Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 3

11-1-2003

Fortune Cookies, Measurement Error, And Experimental Design

Greogry R. Hancock
University of Maryland, ghancock@umd.edu

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 293-305

1538 – 9472/03/$30.00

Fortune Cookies, Measurement Error, And Experimental Design

Gregory R. Hancock
University of Maryland

This article pertains to the theoretical and practical detriments of measurement error in traditional univariate and multivariate experimental design, and points toward modern methods that facilitate greater accuracy in effect size estimates and power in hypothesis testing.

Keywords: measurement error, latent variables, multivariate analysis, experimental design

Introduction

Whichever leg of my post-secondary academic journey, and with whichever campus I have had the privilege of affiliating, the vast majority of my midday meals have ended with a fortune cookie. Since my college days, in fact, I estimate that I have had lunch at some inexpensive Asian restaurant near campus well over a thousand times. My graduate school office mates and the many students and faculty whom I served as teaching assistant might even remember all the little strips of paper taped to the top of my desk, filling the entire surface with fortunes by the time I finished my doctorate.

Gregory R. Hancock is Professor in the Department of Measurement, Statistics and Evaluation at the University of Maryland. His research appears in such journals as *Psychometrika*, *Structural Equation Modeling*, and *Journal of Educational and Behavioral Statistics*. He serves on several journal editorial boards, and regularly conducts workshops around the U.S. Email: ghancock@umd.edu.

Today, a little more reserved in my decorative zeal, though no less so in my meal predilection, I have but a single fortune tacked outside of my office door. Amidst aging cartoons and family pictures is an enlarged photocopy of the one little rectangle of wisdom I have saved over these last decades. It reads:

Love truth
but pardon error.
Lucky Numbers 7, 8, 13, 31, 32, 44

Although my quantitative training precludes me from seeking fortune based on the third line, not so with the first two. Their aphorism seems replete with insight and potential on many levels, personal and professional, with the latter level serving as the inspiration for this article.

Less obtusely, in so many applied statistical analyses there seems to be a schism between the variables we have and the variables we wish we had. This is apparent in statements of theory preceding and justifying those analyses and in the interpretations and purported implications that follow. Educational policy researchers, for example, might analyze measures of teacher’s job satisfaction and absenteeism and then make proclamations regarding the apparent degree of teacher burnout. Those studying child development might start by eliciting new mothers' responses to rating scale items regarding interactions with their infants, and conclude by making inferences about those mothers' emerging maternal warmth. Health care researchers might want an understanding AIDS patients' sense of hopelessness while in group therapy, and choose measures of patients' treatment compliance to help facilitate that understanding. Such is the nature of so much applied research, particularly within the social sciences — constructs of interest such as burnout, maternal warmth, or hopelessness are generally latent, so our analyses seem resigned to rely upon the fallible measured variables as surrogates.

And therein lies the schism, in the operationalization of true constructs as error-laden measured variables. At best the imperfect connection might lead us to a distorted image of the critical relations in a population; at worst we might not even have sufficient power to draw inference at all. Within the context of experimental design specifically, the primary focus of this treatise, the implication is that treatment effectiveness might be severely underestimated, or perhaps even undetected. Of course this is not unknown. In fact, nothing written here will be new knowledge. But it is important and often-overlooked knowledge, bearing clarification and amplification. It will thus be my purpose to drive home the often underestimated (if not entirely disregarded) importance of constructs and measurement error in our univariate and multivariate experimental analyses, and to point the applied researcher toward more modern strategies for dealing with measurement error in experimental design.

Love Truth

The purpose of applied statistics seems to be to gain insight into some truth bearing practical consequence. Drawing upon a few familiar test statistics, we attempt to use observed relations among measured variables in samples to make educated guesses about unobserved relations in the populations of which each sample serves as assumed microcosm. But what, precisely, is the population relation we hope to understand in order to have practical consequence? What is the truth into which we seek insight?

As we learn and practice the many methods huddled under the general linear model umbrella, we typically hold as our goal a correct inference about, and often estimation of, some population relation among observed variables – a true correlation between X and Y (ρ_{XY}), a true predictive relation of X_3 to Y holding X_1 and X_2 constant (β_3), a true standardized effect size for the mean difference between Populations 1 and 2 on Y (d_Y), and so on. But what does any measured X or Y variable really represent, and what information do any relations among such variables convey?

In the physical sciences, variables such as temperature, pressure, mass, and volume, when considered in sufficient quantities, are in their measurement as they are in name. That is, there tends to be a strong correspondence between the measurement and the entity it represents. In the social sciences, some such variables exist as well – biological sex, treatment group assignment, and political party affiliation, for example. Except for data recording or entry errors, we expect each variable to represent precisely that which its name implies. Other social science variables would also *seem* to have such identity, being determinable largely without interference – number of therapy sessions attended, number of children's books in the home, and the like. However, a fundamental question in many disciplines, particularly those in the social sciences, is the following: What is the underlying construct that each variable has been selected to represent?

The univariate scenario

Consider a researcher who is truly interested in a construct contrived here as In-Home Reading Resources. In that case, number of children's books in the home is indeed a fairly proximal operationalization of the construct of interest. As such, estimates regarding population mean differences in number of children's books in the home, or regarding the population relations this measured variable has with other such proximal operationalizations, provides direct insight into some truth for the construct of In-Home Reading Resources. On the other hand, if a researcher is interested in a construct designated as Parental Commitment to Literacy, and has attempted to capture the spirit of this construct using the number of children's books in the home, then we expect the measured variable to be a more distal operationalization of the desired construct. As such, inference about population differences in Parental Commitment to Literacy, or of its relation to other variables (proximally or distally operationalized), is compromised by typical general linear model analyses. Thus, the truths we seek – constructs and their population relations – are often not directly accessible.

The issue at hand, of course, is one of measurement error in our variables. As an indicator of In-Home Reading Resources, number of children's books in the home has virtually no measurement error; when reflecting Parental Commitment to Literacy, however, it has considerable error. Imagine a researcher employing a control and treatment group to draw inference about the impact of a treatment designed to enhance Parental Commitment to Literacy. Figure 1 displays hypothetical population distributions for the measured variable of number of children's books in the home (*Y*), as well as for the latent construct of Parental Commitment to Literacy (η). Notice that while the means of the two populations are expected to be the same for *Y* and η, the relative magnitude of the treatment effect on the Parental Commitment to Literacy construct would be underestimated. The standardized effect size for *Y*, which is the familiar

$$d_Y=(\mu_{1Y}-\mu_{2Y})/\sigma_Y \qquad (1)$$

(Cohen, 1988), is depicted as approximately .65; meanwhile, the standardized effect size for η,

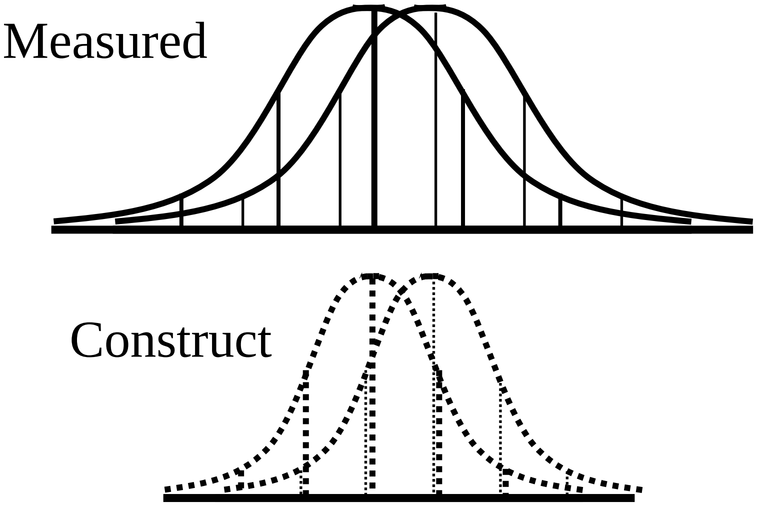
$$d_\eta=(\mu_{1\eta}-\mu_{2\eta})/\sigma_\eta, \qquad (2)$$

is near .95. For this disparity to occur, the construct's standard deviation would have to be 68.4% of the size of standard deviation of *Y*, meaning its variance is roughly 46.8% ($.684^2$) that of *Y*. That is, 46.8% of the variability in *Y* reflects η, while 53.2% is error with respect to the construct of interest. Put directly,

$$d_Y^2 = \rho_{YY} d_\eta^2, \qquad (3)$$

where ρ_{YY} is the reliability of *Y* (.468 in the above example). Thus, while the number of children's books in the home may accurately reflect In-Home Reading Resources, with regard to Parental Commitment to Literacy it could be a relative overestimate or underestimate for any given individual.

Figure 1. Univariate population difference on measured variable and underlying construct.



As mentioned previously, the implications of measurement error for inference are two-fold. First, as seen in Figure 1, we would underestimate the magnitude of the treatment effect on Parental Commitment to Literacy. That is, we would make an incorrect estimate of the truth we seek. Second, the presence of the error variance would decrease the power of a two-sample test to detect the presence of that treatment effect. An understanding of this loss of power may be communicated in terms of additional subjects needed in each group to compensate for the presence of measurement error. Assuming a desired level of power (e.g., .80) and a specific standardized effect size at the construct level (e.g., d_η=.30), the number of subjects per group for a two-sample *z*-test can easily be shown to be:

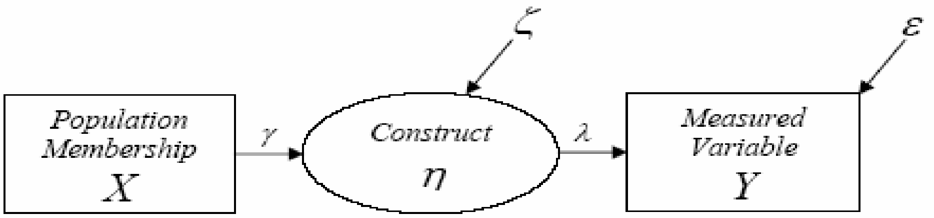
$$n_Y = (1/\rho_{YY})\, n_\eta. \qquad (4)$$

For example, conducting the test using a valid measure with reliability of ρ_{YY} =.50 would require twice as many subjects as a test that could, hypothetically, be conducted directly at the construct level. This result holds for *t*-tests as well for all but the smallest sample sizes, where appreciable changes in the critical value make the relation only approximate. Further, except for very small samples, Equations 3 and 4 hold for *k*-group between-subjects analysis of variance (ANOVA) as well using the more general *k*-group effect size measures (see Cohen, 1988).

The scenario for the univariate outcome may also be depicted symbolically using a path diagram. In Figure 2 we see the measured variable *Y* being defined by two components, the construct of interest η and measurement error ε. The connection between η and *Y*, labeled as λ in Figure 2, symbolically reflects the (square root of the) measured variable's reliability. The stronger the relation λ, the more proximal *Y*'s operationalization of η and thus the less error variance it contains; conversely, the weaker λ, the more distal *Y*'s operationalization of η and thus the more error variance it contains. On the left we see a grouping variable representing population membership and whose influence is being assessed; this could be a single variable for *k*=2 groups, or *k*-1 group code variables for the general *k*-group case.

As depicted, population membership *X* has a potential bearing γ on the construct η underlying the measured variable *Y*, while the remaining variance in η is accounted for by other independent but latent residual influences ζ. Thus, an observed population difference on the measured variable *Y* is actually the attenuated manifestation of a population difference on the true underlying construct of interest. The weaker the connection between the η and *Y* (i.e., the weaker the reliability), the less well the population difference on the construct of interest is propagated to, and thus reflected in, the observed variable.

Figure 2. Path model for univariate case



This simple univariate example underscores two needs regarding truth in experimental design and analysis. First, we must seek measured operationalizations as proximal to their constructs as possible. Certainly in the social sciences perfect operationalization is generally unrealistic, particularly given the vagaries of human behavior, perception, affect, and attitude. Notwithstanding, researchers should expend considerable effort to select or construct the most valid and reliable measures feasible. Second, to the extent that measurement error remains, we must employ analytic methods that maximize the accuracy of inference and estimation, thereby portraying population truths with the greatest clarity. These analytic methods must, to every extent possible, penetrate the measurement noise to achieve the same fidelity to truth as the theoretical questions that preceded and the practical proclamations we hope to follow. One attempt to do so lies within a multivariate scenario.

The multivariate scenario

Researchers often attempt to enhance their ability to make inference about population differences by gathering several pieces of evidence to be employed within a multivariate experimental design. In multivariate analysis of variance (MANOVA) with outcome measures Y_1 through Y_m, the hope is that the signal of population differences on some combination of variables will be detected above the noise of their measurement error. This portion of the article will address MANOVA in the presence of measurement error, and highlight its somewhat misguided attempt to get closer to truth.

Consider the multivariate scenario with k=2 populations, often analyzed using Hotelling's T^2. An example is depicted in Figure 3 using m=2 outcomes for simplicity, and with extremely large population differences for clarity. As before, assume that each Y_i measure is an operationalization of its own specific construct η_i, with individual standardized effect sizes of d_{Y_i} and d_{η_i} for the univariate measured and latent population mean differences, respectively. The assessment of the multivariate population difference between centroids $\boldsymbol{\mu}_{Y1}$ and $\boldsymbol{\mu}_{Y2}$ is tantamount to evaluating the univariate mean difference on the maximally differentiating discriminant function W=w_1Y_1 + w_2Y_2 = $\mathbf{w}'\mathbf{Y}$, with weights $\mathbf{w}$ commonly (but not necessarily) chosen so the within-group variance σ_W^2 =$\mathbf{w}'\boldsymbol{\Sigma}_Y\mathbf{w}$ equals 1. Observed and latent variable distributions on each Y_i axis, as well as on the W axis, are depicted in Figure 3.

Given that W is a linear combination of the observed variables, the measurement error of each Y_i is propagated to the linear composite W. The standardized effect size along the W axis, the effect of interest in MANOVA, is d_W=(μ_{1W}-μ_{2W})/σ_W; it appears as approximately 3. The square of this effect size, d_W^2, may be computed as the squared Mahalanobis' distance

$$D_W^2 = [\boldsymbol{\mu}_{1Y} - \boldsymbol{\mu}_{2Y}]'\boldsymbol{\Sigma}_{Y_{within}}^{-1}[\boldsymbol{\mu}_{1Y} - \boldsymbol{\mu}_{2Y}], \qquad (5)$$

where $\boldsymbol{\Sigma}_{Y_{within}}$ is the pooled (within-group) variance-covariance matrix reflecting the observed Y_i measures' m-dimensional dispersion and within lurks the influence of measurement error. Specifically, $\boldsymbol{\Sigma}_{Y_{within}} = \boldsymbol{\Sigma}_{\eta_{within}} + \boldsymbol{\Sigma}_{\varepsilon_{within}}$, where $\boldsymbol{\Sigma}_{\eta_{within}}$ is the pooled (within-groups) variance-covariance matrix of the specific constructs η_i and $\boldsymbol{\Sigma}_{\varepsilon_{within}}$ is a diagonal matrix of within-group error variances, assumed independent and each equal to $\sigma_{Y_i}^2(1-\rho_{Y_iY_i})$. Thus,

$$D_W^2 = [\boldsymbol{\mu}_{1Y} - \boldsymbol{\mu}_{2Y}]'[\boldsymbol{\Sigma}_{\eta_{within}} + \boldsymbol{\Sigma}_{\varepsilon_{within}}]^{-1}[\boldsymbol{\mu}_{1Y} - \boldsymbol{\mu}_{2Y}]. \ (6)$$

As seen in Figure 3, the population mean difference on the W axis mirrors the univariate case, where the standardized effect size on the measured composite W underestimates the standardized effect size on the corresponding underlying construct. In this case, the construct underlying W, denoted here as η_W, is a linear combination of the η_i constructs underlying the respective measured Y_i variables: η_W =$w_1\eta_1$ + $w_2\eta_2$ = $\mathbf{w}'\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is the vector of η_i constructs. Whereas the measured standardized

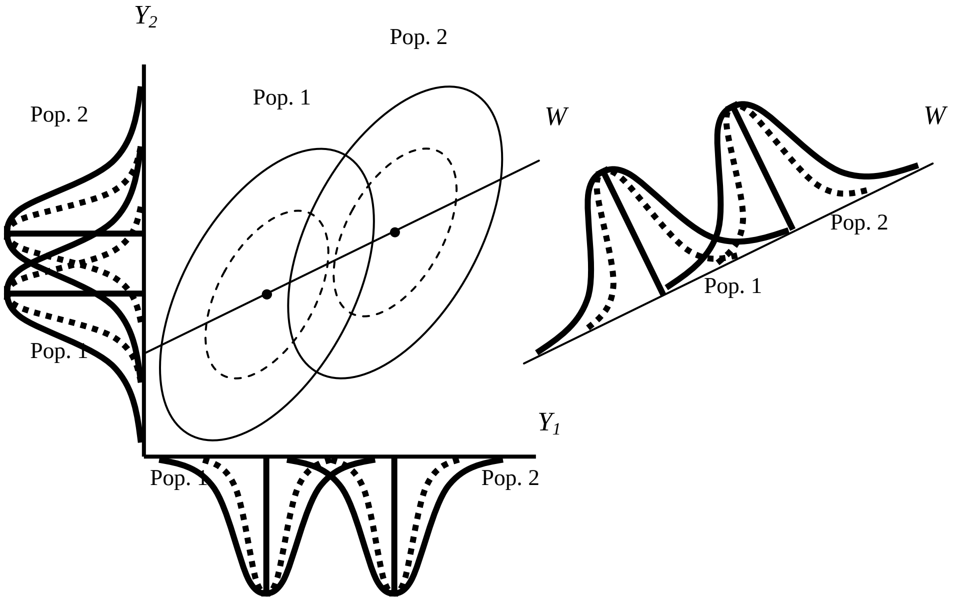


Figure 3. Multivariate population difference on measured variables and underlying constructs.

effect size on *W* was depicted as near 3, the latent standardized effect size for η_W, $d_{\eta_W} = (\mu_{1\eta} - \mu_{2\eta}) / \sigma_{\eta_W}$, is approximately 5. The square of this effect size, $d^2_{\eta_W}$, is also the squared Mahalanobis' distance

$$D^2_{\eta_W} = [\boldsymbol{\mu}_{1\eta} - \boldsymbol{\mu}_{2\eta}]'\boldsymbol{\Sigma}^{-1}_{\eta_{within}}[\boldsymbol{\mu}_{1\eta} - \boldsymbol{\mu}_{2\eta}], \qquad (7)$$

which corresponds to Equation 6 with the error variance $\boldsymbol{\Sigma}_{\varepsilon_{within}}$ removed. In fact, the reliability of the composite *W* could be determined as $D^2_W / D^2_{\eta_W}$, which is just a multivariate restatement and rearrangement of Equation 3.

To get a sense of the impact of measurement error on the multivariate effect size, consider a simple scenario in which the η_i constructs are uncorrelated (and hence so too are the Y_i variables). In this case the matrix $\boldsymbol{\Sigma}_{Y_{within}}$ is diagonal, and thus Equation 5 may be shown to simplify to

$$D^2_W = \mathbf{d}_Y{'}\mathbf{d}_Y = \sum_{i=1}^{m} d^2_{Y_i}, \qquad (8)$$

where $\mathbf{d}_Y$ is the vector of standardized effect sizes for Y_i (i=1…m) as per Equation 1. The same logic would also yield a parallel result for the latent effect size:

$$D^2_{\eta_W} = \mathbf{d}_\eta{'}\mathbf{d}_\eta = \sum_{i=1}^{m} d^2_{\eta_i}, \qquad (9)$$

where $\mathbf{d}_\eta$ is the vector of latent standardized effect sizes for η_i (i=1…m) as per Equation 2. Taking each Y_i variable's measurement error into account following Equation 3, Equation 8 yields

$$D^2_W = \sum_{i=1}^{m} d^2_{\eta_i}(\rho_{Y_iY_i}). \qquad (10)$$

If all Y_i variables were of the same reliability ρ_{YY}, it further follows that

$$D_W^2 = (\rho_{YY})D_{\eta_W}^2 \qquad (11)$$

(again, given uncorrelated η_i constructs and homogeneous reliabilities). Assuming a desired level of power (e.g., .80) and a specific effect size at the latent multivariate level (e.g., d_{η_W} =.30), the number of subjects per group for a two-sample test can be shown to be inversely proportional to measured variable reliability for all but the smallest sample sizes (in this highly restrictive example). That is,

$$n_W = (1/\rho_{YY})n_{\eta_W}\,. \qquad (12)$$

More generally, given any correlational pattern among the η_i constructs (and resulting attenuated correlations among the Y_i variables), the resulting reliability ρ_{WW} of the composite W would yield the corresponding relation

$$n_W = (1/\rho_{WW})n_{\eta_W}\,. \qquad (13)$$

Thus, the more reliable the composite *W*, the more MANOVA's power tends toward that of a theoretical test directly on the underlying construct.

In the univariate case, two implications of measurement error were highlighted: underestimating the magnitude of the treatment effect on the underlying construct of interest, and decreased power to detect the treatment effect. As illustrated, these hold as well for the multivariate case. However, while we may tend to gain power by accommodating multiple measured variables simultaneously, it is here that we must remind ourselves of our purpose, of precisely what truth we seek. That is – what, exactly, is the construct of interest in MANOVA?

Figure 4, a conceptual path diagram for the multivariate case, will help this discussion. On the left is a group code variable (e.g., dummy) representing population membership and whose influence is being assessed. As depicted, population membership has a potential bearing on the η_i constructs underlying the measured Y_i variables. Portions of the constructs not explained by population membership are represented in the latent residual influences ζ_i, which are likely to be correlated (shown in Figure 4 by shared two-headed arrows). Population differences on the measured variables are the observable manifestations of differences on the true underlying constructs of interest. The connection between each η_i and Y_i reflects the (square root of the) reliability of each variable; the weaker such a relation the less well the population differences on a construct are propagated to, and thus reflected in, the observed variables. As a result of each variable's imperfect operationalization of its construct, error ε_i contributes to the variable as well. Finally, in the case of multiple outcomes, a discriminant function *W* is represented as a composite of the measured variables. The weights determining this composite are optimal in the sense that they maximize the relation between *W* and *X*. Note that *W*, as a weighted sum of measured variables, is also a weighted sum of constructs and errors. That is, unless all variables are perfect operationalizations of their constructs, the composite *W* will contain measurement error which thus hampers its ability to reflect population differences propagated by *X*.

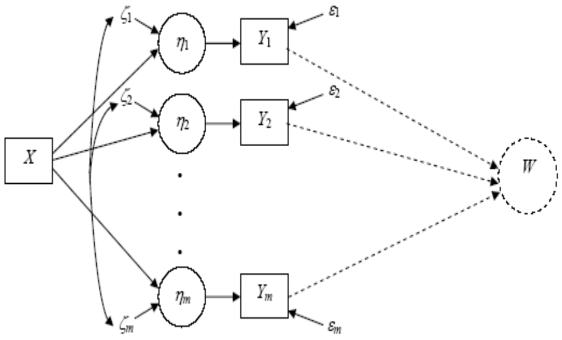


Figure 4. Path model for multivariate case, with *m* constructs.

So if *W* contains measurement error, with respect to what construct does that measurement error exist? The answer, as utilized previously, is the composite implicitly formed by MANOVA of the constructs underlying the variables. But what truth does a composite of univariate constructs represent? To this critical question there seems to be three common answers, none of which is entirely satisfactory. Each will be presented in turn, along with the concerns it inspires.

Position 1: The composite is not itself intended to be a construct; rather, it is merely a vehicle for the simultaneous examination of the *m* individual constructs of interest.

Response 1: If the separate constructs are of interest, then a MANOVA is inconsistent with that interest. Rather, a collection of individual ANOVAs, however seemingly inelegant, would address each construct directly. An omnibus MANOVA is not generally appropriate as a Type I error control mechanism since a single false univariate null hypothesis renders the multivariate null hypothesis false, and thus control over other true univariate ("partial") nulls becomes ungoverned. If one wishes to invoke an error control mechanism at the level of the constructs of interest, such as that of Bonferroni or his descendants, it may be applied across ANOVAs.

Position 2: The univariate constructs are facets of a single meaningful whole, as represented by the discriminant function and upon which knowledge of population differences is sought.

Response 2: Measured variables having a deterministic and defining bearing on a construct have been referred to as constituting an *emergent variable system* (e.g., Bollen & Lennox, 1991; Cohen, Cohen, Teresi, Marchi, & Velez, 1990). For example, one could imagine an unmeasured construct representing stress, contributed to and defined by such variables as relationship with parents, relationship with spouse, and demands of the workplace. In this case population differences in stress might indeed be of interest.

However, the formation of the discriminant function is not done in a manner reflecting any relative theoretical contributions of the three measured variables. If population differences existed only in terms of demands of the workplace, for example, then the discriminant function would be composed of only that variable. But does that then mean that stress is only a function of demands in the workplace? Surely not. Thus, while the notion is reasonable that variables combine to define a composite with a meaningful underlying construct, those variables' combination is not informed by the theoretical soundness of the construct, but rather only by measured variable mean differences. Forming a meaningful composite and then conducting an ANOVA on the resulting scores would seem more consistent with the beliefs underlying this variable system.

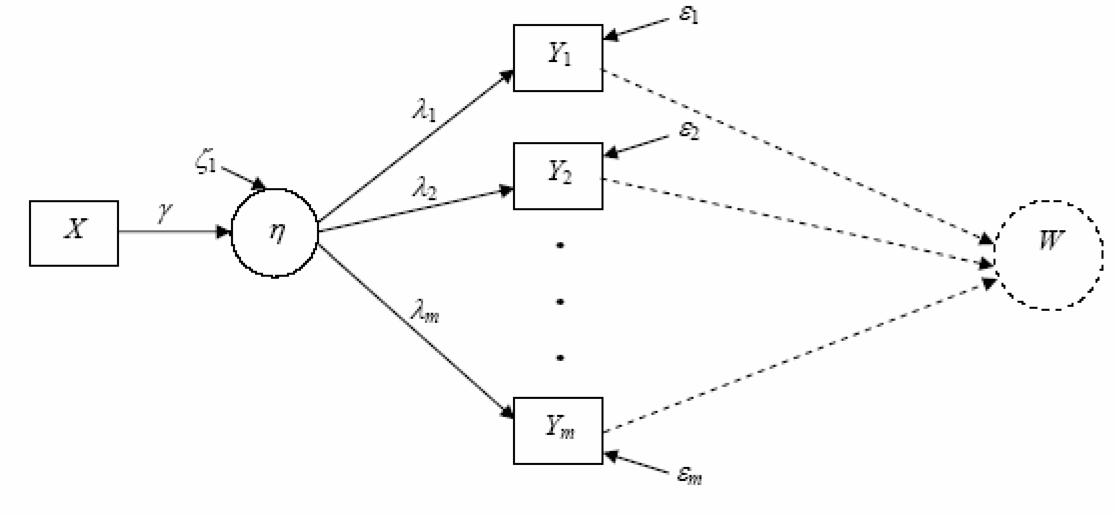
Position 3: The univariate constructs are actually a single meaningful underlying construct; the discriminant function represents that construct and allows for the assessment of population differences thereon.

Response 3: Contrary to the emergent variable system described in Response 2, the variable system here is latent. That is, all measured variables are believed to be undergirded by the same construct (but perhaps varying in the quality of their reflection), and it is on this common construct that inference is desired. Still, although a single construct exists, MANOVA remains clouded in its ability to address this construct directly.

Consider Figure 5, where X codes population membership and has a potential bearing γ on the common construct η underlying the measured Y_i variables. Thus, population mean differences on the measured variables are the observable manifestations of a population difference on the true underlying construct of interest. Again, the connections between the η construct and Y_i variables (λ_i) embody the (square root of the) reliability of each variable; the weaker such a relations the less well the group differences will be reflected in the observed variables. Finally, the discriminant function W is again shown as an optimal composite of the measured variables, where every variable in the composite contributes some part η and some part ε_i. So the discriminant function has succeeded to some extent in being a reflection of a construct of interest; however, it has still failed to eradicate error.

Further, the function has used group mean differences to guide its definition rather than proximity of construct operationalization. Thus, even if a single common construct underlies the measured variables, measurement error within this multivariate approach will continue to compromise the accuracy of a treatment effect's assessment as well as the power to detect that effect. That is, we must continue the search for methods that attempt to pardon error.

Figure 5. Path model for multivariate case, with one construct.



Pardon Error

Having cursed the varying degrees of darkness inherent in traditional univariate and multivariate experimental analyses, I now wish to light a candle – or more accurately, introduce the candle others have lit (e.g., Muthén, 1989; Sörbom, 1974). The foundation for this illumination may be seen in Figure 5, already presented. Our real goal is not to be able to detect an overall relation between the population membership X and the discriminant function W, but rather between X and the construct η. That is, we desire an estimate of the path denoted as γ, making the discriminant function W irrelevant. Fortunately, under the umbrella of structural equation modeling, a clearer attempt at a solution exists.

In Figure 5 the relations between the construct and its measured operationalizations may be expressed in a system of m structural equations of the form $Y_i = \lambda_i\eta + \varepsilon_i$ (i=1…m). These *measurement equations* may in turn be represented collectively as

$$\mathbf{Y} = \Lambda\eta + \boldsymbol{\varepsilon}, \quad (14)$$

where $\mathbf{Y}$ is a subject's mx1 vector of Y_i scores, Λ is an mx1 vector of unstandardized factor loadings generally assumed to hold for all subjects in both populations (homogeneity of measurement), and $\boldsymbol{\varepsilon}$ is a subject's mx1 vector of ε_i measured variable residuals. More interestingly, the theoretical relation of our current focus is contained in the structural equation relating population membership to the construct,

$$\eta = \gamma X + \zeta. \quad (15)$$

These structural equations, along with the simplifying (but not mandatory) assumption of independence of all exogenous elements (X, $\boldsymbol{\varepsilon}$, and ζ), have implications for the partitioned variance-covariance matrix Σ containing the X and Y_i variables for all populations combined. Specifically, for the Y_i variables alone, Equation 14 implies

$$\Sigma_Y = \Lambda\phi_\eta\Lambda' + \Theta_\varepsilon\,, \quad (16)$$

where ϕ_η is the total construct variance for both populations combined, and Θ_ε is the mxm

variance-covariance matrix for the ε_i residuals. Equation 15 has implications for ϕ_η, such that

$$\phi_\eta = \gamma^2 \sigma_X^2 + \psi \, , \qquad (17)$$

where σ_X^2 is the variance of X and ψ is the variance of the construct residual ζ. That is, ψ is the part of the construct variance that is not explained by population membership; as such, it is the pooled within-groups variance for the construct. Finally, the portion of the covariance matrix relating the vector $\mathbf{Y}$ of Y_i variables to X, as following from Equations 14 and 15, is

$$\mathbf{\Sigma}_{XY} = \gamma\sigma_X^2 \mathbf{\Lambda}' \, . \qquad (18)$$

As implied by the model in Figure 5, the full partitioned matrix for the X and Y_i variables (respectively) is:

$$\mathbf{\Sigma} = \left[\begin{array}{c|c} \sigma_X^2 & \gamma\sigma_X^2 \mathbf{\Lambda}' \\ \hline \mathbf{\Lambda}\gamma\sigma_X^2 & \mathbf{\Lambda}[\gamma^2\sigma_X^2 + \psi]\mathbf{\Lambda}' + \mathbf{\Theta}_\varepsilon \end{array} \right] . \qquad (19)$$

Using maximum likelihood estimation within structural equation modeling (see, e.g., Bollen, 1989), and after fixing one factor loading to a value of 1 so as to give the construct η a unit of measurement (i.e., that of the corresponding indicator variable), population values for all parameters in Equation 19 are chosen so as to maximize the likelihood of the observations giving rise to the sample covariance matrix **S**. After conducting an assessment of the data-model fit as represented by the degree of correspondence between the observed matrix **S** and the expected matrix $\hat{\mathbf{\Sigma}}$ (after substituting the optimum parameter values into Equation 19), satisfactory fit allows one to proceed to the question at hand. That question involves the estimation of, and statistical test of, the population mean difference(s) on the construct η.

For the two-group case, the path from the single dummy variable X to the construct η is an estimate of the population difference on the construct. This path, γ, will also have a maximum likelihood standard error as a by-product of the estimation process, which will allow a statistical test of the difference between the two population means on the construct η. If X is coded 0/1, then a statistically significant and positive estimate of γ implies the population coded X=1 has a higher mean on the construct η, whereas a negative value would imply superiority of the population coded X=0. An interpretation of the value of γ itself is not generally useful because it reflects the metric that η has been assigned by fixing a variable loading to 1. However, given that the pooled within-groups construct variance ψ has been estimated as well, we may derive an estimate of the latent standardized effect size d_η, where

$$d_\eta = \gamma / \sqrt{\psi} \, . \qquad (20)$$

Thus, if a single construct underlies our measured variables, we are able to conduct a statistical test on the construct mean difference as well as estimate the standardized effect size associated with that differences in latent means.

The simple process described above, which may be conducted using any structural equation modeling software (e.g., AMOS, EQS, LISREL, Mplus), is part of a larger class of models known as multiple-indicator multiple-cause (MIMIC) models suggested for assessing latent population differences (Muthén, 1989). The procedure is not without its own assumptions and restrictions, some of which may be softened in a somewhat more complicated strategy known as structured means modeling (Sörbom, 1974). Those details are left for the interested reader, and are summarized didactically elsewhere (e.g., Hancock, in press). More importantly is that these methods exist to put the construct back at center stage, in terms of hypothesis testing and effect size estimation, and as such the theoretical benefits over a MANOVA approach should be clear.

We may also take a practical approach in comparing the MIMIC and MANOVA strategies by determining the sample sizes required to detect a specific latent standardized effect size in order to achieve a desired level of statistical power. In Table 1 we see the cases of m=2, 3, and 4 measured variables, crossed with homogeneous sets of standardized loadings of λ=.4, .6, and .8. The standardized latent effect sizes included were d_η=.2, .5, and .8. In all conditions the necessary sample size was assessed for both MANOVA and the MIMIC approach in order to achieve .80 power using the equivalent of a two-tailed test at the .05 level. For MANOVA, sample size determination in each case followed strategies for Hotelling's T^2 outlined by Cohen (1988; Section 10.3.2.1), while for the MIMIC approach the methods derived by Hancock (2001) were used.

Table 1
Sample Size Required For Two-Group .05-Level Tests With Power=.80

		MIMIC			MANOVA		
		λ=.4	λ=.6	λ=.8	λ=.4	λ=.6	λ=.8
m=2	d_η=.2	1424	742	504	1748	912	619
	d_η=.5	229	120	81	281	148	101
	d_η=.8	90	47	32	111	59	41
m=3	d_η=.2	1080	626	467	1502	871	650
	d_η=.5	174	101	76	242	141	106
	d_η=.8	68	40	30	96	57	43
m=4	d_η=.2	909	568	449	1383	865	684
	d_η=.5	146	92	73	224	141	112
	d_η=.8	58	36	29	89	57	45

Many points are noteworthy in Table 1. As expected, for both the MIMIC and MANOVA methods the necessary sample size decreases as effect size increases (holding all else constant). Specifically, sample size decreases were approximately proportional to corresponding increases in the square of d_η (e.g., from d_η=.2 to d_η=.5, sample size necessary decreases by a multiplicative factor of $.2^2/.5^2$=.16). Sample size also decreases for both methods as loading magnitude increases (holding all else constant). In particular, sample size decreases were approximately proportional to corresponding increases in construct reliability as measured by coefficient H (also known as *maximal reliability*), where for the case of homogeneous loadings H mirrors the Spearman-Brown prophecy formula as

$$H = m\lambda^2/[1+(m-1)\lambda^2] \qquad (21)$$

(see Hancock, 2001). For example, with m=3 variables, H=.276 for λ=.4 and H=.529 for λ=.6; sample size thus decreases by a multiplicative factor of .276/.529=.521 for both the MIMIC and MANOVA strategies. For MANOVA this sample size decrease is due to the increased presence of the construct in the discriminant function; for the MIMIC approach, which already operates at the construct level, this sample size decrease is due to a decrease in the standard error associated with the γ path.

With regard to increasing the number of variables, for the MIMIC strategy sample size decreases correspondingly (holding all else constant); this is because distributional noncentrality varies directly with construct reliability as measured by H (Hancock, 2001), which increases with the addition of any nonzero loading. For MANOVA, sample size decreases with additional variables for λ=.4 and .6, but an increase in required sample size is observed for

λ=.8. This is because at some point additional variables do not contribute sufficient new information about the construct to justify the additional degree of freedom expenditure. This was seen in a supplemental analysis as well using λ=.9 (not shown in Table 1), where for *m*=2, 3 and 4 the necessary sample size per group for MANOVA increased from 540 to 590 to 635, respectively.

Overall, as expected the sample size required for MANOVA was always greater than for MIMIC. For the *m*=2 case MANOVA sample sizes were always about 23% larger than for the MIMIC approach. For *m*=3 that number increased to around 39%, while for *m*=4 required sample sizes for MANOVA were approximately 52% larger than for the MIMIC strategy. Thus, not only has the MIMIC approach's estimation and inference operated directly at the level of the construct of interest, it has done so with the same power for a considerable savings in sample size (or with greater power for the same sample size expenditure). And interestingly, at no point did we need to estimate variables' reliability; this information was implicit within the MIMIC process in the estimation of the λ_i loadings.

Extensions to this latent approach exist both internally and externally, where the former refers to methods for answering the same questions under less restrictive assumptions and the latter refers to methods for addressing more complex questions. With regard to internal extensions, the primary assumption implicit in MIMIC modeling is that, because the data from the groups are combined and only one model results, the same measurement model holds across populations. This includes loadings, construct variance, and error variances. In effect, all sources of covariation among observed variables are assumed to be equal in all populations, making the assumption of identical measurement models tantamount to an assumption of equal variance/covariance matrices (as is actually assumed in MANOVA as well). As alluded to previously, a more flexible approach to assessing latent means exists in structured means modeling (Sörbom, 1974), where only the corresponding loadings are commonly constrained across populations in the complete covariance model. Further, additional flexibility may exist to allow for some loading differences across populations under particular configurations of partial measurement invariance (Byrne, Shavelson, & Muthén, 1989).

Externally, the methods of assessing latent means may be extended greatly. Within the MIMIC framework, the creative use of group code predictors of the latent construct of interest (e.g., dummy variables) can fairly easily facilitate inferences that parallel those of more complex one-way and factorial ANOVA designs. Also, covariates may be introduced along with the group code variables. In fact, like all other variables covariates have underlying constructs; as such, given multiple indicator variables a latent covariate construct may be incorporated into the model along with the group code variables. The disattenuation of measurement error in the covariate provides greater accuracy in the assessment and testing of the covariate's predictive role in the design, as well as of population mean differences on the outcome construct after exacting such latent control.

Seeking Your Fortune

Inspired jointly by ancient wisdom and modern analytical methods, this article has attempted to return our focus to the constructs that underlie our experimental research endeavors. Certainly those constructs must be grounded in observable measures, but the proximity of those measures' operationalization of the construct(s) should be acknowledged and even accommodated. I have attempted to highlight the theoretical and practical costs of imperfect operationalization within traditional experimental analyses, and pointed toward reasonably accessible strategies that circumvent our measures' necessary imperfections.

But there is no free lunch, so to speak. Although the latent variable approaches to experimental design can pardon error and thus attempt to correct for unreliability, researchers are not thereby absolved of expending considerable effort in choosing or constructing quality measures. Poor reliability in measures yields less stability in the constructs and in estimates of their relations with other variables (e.g., group code variables), as well as larger standard errors for the statistical assessment of estimated relations. Thus, the methods described briefly herein serve to complement sound principles of instrument selection and construction.

These methods also signal the potential to reframe other aspects of the multivariate general linear model as well. Although this article has focused on experimental design, the canonical correlation model suffers from some of the same problems as MANOVA. Specifically, while *X* and *Y* variables are generally chosen by researchers with some constructs in mind, *X* and *Y* composites are formed whose primary allegiance is to the maximization of *XY* relations. If one used variables to define constructs in separate *X* and *Y* measurement models, the relations between constructs would be directly couched in theory, disattenuated of measurement error, and detectable with considerably more power than within the canonical framework. Expositions similar to those provided here for experimental design could be crafted, and would be equally compelling.

In sum, although constructs and their relations are the beloved truths that motivate most applied statistics, so many of our analytical efforts are hindered in their inferential estimation and hypothesis testing by our measures' inability to reflect those constructs satisfactorily. The current article has illustrated the detriments of failing to pardon error from our experimental inference, and has directed the applied researcher toward more modern methods that can assist researchers in getting closer to the truths they seek. It is my hope that they will pursue these and related methods as they seek their research fortunes. In the mean time, I believe I have a lunch appointment….

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley & Sons.

Bollen, K. A., & Lennox, R. (1991).Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, P., Cohen, J., Teresi, M., Marchi, M., & Velez, C.N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement*, *14*, 183-196.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373-388.

Hancock, G. R. (in press). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: SAGE Publications.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.