation across all tests) allowed to vary and covary freely. The variation in test-type slopes and intercepts was statistically significant across people. However, the covariance between intercepts and test-type slopes was not significant (τ_{01} = -0.00048; χ^2 = .1, df = 1, p = .75), so this covariance of test-type slopes and intercepts was constrained to zero to simplify subsequent models.

The parameter estimates for the model constraining the covariance to zero are shown in Table 2. Motivation scores for achievement tests were significantly different then motivation scores for developmental tests. Specifically, the grand slope was -.08,

[1] The maximum plausible value was 1.25. Because this exceeded the maximum possible value of 1.0, it was interpreted as equaling 1.0.

which indicates that motivation scores for achievement tests are about .08 percentage points lower than developmental tests. The estimated average motivation score for developmental tests was .82, whereas the estimated average motivation score for achievement tests was .74. Including test-type as a predictor in the HLM was deemed practically meaningful because the random effect for test-type explained 14% of variation in motivation scores across tests within examinees. The significant variance components indicated substantial variation across examinees in effort on developmental tests and achievement tests, with somewhat larger examinee variation in effort put forth on developmental tests relative to achievement tests.

Table 2
Parameter estimates for Research Question: Set 2 model

Fixed Effect	Coefficient	*SE*	t	
Intercept, γ_{00}	0.815	0.008	96.50*	
Test-type, γ_{10}	-0.080	0.007	-11.96*	
Random Effect	Variance Component	*SE*	χ^2	*df*
Intercept, u_{0j}	0.054	0.003	3313.7*	1
Test-type, u_{1j}	0.014	0.002	104.6*	1
Level 1 effect, r_{ij}	0.030	0.0007		

* $p < .0001$

Research Questions: Set 3. The model above was expanded upon by adding level 2 predictors of between examinee variability in motivation on both achievement and developmental tests. The results for this final model are presented in Table 3.A significant relationship was found between gender and examinee motivation on development tests, with male examinees motivation scores being 10% lower than female examinees. Gender explained 4% of the variation among examinees in test-taking effort on the developmental tests. The effect of gender on RTE did not significantly differ across test type. The predicted RTEs by gender and test type for students with SAT scores equal to 1000 are shown in Figure 1. The figure illustrates nicely how motivation is lower for achievement tests relative to developmental tests and for males relative to females.

The relationship between SAT and examinee motivation on development tests was positive, but not statistically nor practically significant (e.g., less than 1% of the variance in effort on developmental tests was attributable to SAT). The effect of SAT on RTE did not significantly differ across test type.

Discussion

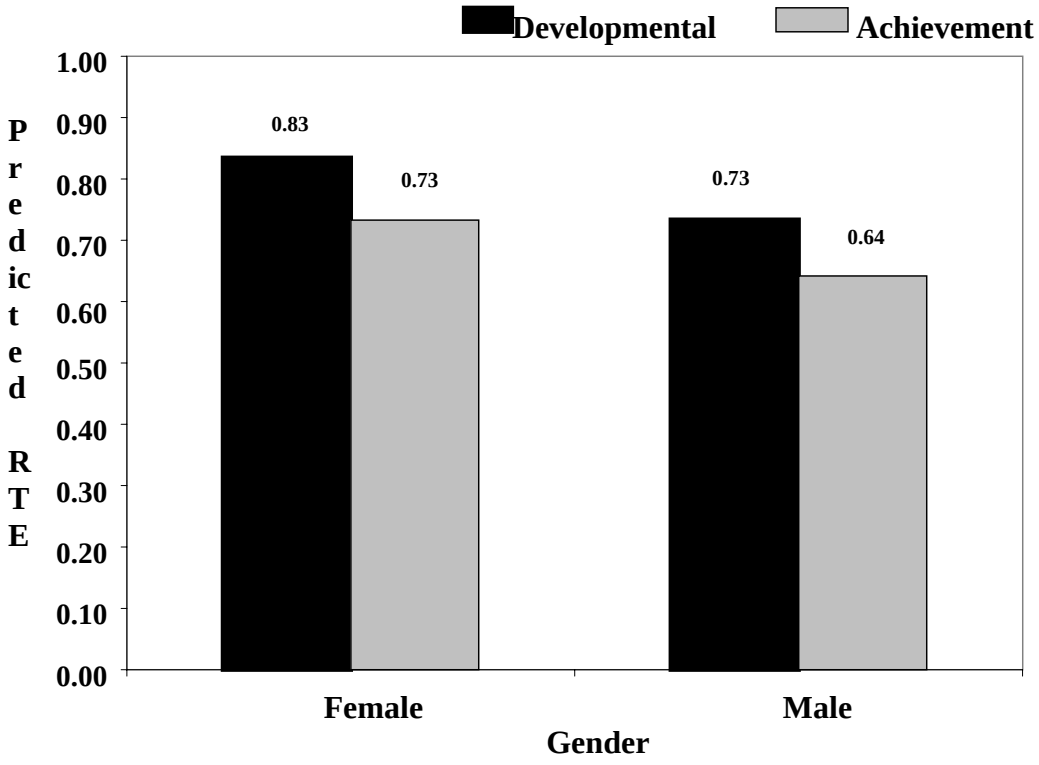
In this study, examinees varied quite a bit from one another in their tendency to engage in low stakes tests. Furthermore, each examinee exhibited sizeable variability in their effort across different tests. Overall, there was more variability in test-taking motivation across examinees (61%) than within examinees (39%). Within examinees, variability in motivation was related to test-type, with achievement tests associated with lower motivation than developmental tests. This finding is not unexpected given that

Table 3
Parameter estimates for Research Questions, Set 3 model

Fixed Effect	Coefficient	*SE*	*t*	*p*
Intercept, γ_{00}	0.694	0.086	8.05	<.0001
Gender, γ_{01}	-0.101	0.017	-6.02	<.0001
SAT, γ_{02}	0.0001	0.000	1.91	0.056
Test-type, γ_{10}	-0.211	0.070	-3.03	0.003
Gender*Test-type, γ_{11}	0.010	0.014	0.75	0.455
SAT*Test-type, γ_{12}	0.0001	0.000	1.81	0.070

Random Effect	Variance Component	*SE*	χ^2	*df*
Intercept, u_{0j}	0.052	0.003	3190*	1
Test-type, u_{1j}	0.014	0.002	104.1*	1
Level 1 effect, r_{ij}	0.030	0.0007		

Figure 1. Predicted RTEs by gender and test type.



achievement items require more interest in and knowledge of a specific content area than developmental tests. Although test-type appears to be a salient predictor of within-examinee motivation (explaining 14% of the within-examinee variance in motivation), there still remains a large amount of variance left to be explained. Specifically, 86% of the within examinee variance remains once controlling for test-type. It is recommended

that future studies incorporate a more comprehensive set of test-level predictor variables in an attempt to explain the large amount of remaining variance.

Examinees differed quite substantially from one another in motivation on both the developmental and achievement tests. Males put forth less effort than females on both types of tests, with difference between male and female effort being the same across test type. Although a statistically significant effect was found for gender, the practical significance of the predictor was small. The minor effect for gender coupled with the lack of effect for SAT necessitates further research investigating a large number of examinee characteristics that can be used to explain effort on low-stakes tests.

Implications of the study

The results from this study support the idea that students vary in their general level of motivation for low-stakes tests, behave differentially across tests, and adjust their level of motivation according to test type. In other words, an examinee may tend to exert high effort, but may also exhibit slightly less or more motivation depending on a particular test. Implications for these findings include that it may be reasonable to take into account whether an examinee had high or low motivation when interpreting test scores, and that test-taking motivation may be alterable despite individual tendencies.

Further, the fact that test-type relates to motivation on a test suggests that test-characteristics are a factor in examinee motivation. If future research continues to uncover test characteristics that are related to higher motivation, these factors could be used to craft test specifications for instruments that are designed to engender maximum motivation in low-stakes assessments contexts.

This study also indicated that examinee characteristics play a role in effort put forth on low-stakes tests. Knowing which examinees are more likely to put forth less effort is useful for targeting motivational interventions to those most in need of the intervention.

Future research

The current study demonstrated the effectiveness of using HLM to model examinee motivation across a battery of tests. Future research aimed at understanding examinee behavior in low-stakes contexts may benefit from using an HLM approach as well. For example, additional models could be proposed that utilize other examinee and test variables. Specifically, it may be fruitful to examine the effect of students' year in school, major area of study, or personality and attitudinal characteristics on test-taking motivation. Also, additional test types could be examined by including constructed response instruments or instruments that utilize innovative items. In general repeating the study with different assessment instruments and a wider array of predictors would be useful.

In addition, the study could be expanded to include additional levels in the HLM model. For example, previous studies have found that proctors have a significant impact on examinee motivation. Thus, modeling the nesting of students within testing rooms may be interesting. Alternatively, the current HLM model of tests and students could be combined with the Wise, Kong and Pastor (2007) HLM model of items and tests to create a 3-level HLM model reflecting item, test, and students characteristics.

Conclusions

Administering tests in a low-stakes context is an integral part of many different scholarly endeavors. Examinees may behave very differently under low-stakes testing conditions than they do in high-stakes testing contexts. Thus, understanding the psychology of the examinee in a low-stakes testing context is imperative to tailoring instruments and test administration procedures for low-stakes assessments. While the current study examined only a few of the many variables that could be affecting examinees motivation, it showed the utility of an HLM approach to these research questions, and provided results that can guide future research.

References

Dainis, A. M., Swerdzewski, P. J., & Harmes, J. C. (2007, April). *The effect of innovative item placement on computer-based test motivation and performance.* Poster session to be presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7,* 311-326.

Kong, X. J., Bhola, D. S., & Wise, S. L. (2005). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education,* Montreal, Quebec, Canada.

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April). *Motivational effects of praise in response time-based feedback: A follow-up study of the effort-monitoring CBT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Raudenbush, S. W. & Bryk, A.S. (2002). Hierarchical Linear Models: Application and Data Analysis Methods 2nd ed. Sage Publications, Thousand, Oaks, CA.

Swerdzewski, P. J., Dainis, A. M., Finney, S. J., & Harmes, J. C. (2007, April). *Skipping the test: Using evidence to inform policy related to those students who avoid taking low-stakes assessments in college.* Poster session to be presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14* (1), 8-9.

Sundre, D. L. & Kitsantas, A. L. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29* (1), 6-26.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wise, S. L. , Kong, J. K., and Pastor, D. A. (2007) *Understanding correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated model. *Journal of Educational Measurement*, *43*, 19-38.

Appendix

Model Equations

Research Questions: Set 1

$$Y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$r_{ij} \sim N(0, \sigma^2)$$
$$u_{0j} \sim N(0, \tau_{00})$$

Research Questions: Set 2

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Test-type)_j + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$r_{ij} \sim N(0, \sigma^2)$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{01} & \tau_{11} \end{bmatrix}$$

Note. In final model for this set, τ_{01} was constrained to zero.

Research Questions: Set 3

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Test-type)_j + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(gender) + \gamma_{02}(SAT) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(gender) + \gamma_{12}(SAT)u_{1j}$$

$$r_{ij} \sim N(0, \sigma^2)$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{01} & \tau_{11} \end{bmatrix}$$

Note. In final model for this set, τ_{01} was constrained to zero.