Beaubien, J. M., Hamman, W. R., Holt, R. W., & Boehm-Davis, D. A. (2001). The application of hierarchical linear modeling (HLM) techniques in commercial aviation research. *Proceedings of the 11th Annual Symposium on Aviation Psychology*, Columbus, OH: The Ohio State University Press.

THE APPLICATION OF HIERARCHICAL LINEAR MODELING (HLM) TECHNIQUES TO COMMERCIAL AVIATION RESEARCH

J. Matthew Beaubien
American Institutes for Research
Washington, DC

Capt. William R. Hamman
United Air Lines
Denver, Colorado

Dr. Robert W. Holt
George Mason University
Fairfax, Virginia

Dr. Deborah A. Boehm-Davis
George Mason University
Fairfax, Virginia

ABSTRACT

Given the inherently hierarchical nature of organizational reality, researchers have always been interested in how to best analyze data that reside at multiple levels of analysis (e.g., at the individual and crew level). Recently, a class of statistical techniques has been developed for just this purpose. The most popular of these techniques is known as Hierarchical Linear Modeling (HLM). The purpose of this paper is to provide an introduction to HLM, to describe HLM's statistical assumptions, to speculate about the effects of violating these assumptions, and to propose viable solutions for aviation researchers so that they can apply this versatile statistical technique in their own research.

AN INTRODUCTION TO HLM

Hierarchically-nested data are ubiquitous in commercial aviation research. For example, pilots are nested within crews, crews are nested within domiciles, and domiciles are nested within fleets. Longitudinal and time-series data can also be considered a special form of nested data. For example, when multiple observations are available for individual pilots, and the number of observations varies across pilots, the observations can be considered nested within pilots (Bryk & Raudenbush, 1992).

Prior to the development of techniques such as HLM, there were two main ways to analyze multi-level data. Unfortunately, both techniques violate a number of critical statistical assumptions. In the paragraphs that follow, we compare and contrast these approaches with HLM. For the sake of simplicity, we will use an example that has two independent variables (situational awareness and crew cohesion) and one dependent variable (maneuver proficiency). Situational awareness and maneuver proficiency are assumed to be individual-level variables, while cohesion is assumed to be a crew-level variable.

Approach #1 (Individual Level of Analysis)

This approach is conceptualized and analyzed entirely at the individual level of analysis. All analyses are calculated based on the total number of pilots. However, because cohesion is a crew-level variable, each pilot receives his/her crew's score. Unfortunately, this approach violates the ordinary least squares (OLS) assumption that all observations are independent of one another (Berry, 1993).

Previous research suggests that even small degrees of non-independence can lead to biased parameter estimates (Bliese, 1998; Ostroff, 1993). Moreover, because the number of pilots necessarily exceeds the number of crews, the standard errors for crew-level variables will be underestimated, thereby leading to spuriously high levels of statistical significance.

Example #1 (Individual Level of Analysis)

When using OLS regression, each crew's cohesion score is assigned to both pilots within that crew. For example, if Crew A had a high level of cohesion (4.5 on a 5-point scale), both members of Crew A would receive their individual scores for situational awareness and maneuver proficiency, but the crew's score of 4.5 for cohesion. Similarly, if Crew B had a low level of cohesion (2.5 on a 5-point scale), both members of Crew B would receive their individual scores for situational awareness and technical proficiency, but the crew's score of 2.5 for cohesion. An example data file might look something like this:

CREW	POSITION	SA	COHESION	PROFICIENCY
A	CAPT	4.0	4.5	2.0
A	FO	2.0	4.5	3.0
B	CAPT	3.0	2.5	4.0
B	FO	2.0	2.5	2.0
C	CAPT	2.0	3.0	5.0
C	FO	4.0	3.0	4.0

The OLS equation that represents this relationship would appear something like this:

$$\text{Proficiency} = b_0 + b_1 \text{ Sit. Awareness} + b_2 \text{ Cohesion} + r$$

This is a typical OLS regression equation. The dependent variable, individual maneuver proficiency, is modeled as a function of three factors: individual situational awareness (b_1), crew cohesion (b_2), and a residual term (r) that represents errors of prediction. The intercept (b_0) is merely a statistical necessity; it represents the "average" level of maneuver proficiency for a hypothetical pilot with zero situational awareness and whose crew has zero cohesion. As stated earlier, because the observations are non-independent and calculated using the wrong sample size, the paramater estimate and standard error for cohesion are likely to be biased.

Approach #2 (Crew Level of Analysis)

One way to avoid the previously-described problem is to use data at only one level of analysis. This approach is conceptualized and analyzed entirely at the crew level. All analyses are calculated based on the number of crews. However, because situational awareness and maneuver proficiency are individual-level variables, each crew receives the mean of its pilots' scores on these variables.

While this technique obviates the problems associated with non-independence, it creates an entirely different set of problems. For example, by aggregating all variables to the crew level, meaningful within-crew variance is ignored, thereby precluding the detection of theoretically valid relationships at the individual level. In addition, because the number of crews is necessarily smaller than the number of individual pilots, this approach can result in low levels of statistical power. Finally, when individual level variables are aggregated to the crew level, the researcher may be left with variables of questionable construct validity (Hofmann, Griffin, & Gavin, 2000).

Example #2 (Crew Level of Analysis)

When using OLS regression to examine the effect situation awareness and crew cohesion on maneuver proficiency, all individual-level variables are first aggregated to the crew level. For example, if Crew A has two pilots, and their individual levels of situational awareness are 4.0 and 2.0 (on a 5-point scale), the crew as a whole would receive a score of 3.0. Similarly, each crew would be assigned its respective means for situation awareness, cohesion, and maneuver proficiency. An example data file might look something like this:

CREW	POSITION	SA	COHESION	PROFICIENCY
A	BOTH	3.0	4.5	2.5
B	BOTH	2.5	2.5	3.0
C	BOTH	3.0	3.0	4.5

The OLS equation that represents this relationship would appear something like this:

$$\text{Proficiency} = b_0 + b_1 \text{ Sit. Awareness} + b_2 \text{ Cohesion} + r$$

Again, this is a typical OLS regression equation. The dependent variable, maneuver proficiency (aggregated to the crew level), is modeled as a function of three factors: aggregated individual levels of situational awareness, crew cohesion, and random error. As before, the intercept is a statistical necessity that represents the "average" level of maneuver proficiency for a hypothetical crew with zero situational awareness and zero cohesion. As stated earlier, because all estimates are based on the total number of crews, meaningful within-crew variance is lost and statistical power is reduced.

Approach #3 (Hierarchical Linear Modeling)

A third approach is to use a multi-level variance decomposition technique such as HLM that simultaneously performs both individual (level-1) and crew level (level-2) analyses. At the individual level, HLM calculates a separate OLS regression equation for each crew between the individual-level predictor(s) and the individual-level criterion. Because this is a traditional OLS regression, there will be an intercept term and one or more slope terms, depending on the number of predictors. Because there are multiple crews, there will most likely be between-crew variance in these intercepts and slopes.

Next, HLM uses the intercepts and slopes from the individual-level model as dependent variables in a subsequent crew-level analysis. In the level-2 analysis, crew level variables are used to predict the level-1 intercepts and slopes using the expectation maximization (EM) algorithm. When the level-1 intercept is used as the dependent variable, the analysis becomes very similar to a hierarchical regression of main effects. When the level-1 slope is used as the dependent variable, the analysis becomes very similar to a moderated regression.

Example #3 (Hierarchical Linear Modeling)

Unlike the previous two techniques, HLM uses all variables in their original form. Going back to our previous example, situation awareness and maneuver proficiency remain at the individual level, while cohesion remains at the crew level. An example data file might look something like this:

CREW	POSITION	SA	COHESION	PROFICIENCY
A	CAPT	4.0	4.5	2.0
A	FO	2.0	4.5	3.0
B	CAPT	3.0	2.5	4.0
B	FO	2.0	2.5	2.0
C	CAPT	2.0	3.0	5.0
C	FO	4.0	3.0	4.0

When statistical analyses are performed, the relative amounts of within- and between-crew variance in the criterion variable are first partitioned. Next, all individual-level analyses are based on the number of individuals, and are compared to the amount of within-crew variance. Finally, all crew-level analyses are based on the number of crews, and are compared to the amount of between-crew variance.

Because the non-independence among individual- and crew-level predictors can be calculated, HLM also computes cross-level relationships. The HLM equations that represent these relationships would appear something like this:

Level 1: $\text{Proficiency}_{ij} = b_{0j} + b_{1j}\text{ Situational Awareness}_{ij} + r_{ij}$

Level 2: $b_{0j} = \tilde{a}_{00} + \tilde{a}_{01}\text{Crew Cohesion}_j + U_{0j}$
$b_{1j} = \tilde{a}_{10} + \tilde{a}_{11}\text{Crew Cohesion}_j + U_{0j}$

At the individual level (level-1), individual maneuver proficiency is modeled as a function of two factors: individual levels of situational awareness (b_{1j}) and within-crew error (r_{ij}). As before, the intercept (b_{0j}) is a statistical necessity, representing the "average level" of maneuver proficiency for a hypothetical pilot with zero situational awareness. Although this is a typical OLS regression equation that has been performed at the individual level, separate OLS regression equations are computed for each crew.

It must be remembered that crews vary along a number of dimensions. For example, due to the quasi-random pairings of captains and first officers, crews will vary in their average level of maneuver proficiency. Similarly, crews will also vary in the degree to which the individual-level predictor(s) predict the individual-level criterion. As a result, there is likely to be significant between-crew variance in their level-1 intercepts and slopes.

At level-2, the first analysis focuses on predicting the intercepts from the level-1 analysis. This "intercept as outcome" model (b_{0j}) represents the extent to which crew cohesion predicts individual maneuver proficiency after controlling for individual levels of situational awareness. It varies as a function of two factors: crew cohesion ($\tilde{a}_{01}$) and between-crew error (U_{0j}). The intercept ($\tilde{a}_{00}$) represents the "average" level of maneuver proficiency for a hypothetical crew with zero cohesion.

The second level-2 analysis focuses on predicting the slopes from the level-1 analysis. This "slope as outcome" model represents the moderating effect of crew cohesion on the situational awareness-maneuver proficiency relationship. It varies as a function of two factors: crew cohesion ($\tilde{a}_{11}$) and between-crew error (U_{0j}). The intercept ($\tilde{a}_{10}$) represents the "average" level of maneuver proficiency for a hypothetical crew with zero cohesion.

Summary

Multi-level variance decomposition techniques such as HLM offer a number of advantages over traditional analysis techniques such as ANOVA and regression. First, because HLM separates out the criterion variance into within- and between-crew components, error terms are not systematically biased. This leads to more accurate effect size estimates and standard errors. Second, because HLM uses all available information, meaningful variance is not wasted. Finally, HLM allows for testing cross-level effects.

Despite its advantages, HLM is based on a number of statistical assumptions, some of which may or may not be tenable. In the section that follows, these issues will be explored in detail.

HLM's ASSUMPTIONS (AND THE CONSEQUENCES OF VIOLATING THEM)

Between- and Within-Crew Variance

Before any multi-level analysis can be conducted, a number of theoretical and statistical assumptions must be tested. First, key variables in the data set must contain sufficient amounts of within- and between-crew variance. Quite simply, if there is not a sufficient amount of within-crew variance, then the individual level (level-1) predictor(s) will not exhibit significant relationships with the individual level criterion. Likewise, there must be a significant amount of between-crew variance. Otherwise, there will be no significant relationships among the crew level (level-2) predictor(s) and the individual level (level-1) slopes and intercepts. Estimates regarding the relative amount of within- and between-crew variance are typically assessed via intra-class correlations (Bryk & Raudenbush, 1992).

Unfortunately, there is little consensus regarding what constitutes a "sufficient" amount of within- and between-crew variance. Traditionally, such estimates have been described in terms of statistical significance levels, rather than based upon an absolute criterion. This is problematic for two reasons. First, the amount of variance that is deemed statistically sufficient may vary as a function of sample size. More specifically, with larger sample sizes (i.e., high levels of statistical power), even trivial amounts of between crew variance may be statistically significant.

Second, and perhaps more importantly, the amount of variance to be explained must have some theoretical meaning. For example, suppose that the criterion variable contains a disproportionate mix of variance (e.g., 10% individual level, 90% crew level). If, after controlling for all individual-level predictors, the crew-level independent variables predict significant amounts of the remaining criterion variance, the result may be statistically significant but practically meaningless. As a result, it is incumbent upon the researcher to specify the level of analysis of each variable *a priori* and confirm this by statistical means before proceeding with the analyses (Bryk & Raudenbush, 1992).

Issues of Aggregation

Unless derived by group consensus, all individual-level estimates of crew-level phenomena must demonstrate significant within-group agreement, such as via estimates of r_{wg} (James, Demaree, & Wolf, 1984, 1993). If crew members don't agree (i.e., if their individual responses are not readily interchangeable with one another) then it is questionable as to whether a true crew-level effect is being observed. Unfortunately, there is disagreement regarding the "acceptable level" of within group agreement. While Nunnally (1978) originally hinted that levels of agreement as low as .50 may be acceptable for research purposes, higher levels of agreement (approximately .90) are typically required for applied purposes.

However, recent years have witnessed the use of variables with somewhat lower levels of within-group agreement. This is important, because substantial within-group disagreement (e.g., values as low as .50) represents a theoretical problem for applied researchers. Specifically, low levels of agreement may indicate within-group polarization, which is the exact opposite of agreement. In most cases, the crew-level variable should be dropped from the model, as it cannot be adequately tested. Alternatively, the level of agreement can be included as a separate crew-level predictor.

Methods of Measurement

There is considerable debate regarding how best to measure crew-level phenomena such as cohesion. Several researchers have argued that scale items should be posed and answered at the individual level, for example by requiring each individual to independently estimate his/her belief about him/herself. Others have argued that questions should be posed at the crew level and answered at the individual level, for example by asking individuals to independently estimate their crewmembers' perceptions. Still others have attempted to obviate the entire issue by using consensus measures (Gibson, Randel, & Earley, 1996).

Unfortunately, each technique has its drawbacks. Technically, inquiring about individual perceptions does not address a crew-level phenomenon, even if statistically significant levels of within-crew agreement are observed. At the same time, asking individuals to estimate collectively held beliefs may require information that they do not possess. Finally, consensus measures can lead to powerful members in the crew exerting their beliefs on less-powerful members, thereby effectively creating in an individual judgment.

Prior Interaction

If the underlying theory specifies that crew interaction is the sole cause of the crew-level variables, the observed levels of within-group agreement need to be based on the crewmembers' prior interaction. Otherwise, common background experiences among the pool of potential crew members (e.g., training, organizational culture), may result in statistically significant levels of within-crew agreement, even with quasi-random pairings. Because researchers often fail to assess the "sharedness" of their predictor constructs at the outset of the crews' formation, it is virtually impossible to determine whether "shared" effects (measured at a later point in time) are the result of the crewmembers' interaction, or are due to other factors. To remove these common background effects, residualized agreement values may be used. Specifically, within-crew agreement values can be calculated immediately upon the crews' formation. These values may then be statistically controlled for when the construct is measured at some point later during the crews' lifecycle.

Sample Size and Crew Size

Recent empirical work suggests that the usefulness of hierarchical linear modeling techniques may be limited by the overall sample size. Specifically, for the to EM algorithm obtain statistical convergence, it is necessary to have a large number of crews (Pollack, 1998). For large carriers, this may not represent a problem. For smaller carriers, however, this may require combining crews across multiple fleets.

Crew size may also be an issue. Previous research employing group sizes as low as three have led to difficulties for the HLM program in estimating within-crew variance, especially when the respondents answer the questions very similarly (Pollack, 1998). This has startling implications for commercial aviation research, because all new aircraft are certified for two-person crews. Nevertheless, HLM may still be applicable to other types of flight-related crews, such as flight attendants, maintenance crews, and dispatch teams, although these groups are less likely to be represented in the commercial aviation research literature. Similarly, HLM may be applied to flight crews, if the definition of a flight crew is expanded to include both the pilot crew and the cabin crew.

Range Restriction

Like every other statistical technique, HLM requires that the predictor and criterion variables be approximately normally distributed (Bryk & Raudenbush, 1992). According to modern statistical theory, range restriction decreases the variance of observed variables. Decreased variances, in turn lead to decreased covariances, thereby operating against detecting empirical relationships. At the current time, the exact biasing effects of range restriction on cross-level relationships is unknown. However, given that normally distributed variables are somewhat uncommon in organizational settings (i.e., to the extent that the organization's recruitment, selection, training, performance evaluation, and termination programs are working properly) it is improbable that individual and crew performance ratings will be normally distributed (Murphy & Cleveland, 1995). Further, given that this is not a statistical artifact, but is rather a true organizational phenomenon, it does not make sense to "correct" such correlations.

However, range restriction may be reduced by programs that are designed to increase the sensitivity of the evaluation process (Holt, 2001). For example, rater training programs can be developed to assist pilot instructors in making fine discriminations among performance levels. Assessment procedures can also be revised to encourage greater variability in performance ratings.

The Criterion Problem

One of the most damaging criticisms is that HLM requires the dependent variable to be operationalized at the lowest level of analysis. More specifically, if individual- and crew-level data are to be analyzed, then the criterion must be measured at the individual level of analysis. This represents a practical problem, because crew performance is typically more important to aviation researchers than individual performance (i.e., because both pilots share a common fate).

One option is to collect multiple observations of individual and crew-level data over time, and to use residualized measures of crew performance as the criterion. For example, if individual situation awareness and crew performance are measured at time 1 and time 2, time 1 measures of both constructs can be partialled out of crew performance at time 2. To date, however, only one published study has attempted such a feat (Griffin, 1997). Further, it is somewhat unclear exactly what such residualized measures are in fact measuring. Quite simply, because residualized scores are used, we know what they are not measuring (e.g., previous level-1 and level-2 effects), but it less clear what they are measuring. Finally, it may be difficult to collect large numbers of longitudinal, cross-level data because of time constraints, attrition, and other limited organizational resources.

CONCLUSIONS AND RECOMMENDATIONS

In this paper, we have attempted to articulate the strengths and weaknesses of multi-level variance decomposition techniques such as HLM. Given the inherently hierarchical nature of organizational phenomena, we believe that such techniques may be meaningfully applied to a number of research domains in the field of commercial aviation.

Many of the previously-identified criticisms are not limited to HLM *per se*. Rather, many are common to all data analytic techniques, and may simply represent incongruities between the practice of organizational research and the statistical requirements. Despite previous arguments to the contrary (Bryk & Raudenbush, 1992), it would appear that HLM places just as many assumptions and requirements on the researcher as do single-level data analysis techniques such as multiple regression or ANOVA.

While these arguments do not diminish the value of multi-level variance decomposition models, it does leave room for improvement. Perhaps the problem is not with the statistical modeling technique. Perhaps the problem lies in the way we as a field conduct research. Perhaps the answer lies somewhere in between.

As noted earlier, HLM is an extremely flexible data analysis technique. Because of its flexibility, some might argue that HLM's greatest value does not even involve the analysis of individual- and crew-level data. Rather, HLM's true value may involve the analysis of repeated measurements over time. For example, given the voluminous amount of highly-reliable FOQA-type data that can be obtained from flight simulators (e.g., number of exceedences), HLM may help aviation researchers estimate individual performance trajectories for complex, technical maneuvers. By modeling performance decrements over time, aviation researchers may be able to estimate optimal re-training intervals on a maneuver-by-maneuver basis.

REFERENCES

Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*(4), 355-373.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Griffin, M. A. (1997). Interaction between individuals and situations: Using HLM procedures to estimate reciprocal relationships. *Journal of Management, 23*(6), 759-773.

Hofmann, D. A., Griffin, M., A., & Gavin, M. B. (2000). The application of hierarchical linear modeling to organizational research. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 467-511). San Francisco: Jossey-Bass.

Holt, R. W. (2001). *Scientific information systems*. Aldershot, UK: Ashgate.

James, L. R., Demaree, R. G., & Wolf, G. (1993). rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*(2), 306-309.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Nunnally, J. C. (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology, 78*, 569-582.