Industrial and Organizational Psychology, **1** (2008), 167–173.

Integrating Traditional Perspectives on Error in Ratings: Capitalizing on Advances in Mixed-Effects Modeling

DAN J. PUTKA, MICHAEL INGERICK, AND RODNEY A. McCLOY
Human Resources Research Organization

The purpose of this commentary is to raise awareness among Industrial and Organizational (I–O) researchers and practitioners regarding how linear mixed models (LMMs) can provide a framework for integrating traditional perspectives on error in performance ratings. The types of rating models discussed by Murphy (2008) were largely formulated long before modern methods for fitting mixed models were established and incorporated into common statistical software (Littell, Milliken, Stroup, & Wolfinger, 1996; SPSS, Inc., 2005). Although the application of certain classes of LMMs has found its way into the I–O literature—most notably hierarchal linear models (e.g., Bliese, 2002)—application of the more general LMM has yet to cross over into I–O research (Searle, Casella, & McCulloch, 1992). This state of affairs is unfortunate. The remainder of our response details on how LMMs can be used to integrate historically distinct perspectives on error in ratings and the value of doing so.

Traditional Perspectives on Error in Ratings

Historically, organizational researchers have maintained two perspectives on error in ratings: (a) a reliability-based perspective, based largely on Classical Test Theory (CTT), that views error in ratings as random and unexplainable (Lord & Novick, 1968) and (b) a validity-based perspective that views error in ratings as resulting from systematic, construct-irrelevant sources of variance such as specific characteristics of raters, ratees, and contexts in which ratings are made (Landy & Farr, 1980). These perspectives are realized in the three types of ratings models discussed by Murphy: The reliability-based perspective is reflected in Murphy's one-factor model and the validity-based perspective is reflected in Murphy's multifactor and mediated models.

Researchers have generally avoided integrating reliability- and validity-based perspectives on error in ratings, in part because of the methodological and pedagogical traditions that have evolved around them (Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000). Although examples of integrated ratings models exist, they still tend to treat systematic sources of variation (both valid and invalid) and "measurement error" as distinct, unrelated entities (Lance, Baxter, & Mahan, 2006). Such a distinction is clearly reflected in Murphy's

Correspondence concerning this article should be addressed to Dan J. Putka. E-mail: dputka@humrro.org

Address: Human Resources Research Organization, 66 Canal Center Plaza, Suite 400, Alexandria, VA 22314

Dan J. Putka, Michael Ingerick, and Rodney A. McCloy, Human Resources Research Organization.

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press

figure 1, where sources of validity (and invalidity) are depicted on the left-hand side of the models and an omnibus measurement error term is depicted on the right-hand side of the models.

The Need for an Integrated Framework

Past researchers have acknowledged that the distinction between reliability- and validity-based perspectives on error becomes blurred when one considers measurement error through the lens of generalizability theory (G-theory) (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Murphy & DeShon, 2000). Under G-theory, measurement error can be multifaceted in nature, with the structure of that error reflecting the conditions of measurement (e.g., raters, items, occasions) across which one wishes to generalize a given measurement procedure (Cronbach et al., 1972). Unlike CTT, G-theory neither assumes nor implies that variance across measurement conditions is unexplainable—and therefore unrelated to any external, substantive variables. Indeed, Cronbach et al. (1972) clearly recognized this:

> Suppose, for example, it is found that in peer ratings there is a substantial subject-rater interaction component … the finding should impel the investigator to ask what rater characteristics contribute to [*explain variance attributable to*] the interaction (p. 382, bracketed text added for clarity).

The above statements are not meant to be an indictment of CTT, rather they are just meant to illustrate that relative to G-theory, CTT implies a relatively narrow definition of measurement error. Remember, the foundations of CTT emerged during a period when items on cognitive ability tests and occasions of measurement were the primary measurement facets of interest, *not* raters (Spearman, 1910). Raters, unlike test items, *cognize and reason*; as such measurement designs that involve raters introduce several potential sources of variance (some that are likely predictable, and others that are essentially unpredictable), that do not really have analogues when test items or occasions are the only facets of measurement (Murphy & DeShon, 2000). Thus, although the perspective that CTT offers on error may be sufficient when test items or occasions are the only facets of measurement, it is arguably not as well suited for designs involving raters as a facet of measurement. As the quote from Cronbach et al. implies, G-theory is a bit more liberal in its definition of error relative to CTT and does not close the door on the possibility of explaining "error" in ratings—it leaves the question open for researchers to address.

In light of the observations above, it is somewhat ironic that G-theory offers little in the way of modeling the *substantive* basis of (a) variance across measurement conditions (e.g., raters, items, occasions) and (b) variance attributable to interactions between such conditions and one's objects of measurement (e.g., variance attributable to Ratee × Rater interaction effects). This is not an indictment of G-theory per se. Rather, it reflects the fact that the variance partitioning that occurs in the context of G-theory is designed simply to estimate the magnitude of variance components reflecting facets of one's measurement design (which is consistent with the purpose of G-theory), *not to explain the substantive nature of the variability* attributable to those components. Thus, the traditional G-theory model offers little insight into the contribution of specific, construct-irrelevant sources of variance in ratings such as those noted in Murphy's second and third types of models.

How Can the LMM Framework Help?

Conceptually, the benefit of using LMMs for modeling ratings is that they offer a method not only for partitioning variance in ratings into components attributable to facets of one's measurement design (as G-theory does) but also for simultaneously modeling the impact of substantive variables (e.g., characteristics of rater–ratee pairs or of individual raters) on each of those variance components. That is, rather than treating

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press

substantive contaminants and error arising from one's sampling of measurement conditions as distinct entities, researchers can use LMMs to estimate the degree to which the former can explain components of the latter. By adopting an LMM approach to modeling variance in ratings, researchers can begin to classify substantive sources of variation in ratings (e.g., the variables on the left side of Murphy's figure 1) in terms of the pathways (components) through which they influence ratings. As such, LMMs offer a framework for integrating reliability- and validity-based perspectives on error in ratings.

How Do LMMs Differ From Models Commonly Used in I–O?

To help clarify how LMMs "work," it is useful to illustrate their relation to the general linear model (GLM) with which most researchers are familiar. Recall that the GLM encompasses both analysis of variance (ANOVA) and simple linear regression models that have been used by researchers for nearly a century. The GLM can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is an $N \times 1$ column vector of observed values on the outcome of interest (e.g., job performance ratings for N ratees), $\boldsymbol{X}$ is an $N \times k$ matrix of observed values for $k - 1$ predictor variables (the first column of values in $\boldsymbol{X}$ consists of 1s, which allows an intercept to be included in the model; subsequent columns represent predictors, such as ratee conscientiousness), $\boldsymbol{b}$ is a $k \times 1$ column vector of fixed-effect parameters (i.e., the model's intercept and slope coefficients for each predictor variable), and $\boldsymbol{e}$ is the $N \times 1$ column vector of residual errors of prediction. In nonmatrix form, this translates into:

$$y_i = b_0 + \sum_{m=1}^{k-1} b_m x_{mi} + e_i, \tag{2}$$

where the "i" subscript indexes the ith ratee and the "m" subscript indexes the mth predictor variable.

The LMM, in contrast, is a more general version of the GLM and can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{Z}$ is the random-effects design matrix (explained below) and $\boldsymbol{u}$ is a vector of random-effect parameters (also explained below). To illustrate the meaning and purpose of $\boldsymbol{Z}$ and $\boldsymbol{u}$ in the LMM, consider the following mock example.

Assume we have job performance ratings for three ratees, each of whom was rated by two raters. Further, assume that each ratee was rated by the same two raters (i.e., a fully crossed measurement design) and that we have measures of (a) each rater's agreeableness (x_1) and (b) each rater–ratee pair's similarity in terms of work values (x_2) that are entered into the model as predictors. If we were to model the performance ratings using the LMM, Equation 3 would take on the following form:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & x_{1\cdot1} & x_{211} \\ 1 & x_{1\cdot2} & x_{212} \\ 1 & x_{1\cdot1} & x_{221} \\ 1 & x_{1\cdot2} & x_{222} \\ 1 & x_{1\cdot1} & x_{231} \\ 1 & x_{1\cdot2} & x_{232} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1\cdot} \\ u_{2\cdot} \\ u_{3\cdot} \\ u_{\cdot1} \\ u_{\cdot2} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}. \tag{4}$$

In the above formulation, the y_{ij} are job performance ratings for the ith ratee as made by the jth rater. The $x_{1\cdot j}$ are values on the rater agreeableness measure for the jth rater (note that these values do not vary across ratees)

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press

and x_{2ij} are values on the rater–ratee work value similarity measure (note that these values vary across ratees and raters). The purpose of random-effects design matrix (***Z***) is to specify the structure of the measurement design underlying the ratings data. In this example, the first three columns of ***Z*** indicate which of the six ratings are for ratees, 1, 2, and 3, respectively, and the last two columns of ***Z*** indicate which of the ratings were made by raters 1 and 2, respectively.[1] The ***u*** column vector by which ***Z*** is postmultiplied holds the random-effect parameters associated with each ratee and rater in one's design. Essentially, these random effects reflect unmodeled sources of consistency in ratings associated with a given ratee (ratee main effects) and a given rater (rater main effects). Labeling these effects as "random" signifies the notion that the ratees and raters participating in our study represent a sample from some broader population of ratees and raters, *not* that the variance in ratings across ratees or raters is unexplainable. For example, addition of rater-level or rater and ratee-level variables to the model, such as rater agreeableness, or rater–ratee work value similarity, might explain some of the variation in those effects. In nonmatrix form, Equation 4 translates into:

$$y_{ij} = b_0 + b_1 x_{1\cdot j} + b_2 x_{2ij} + u_{i\cdot} + u_{\cdot j} + e_{ij}. \quad (5)$$

Another important difference to note between the GLM and the LMM in the context of modeling ratings is that instead of modeling ratings that have been averaged across raters for each ratee, use of the LMM involves modeling *disaggregated* ratings (i.e., ratings for each rater–ratee pair are treated as separate observations). If we were to fit a simple GLM to data with such a structure, we would violate the assumption of independence of residual errors of prediction because some of the observations would share a ratee in common and other observations would share a rater in common (Bliese & Hanges, 2004). The addition of the random-effect portion of the LMM (***Zu***) not only accounts for these sources of nonindependence but also allows for estimation of variance components associated with different facets of one's measurement design (e.g., variance attributable to ratee main effects, rater main effects, etc.). These variance components serve as the building blocks for generalizability coefficients and intraclass correlations (Cronbach et al., 1972; McGraw & Wong, 1996).

Although we highlight differences between the LMM and the GLM using matrix notation, the example above is for pedagogical purposes only. Fitting these models does not require use or knowledge of matrix algebra. These models can be estimated easily using mixed-model procedures widely available in statistical software. For example, both SAS and SPSS have mixed-model procedures that allow one to fit LMMs and offer extensive formal documentation on how to fit such models (e.g., Littell et al., 1996; SPSS, Inc., 2005).

An Example of Integrating Perspectives on Error Via the LMM

To illustrate how the LMM integrates various perspectives on error in ratings, it is useful to draw some parallels to traditional applications of G-theory and hierarchal linear modeling (HLM). As we note below, the LMM represents a more general version of (a) HLMs commonly used by organizational researchers to model multilevel data (Bliese, 2002) and (b) random-effect ANOVA models that underlie G-theory (Cronbach et al., 1972).

First, consider what a typical decomposition of ratings based on G-theory provides us. In G-theory, we fit a reduced version of the LMM that typically omits all fixed-effect parameters except for the intercept term,

[1] Although our example illustrates how ***Z*** may be structured when modeling ratings arising from a fully crossed measurement design, LMMs afford researchers complete flexibility for structuring ***Z*** depending on the measurement design they encounter in practice. For example, ***Z*** can be structured to fit a model for ratings arising from a design in which raters are nested within ratees (i.e., each ratee is rated by a unique, nonoverlapping set of raters) or a design that is more ill structured in nature (e.g., each ratee is rated by a partially overlapping set of raters).

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press

which essentially leaves us with a random-effect ANOVA model. In terms of Equation 4, the $\boldsymbol{X}$ matrix becomes a column vector consisting only of 1s, the $\boldsymbol{b}$ vector becomes a single value reflecting the model intercept (essentially, the grand mean rating across rater–ratee pairs), and the structure of the random-effect design matrix ($\boldsymbol{Z}$) mimics the measurement design underlying the ratings (as shown in the example above). In terms of Equation 5, this simply amounts to removing the $b_1x_{1\cdot j} + b_2x_{2ij}$ portion of the model. If we were to fit such a random-effect model to the performance ratings in the example above, it would partition variance in ratings into three components: (a) variance attributable to ratee main effects ($\sigma^2{}_T$), (b) variance attributable to rater main effects ($\sigma^2{}_R$), and (c) variance attributable to the combination of the Ratee × Rater interaction and residual effects ($\sigma^2{}_{TR,\boldsymbol{e}}$). Although such a decomposition of variance is useful from a G-theory perspective, it does not address the question of what substantive variables (if any) explain variance attributable to each of these components. However, such a question can be answered easily within the context of the LMM.

To estimate the impact that substantive variables such as rater agreeableness and rater–ratee similarity have on *each* component of variance in ratings (i.e., $\sigma^2{}_T$, $\sigma^2{}_R$, $\sigma^2{}_{TR,\boldsymbol{e}}$), one would follow a two-step process. First, as described above, one would fit a model that omitted all substantive predictor variables and let the random-effect portion of the model reflect the structure of one's measurement design. The resulting model would provide *unconditional* estimates of the variance components underlying the ratings (i.e., precisely what is provided via G-theory). Next, one would add the substantive predictor variables of interest (e.g., rater agreeableness, rater–ratee work value similarity) to the aforementioned model as fixed predictor variables (e.g., Equations 4 and 5). This second model would produce estimates of $\sigma^2{}_T$, $\sigma^2{}_R$, $\sigma^2{}_{TR,\boldsymbol{e}}$ that are *conditional* on the substantive predictor variables—that is, the components would reflect decomposition of the variance in ratings after removing variance accounted for by the predictors (Searle et al., 1992). The magnitude of these conditional variance component estimates could then be compared to the unconditonal variance component estimates from the first model to calculate the proportion of variance attributable to each component that was explained when the substantive predictor variables of interest were added to the model (Snijders & Bosker, 1944). If a substantive predictor variable explains variance attributable to a component, that component should become smaller when the predictor variable is added to the model. Table 1 provides an example of SAS PROC MIXED code for fitting the series of models described in the example above. As alluded to earlier, the data set on which this code is run would need to be structured such that each row in the data set corresponds to a unique rater–ratee pair.

Readers familiar with the HLM literature may observe that the two-step process described above is analogous to standard practice when modeling multilevel data (Bliese, 2002). A typical first step in applications of HLMs is to fit a null model that provides unconditional variance component estimates for each level of the model (e.g., individual and group levels). These unconditional components serve as a baseline for

Table 1. *Example of SAS PROC MIXED Code for Fitting Linear Mixed Models*

Step 1: fit the random-effects model

```
proc mixed method = REML;
   class ratee_id rater_id;
   model rating = ;
   random ratee_id rater_id;
run;
```

Step 2: fit the mixed model

```
proc mixed method = REML;
   class ratee_id rater_id;
   model rating = rater_agreeableness
                  rater_ratee_similarity;
   random ratee_id rater_id;
run;
```

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press

judging the contribution of fixed individual- and group-level predictors that are added in subsequent steps of the modeling process. The primary difference between the typical HLM modeling strategy and the application of LMMs proposed here is that instead of attempting to explain variance in an outcome variable at different *levels of analysis*, we are attempting to explain variance associated with different *facets of our measurement design*. Furthermore, unlike current HLMs in vogue in the I–O literature, LMMs do not practically limit us to modeling nested data structures. The LMM is flexible enough to easily model data structures involving nearly any possible combination of nested and crossed design factors imaginable. Taken together, the LMM provides researchers with a tool that substantially extends the modeling framework underlying G-theory and current applications of multilevel modeling.

Leveraging the LMM to Refine the Classification of Errors in Ratings

The preceding sections illustrate how LMMs can provide researchers with a powerful tool for integrating reliability- and validity-based perspectives on error in ratings. One key implication of using LMMs in this manner is that they can offer researchers a more refined way to classify error in ratings. For example, the LMM can allow researchers to differentiate between variables that act as (a) *independent* contaminants of ratings via rater main effect variance (e.g., they explain differences in rater leniency/severity but do not explain any ratee main effect [true score] variance) or via rater–ratee interaction effect variance (e.g., they explain the idiosyncrasies in raters' ratings of individual ratees but do not explain any ratee main effect variance) and (b) *nonindependent* contaminants (e.g., they explain variance not only in one or both of the rater-related variance components above but also ratee main effect [true score] variance). Being able to identify and minimize independent contaminants would improve the reliability of ratings, whereas identifying and minimizing nonindependent contaminants could potentially reduce the reliability of ratings (as traditionally defined). Regardless, these examples illustrate how LMMs can be used to refine the classification of errors in ratings and, in turn, clarify the implications such errors have for the reliability of the resulting measure.

Summary

The LMM provides researchers with a tool to quantify and classify the contribution of various sources of error to components of variance that underlie ratings. Although far from being a comprehensive treatment of applying LMMs for this purpose, we hope that this commentary encourages researchers and practitioners to explore the potential of LMMs for improving our understanding of performance ratings. We realize that simply improving our quantification and classification of errors in ratings will not ameliorate them. However, the hope is that a better understanding of the variability in ratings will enable us to design more effective rating systems.

References

Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 401–445). San Francisco: Jossey-Bass.

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as thought they were independent. *Organizational Research Methods*, *7*, 400–417.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Lance, C. E., Baxter, D., & Mahan, R. P. (2006). Evaluation of alternative perspectives on source effects in multisource performance measures. In W. Bennett, C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 49–76). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72–107.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 148–160.

Murphy, K. R., & DeShon, R. (2000). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913–924.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901–912.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological methods & research*, *22*, 342–363.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

SPSS, Inc. (2005). *Linear mixed-effect modeling in SPSS: An introduction to the mixed procedure* (Technical Report LMEMWP-0305). Chicago, IL: Author.

https://doi.org/10.1111/j.1754-9434.2008.00032.x Published online by Cambridge University Press