School of Economic Sciences

Working Paper Series
WP 2010-15

Measurement Invariance and Response Bias: A Stochastic Frontier Approach

By

Robert Rosenman, Vidhura Tennekoon, and Laura G. Hill

September 2010

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.

Measurement Invariance and Response Bias: A Stochastic Frontier Approach

Robert Rosenman, Vidhura Tennekoon, and Laura G. Hill

Washington State University

Author Note

Robert Rosenman and Vidhura Tennekoon, School of Economic Sciences, Washington State University; Laura G. Hill, Department of Human Development, Washington State University. This research was supported in part by a grant from the National Institute of Drug Abuse (R21-DA 025139-01Al). We thank the parents and facilitators who participated in the program evaluation. Correspondence concerning this article should be addressed to Robert Rosenman, School of Economic Sciences, Washington State University, Pullman WA 99164. Email: yamaka@wsu.edu.

Abstract

The goals of the present paper were to assess measurement invariance using a common econometric method   and to illustrate the approach with self-reported measures of parenting behaviors before and after a family intervention. Most recent literature on measurement invariance (MI) in psychological research 1) explores the use of structural equation modeling (SEM) and confirmatory factor analysis to identify measurement invariance, and 2) tests for measurement invariance across groups rather than across time. We use method, Stochastic Frontier Estimation, or SFE, to identify response bias and covariates of response bias both across individuals at a single point in time and across two measurement occasions (before and after participation in a family intervention). We examined the effects of participant demographics (*N* = 1437) on response bias; gender and race/ethnicity were related to magnitude of bias and to changes in bias across time, and bias was lower at posttest than at pretest. We discuss analytic advantages and disadvantages of SFE relative to SEM approaches and note that the technique may be particularly useful in addressing the problem of “response shift bias” or “recalibration” in program evaluation -- that is, a shift in metric from before to after an intervention which is caused by the intervention itself and may lead to underestimates of program effects.

KEYWORDS: Measurement invariance, measurement equivalence, response bias, response-shift bias, stochastic frontier analysis

Measurement Invariance and Response Bias: A Stochastic Frontier Approach

The goals of the present paper were to assess measurement invariance using a common econometric tool (Stochastic Frontier Estimation SFE), and to illustrate the approach with self-reported measures of parenting behaviors before and after a family intervention. Approaches to assessing measurement equivalence or measurement invariance (ME/I) have increased in pace with advances in software capabilities and have been extensively documented in recent years (for a review, see Vandenberg & Lance, 2000). The bulk of the literature on ME/I is devoted to structural equation modeling (SEM) and confirmatory factor analysis (CFA) or some extension of CFA, and the most common application is the test of ME/I across groups (e.g. gender, ethnic groups, control and treatment groups). A smaller literature has examined the problem of testing for ME/I in longitudinal studies, in which a measured attribute may change over time as a result of developmental processes (cf. Duncan, Duncan & Stoolmiller, 1994) or illness and changes in physical functioning (Oort, Visser, & Sprangers, 2009; Visser, Oort, & Sprangers, 2005).

A special problem of ME/I occurs when outcomes are self reported and a respondent's frame of reference changes across measurement points, especially if the changed frame of reference is a function of treatment or intervention. More specifically, an intervention may change respondents' understanding or awareness of the target concept and the estimation of their level of functioning with respect to the concept (Sprangers & Hoogstraten, 1989), thus violating measurement invariance. In fact, some treatments or interventions are <u>intended</u> to change how respondents look at the target concept. The bias arising from the change of frame of reference of respondents between two measurement points is commonly known as "response shift bias" (Howard, 1980). Response shift bias has been a particular concern of program evaluators for several decades, because a downward revision of one's abilities may occur when more rigorous internal standards are developed as a function of knowledge gained during the intervention; thus, a self-reported assessment at posttest may be lower than at pretest, even though knowledge or abilities targeted by an intervention have actually increased.

A common practice in program evaluation, used to minimize the problem of response shift, is to administer “retrospective pretests”, in which respondents estimate their pre-intervention functioning after completion of a program for comparison with posttest ratings (Bray, Maxwell, & Howard, 1984; Campbell, Stanley, & Gage, 1963). Because retrospective pretests are completed at the same time as posttests, presumably respondents’ metric for the two measures are the same and the assumption of ME/I holds. However, retrospective pretests are subject to a host of cognitive biases (Hill & Betz, 2005), including degradation of memory, effort justification bias, and the “Lake Woebegon” effect (people’s tendency to enhance present functioning in comparison to past functioning)(Kruger, 1999).

Past researchers have demonstrated the presence of response-shift bias by comparing self-ratings to objective performance and showing that self-ratings decreased from pre-intervention to post-intervention even though objective performance increased. However, in community-based program evaluation, experimental tests are impractical or impossible, and in such cases statistical approaches may be useful (Schwartz & Sprangers, 1999). In recent years, researchers have applied SEM to the problem of decomposing error in order to identify response shift bias. Oort (2005) described a statistical approach that identifies three types of response shifts that bias the aggregate estimate of true change: “Reconceptualization” means that the intervention has changed the factors which explain the self-rating; “reprioritization” means the intervention has changed the relative importance of the same factors which explain the self-rating; and “recalibration” is a change in the internal standards used by the respondent after the intervention, where this change is attributed to the intervention. In an application on cancer patients undergoing invasive surgery, Oort, Visser & Sprangers (2005) found that only recalibration changed the overall measure of the intervention on self-ratings of health-related quality of life.

In the present paper we suggest a different statistical approach to revealing and measuring the existence of response differences related to changes in respondents' frame of reference and, perhaps more importantly, identifying covariates of these differences. The suggested approach is based on Stochastic Frontier Estimation (Aigner, et al., 1977; Battese & Coelli, 1995; Meeusen & van den Broeck, 1977), a technique widely used in economics and operational research, and can be used to identify response shift when objective measurement is not feasible.

Our approach has two significant advantages over that proposed by Oort and colleagues- (2005). His approach requires a minimum of two temporal sets of observations on the self-rating of interest as well as multiple measures of the item to be rated, and reveals only *aggregate changes* in the responses. SFE, to its credit, can identify response *differences* across *individuals* (as opposed to simply aggregate response shifts) with a single temporal observation and a single measure, so is much less data intensive. Moreover, since it identifies differences at the individual level, it allows the analyst to identify not only that responses differ by individual, but what characteristics are at the root cause of the differences, allowing SFE to be used to systematically identify equivalents to Oort's three types of response shifts. Of course, changes at the individual level can be aggregated to a measure comparable to what Oort's reveals, and response *change* can be identified as long as more than one temporal observation is available for the respondents. SFE again has an advantage because such changes can be assessed for the individual, and the root causes again identified.

What may superficially be seen as two disadvantages to SFE when compared to SEM approaches are actually common to both methods. First, both measure response (and therefore response shift) against a common subjective metric established by the norm of the data. In fact, we term any systematic difference by an individual from this normal “response bias” (which of course, is different from response shift bias), and SFE allows analysts to identify covariates to this bias. With both SEM and SFE, if an objective metric exists, the difference between the self-rating and the objective measure is easily established. A second apparent disadvantage is that SFE requires a specific assumption of bias existing in only one direction (although it is possible to test this assumption statistically). SEM can reveal response shift without such a strong assumption, allowing individual respondents to change their bias in different directions – but aggregate changes become manifest only if “many respondents experience the same shift in the same direction” (Oort, 2005, p. 595). Hence, operationally the assumptions are nearly equivalent.

Response Bias, Response Shift, Response Bias Shift, and Response Shift Bias

Measurement invariance, a prerequisite for an unbiased estimate of a variable of interest, can be formally defined following Mellenbergh (1989) as $f\ (X \mid T = t, V = v) = f\ (X \mid T = t)$ where X denotes the observed measurement, T is the true attribute measured using X, and V represents the variables other than T . Mellenbergh (1989) introduced the above concept, the Principle of Conditional Independence (PCI), to define item bias. Oort (1991) showed that PCI subsumes a whole range of measurement issues. However, when X is a self-reported outcome and V includes (often unobserved) variables affecting the frame of reference used by respondents, PCI is not assured. Within this context, *response bias* is simply the case that $f\ (X \mid T = t, V = v) \neq f\ (X \mid T = t).$ The bias is upward if $f\ (X \mid T = t, V = v) > f\ (X \mid T = t)$ and downward if the inequality goes the other way.

Response shift can occur independently of response bias – in fact, response shift is usually the goal of the treatment or intervention. *Response shift* would mean that $f_1(X \mid T = t_1) \neq f_2(X \mid T = t_2)$ where the subscripts indicate the measured attribute before and after intervention; in other words, the change caused (and hopefully intended) by the intervention. *Response bias shift* is a change in measurement of the degree of bias from before to after treatment. Mathematically this would mean $\left[f_1(X \mid T = t, V = v) - f_1(X \mid T = t)\right] \neq \left[f_2(X \mid T = t, V = v) - f_2(X \mid T = t)\right]$ so the change in the bias from before and after the intervention confounds the measurement of the actual response shift. Unless the response bias shift is on average equal to 0, the result would be *response shift bias* (the downward revision of abilities, or recalibration in Oort's terminology) as an analyst tries to measure the average change caused by the intervention.

Measuring Response Bias

Our approach for measuring response bias and response bias shift is based on the Battese and Coelli (1995) adaptation of the stochastic frontier model (SFE) independently proposed by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977). Let

(1) $$Y_i = X_i\beta + e_i$$

where Y_{it} is the (latent) response variable, X_i are variables that explain the response. The term e_i is stochastic and can be decomposed as $e_i = \varepsilon_i - u_i$ where ε_i is a random error distributed iid $N(0, \sigma_\varepsilon^2)$ and u_i is a non-negative (truncated at zero) random variable which accounts for response shift away from a "frontier" response level and is distributed $N(\mu_i, \sigma_u^2)$ independent of ε_{it}. Moreover

(2) $$\mu_i = z_i\delta$$

where the z_i are variables that explain the specific deviation from the response frontier. Subscript i indexes the individual observation and, if relevant, we can add a t subscript for panel data. Under this characterization the "frontier" response is the response expected of all respondents sharing the same X_i. Thus, the expected response of observation i is given by

(3) $$E(Y_i) = X_i\beta - z_i\delta.$$

In equation (3) $z_i\delta$ identifies the one-sided observation-specific deviation from the response frontier. We note that $z_i\delta$ is the expected value of u_i. Hence, it represents the observation-specific response bias from the normal relationship between the Xs and Y. Treatment can affect both frontier response and the response bias. It affects the frontier response by changing the $X_i\beta$ part of (3). It changes the response bias by changing the $z_i\delta$ part of (3).

Identification requires that bias be one-sided – respondents either overestimate their response level or underestimate it.[1] Hence, researchers should have a prior belief of the direction of bias that is expected. Statistical analysis may be used to test this prior belief.

Using SFE for Measuring Response Bias Shift

When respondents undergo a treatment the response can change for two reasons. First, the treatment may have an effect (in the sense of changing a response, not necessarily in the desired direction). Such changes would be reflected in estimated βs in equation (3). A second

[1] If panel data are used the model can be identified as long as $E(\mu_{it})$ is constant for all t and does not equal 0.

reason that the response might change is that the response bias increases or decreases, which would be reflected by changes in the δs.

We considered two ways to use SFE to simultaneously estimate response bias shift. The first approach involves two separate SF estimations, the first for families before they have the program or treatment and the second for the same families after they have had the treatment. With this approach, as noted above, response *bias* for observation *i* is measured separately for each period by $z_i\delta$. Response *bias shift* with this approach is characterized by a change in the value of δs from before the program or treatment to after the program or treatment. Changes in the β are attributable to the program, and the change in the δ are attributable to response shift.

The problem with this first approach is that the stochastic frontier [the $X_i\beta$ part of (3)] is measured relative to the data; thus the determination of u_i is dependent on the data used. More specifically, u_i is a relative rather than absolute measure. Hence, since different data are used for the before- and after-treatment measurements, the scale of the bias estimate may change. To address this problem we instead estimated a single SFE using pre and post responses as separate observations. This allows the bias to be measured against a single relative metric. Treatment (as a dummy variable) enters the model as an X_i and also as a z_i. The estimated β coefficient on treatment (as one of the X_i) indicates the effect treatment has on functioning. The estimated δ coefficient on treatment(this time as one of the z_i) indicates how treatment has changed response bias. If $\delta = 0$ the response bias, if it exists, is not affected by the treatment. Cross terms of treatment and other variables (that is, slope dummy variables) may be used if the treatment is thought to change the general way these other variables interact with functioning.

An Application to Evaluation of a Family Intervention

We applied stochastic frontier to examine ME/I in program evaluations of a popular, evidence-based family intervention (the Strengthening Families Program for Parents and Youth 10-14, or SFP)(Kumpfer, Molgaard, & Spoth, 1996). Families attend SFP once a week for seven weeks and engage in activities designed to improve family communication, decrease harsh parenting practices, and increase parents' family management skills. At the beginning and end of a program, parents report their level of agreement with various statements related to skills and behaviors targeted by the intervention (e.g. "I have clear and specific rules about my child's association with peers who use alcohol"). Consistent with the literature on response shift, we expected that the assumption of ME/I across measurement occasions would be violated due to parents' changing frame of reference over the course of the program. Specifically, we hypothesized that non-random bias would be greater at pretest than at posttest as parents changed their standards about intervention-targeted behaviors and became more conservative in their self ratings.

Method

Sample

Our data consisted of 1437 parents who attended 94 SFP cycles in Washington State and Oregon from 2005 through 2009. Twenty-five percent of the participants identified themselves as male, 72% as female, and 3% did not report gender. Twenty-seven percent of the participants identified themselves as Hispanic/Latino, 60% as White, 2% as Black; 4% as American Indian/Alaska Native, 3% as other or multiple race/ethnicity, and 3% did not report race/ethnicity. Almost 74% of the households included a partner or spouse of the attending parent, and 19% reported not having a spouse or partner. For almost 8% of the sample the presence of a partner or spouse is unknown. Over 62% of our observations are from Washington State, with the remainder from Oregon.

Measures

The outcome measure consisted of 13 items assessing parenting behaviors targeted by the intervention, including communication about substance use, general communication, involvement of children in family activities and decisions, and family conflict. Items were designed by researchers of the program's efficacy trial and information about the scale has been reported on elsewhere (Spoth, Redmond, Haggerty & Ward, 1995; Spoth, Redmond & Shin, 1998). Cronbach's alpha in the current data was .85 at both pretest and posttest. Items were scored on a 5-point Likert-type scale ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree").

Variables used in the analysis, including definitions and summary statistics are presented in Table 1. The average family functioning, as measured by the change in self-assessed parenting behaviors from the pretest to the posttest, increased from 3.98 to 4.27 after participation in SFP.

Procedure

Pencil-and-paper pretests were administered as part of a standard, ongoing program evaluation on the first night of the program, before program content was delivered; posttests were administered on the last night of the program. All data are anonymous; names of program participants are not linked to program evaluations and are unknown to researchers. Procedures for the current study were issued a Certificate of Exemption from the Institutional Review Board of Washington State University.

Results

We used SFE to estimate (pre- or post-treatment) family functioning scores as a function primarily of demographic characteristics. Preliminary analysis indicated that the one-sided errors were downward. When we tried to estimate SFE with one-sided errors upward the procedure failed to converge. The same specification leaving out the constant term converged, but a null hypothesis of one-side errors was rejected with almost certainty. A similar analysis but with the one-sided errors completely random (rather than dependent on treatment and other variables) was also rejected, again with near certainty. Additional preliminary analysis of which variables to include among z_i (including a model using all the explanatory variables) led us to conclude that three variables determined the level of bias in the family functioning assessment – age, Latino/Hispanic ethnicity, and whether or not the functioning measure was a pre-test or post-test assessment.

SFE Total Effects Model

The results of the SFE are shown in Table 2. The Wald χ^2 statistic indicated that the regression was highly significant. Several demographic variables were found to influence the assessment of family functioning with conventional statistical significance. Males gave lower estimates of family functioning than did females and those with unreported gender. All non-White ethnic groups (and those with unreported race/ethnicity) assessed their family's functioning more highly than did White respondents. Most importantly, participation in the Strengthening Families Program increased individuals' assessments of their family's functioning.

We assessed bias, and its change, from the coefficient estimates for the δ parameters where $\mu_i = z_i\delta$. First the overall question was if, in fact, there was a one-sided error. Three measures of unexplained variation are shown in Table 2: $\sigma^2 = E(\varepsilon_i - u_i)^2$ is the variance of the total error, which can be broken down into component parts, $\sigma_u^2 = E(u_i^2)$ and $\sigma_\varepsilon^2 = E(\varepsilon_i^2)$. The statistic $\gamma = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$ gives the percent of total unexplained variation attributable to the one-sided error. To ensure $0 \le \gamma \le 1$ the model was parameterized as the inverse logit of γ and reported as inlgtgamma. Similarly the model estimated the natural log of σ^2, reported as lnsigma2, and used these estimates to derive $\sigma^2, \sigma_\varepsilon^2, \sigma_u^2$ and γ. As seen in the table the estimates for inlgtgamma and lnsigma2 were highly significant. Hence we found strong support for the one-sided variation that we call bias, and we saw that by far the most substantial portion of the unexplained variation in our data came from that source.

Three variables explained the level of bias. Latino/Hispanic respondents on average had more biased estimates of their family functioning. Looking again at equation (3) we see that this means they, relative to other ethnic groups, underestimated their family functioning. However, we found that older participants had smaller biases, thus giving closer estimates of their family's relative functioning. Of primary interest is the estimate of the Treatment effect. Participation in SFP strongly lowered the bias, on average.

The total change in the functioning score averaged 0.295. This total change consisted of two parts as indicated by the following equation:

(4) total change = measured prescore – measured postscore

= (real prevalue – prevalue bias) – (real postvalue – postvalue bias)

= real change – (postvalue bias – prevalue bias)

The term in parentheses is negative (the estimation indicates that treatment lowered the bias). Thus, the total change in the family functioning score underestimated the improvement due to SFP, although the measured post-treatment family functioning was not as large as it would seem from the reported family functioning scores, on average. Table 3 shows the average estimated bias by pre- and post-treatment, and the average change in bias, which was -0.133. Thus, the average improvement in family functioning was underestimated by this amount.

Table 4 shows the results of a regression on bias change and demographic and other characteristics. Males and Black respondents had marginally larger bias changes, while those with race/ethnicity unreported had smaller bias changes. Since the bias change was measured as postscore bias minus prescore bias, this means that the bias changed less, on average, for male and Black respondents, but more, on average, for those whose race was unreported.

SFE Model with Heteroscedastic Error

One alternative to the total effects model which generated the results in Table 2 is a SFE model which allows for heteroscedasticity in ε_i, u_i, or both. More precisely, for this model we maintained equation (3) but had $E(\varepsilon^2) = \omega_\varepsilon w_i$ and $E(u^2) = \omega_u w_i$ where ω_ε and ω_u re parameters to be estimated and w_i are variables that explain the heteroscedasticity. We note that w_i need not be the same in the two expressions, but since elements of ω_ε and ω_u can be zero we lose no generality by showing it as we do, and in fact in our application we used the same variables in both expressions, those that we used to explain μ in the first model. Table 5 reports the results of such a model. In this case the one-sided error we ascribe to bias is evident from statistically significant parameters in the explanatory expressions for σ_u^2.

We note first that the estimates in the main body of the equation were quantitatively and qualitatively very similar to those for the non-heteroscedastic SFE model. The only substantive change is that age was no longer significant at an acceptable p-value, and race unreported had a p-value of 0.1. All signs and magnitudes were similar. Once again, results indicated that participation in SFP (Treatment) strongly improved functioning. Additionally, treatment lowered the variability of both sources of unexplained variation across participants. The decreased unexplained variation due to ε is likely explained by individuals having a better idea of the constructs assessed by scale items. For our purposes, the key statistic here is the coefficient of treatment explaining σ_u^2. The estimated parameter was negative and significant with a p-value=0.03. Since the bias was one-sided we clearly can conclude that going through SFP lowered the variability of the bias significantly. Moreover, these estimates can be used to predict the bias of each observation, and with this model the average bias fell from 0.545 to 0.492, so while the biases were larger with this model, the decrease in the average (-0,63) was about one-half the decrease we saw in the first model.

Discussion

The general question of invariance of measurement, as Horn and McArdle (1992) noted, is whether under different conditions of observing and studying phenomena, the attributes of measures change. Without invariance of measurement we cannot unambiguously interpret any differences in a measured phenomenon, whether inter-temporal or across individuals or groups. Changing response bias is but one reason we might not see invariance of measurement.

Within this context, Vandenberg and Lance (2000) argued that researchers often invoke unstated assumptions about measurement equivalence in conducting tests of substantive hypotheses. Although rarely tested, these assumptions could be checked as extensions to the basic CFA framework, increasing the reliability and validity of such studies. All the possible tests of ME/I that they noted are variants of SEM or CFA, and are conceptually equivalent to the idea of comparing the outcome from multiple samples.

Vandenberg (2002) shed further light on the issue and raised additional questions: *"How do you know that the test for configural invariance (or any other ME/I test) is truly detecting* *changes in or differences between groups in terms of the conceptual frame of reference used to make responses?"* or "*How do you know that failure to support ME/I (i.e., differences exist) is due to differences or shifts in the measurement properties of the instruments, or is an artifact of some other influence?"* The first question, as he argued, is an issue of sensitivity and the second question is one of susceptibility. Schmitt and Kuljanin (2008), who reviewed practices of assessing ME/I for the years since Vandenberg and Lance (2000), indicated that these important questions are yet to be fully addressed.

The SFE approach presented here suggests a methodology which is fundamentally different from the tests and approaches reviewed in Vandenberg and Lance (2000) and Schmitt and Kuljanin (2008). Moreover it is robust to the issues of sensitivity and susceptibility raised by Vandenberg (2002). Unlike previous approaches, SFE is not a mere test of ME/I. Instead it provides a comprehensive tool to identify response *differences* across individuals, in addition to aggregate response shifts. Moreover, it helps to identify what characteristics are at the root cause of the differences, allowing SFE to be used to systematically to identify equivalents to the three types of response shifts presented in Oort et al. (2005). All these relative merits are achieved with a much less data-intensive approach, since SFE only requires a single temporal observation and a single measure. Of course, the response *change* can be identified as long as more than one temporal observation is available for the respondents. SFE again has an advantage because such changes can be assessed for the individual, and the root causes again identified.

An additional complication not addressed in existing alternative methodologies, discussed in Borsboom and colleagues (2008), is that *" measurement invariance (defined in terms of an invariant measurement model in different groups) is generally inconsistent with selection invariance (defined in terms of equal sensitivity and specificity across groups)"*(p.75).

The issue, to a large extent, is a problem of existing methodologies. However, SFE can address any selection invariance problem by incorporating a Heckman (1979) correction (or other corrections, such as instrumental variables) in the model.

The SFE method, however, is not without problems. The main limitation is that the estimates rely on assumptions about the distributions of the two error components. Model identification requires that one of the error terms, the bias term ($z_i\delta$) in our application, to be one-sided. This, however, is not as strong an assumption as it looks, for two reasons. First, the researcher has the option of choosing the direction of bias, either as non-negative or non-positive; often there is prior information or theory that indicates the most likely direction. Second, the validity of the assumption can be tested statistically.

As we noted earlier, the potential for response-shift bias, or downward recalibration of self-ratings as a function of material or skills learned in an intervention, has long been a concern to program evaluators as it may result in underestimates of program effectiveness (Howard & Dailey, 1979; Norman, 2003; Pratt, McGuigan & Katzev, 2000; Sprangers, 1989). However, in the absence of an objective performance measurement, it has not been possible to determine whether lower posttest scores truly represent response-shift bias or instead an actual decrement in targeted behaviors or knowledge (i.e. an iatrogenic effect of treatment). By allowing evaluators to test for a decrease in response bias from pretest to posttest, SFE provides a means of resolving this conundrum.

Conclusions

We presented SFE as a method to identify one cause of measurement variance, response bias and changes in response bias, within the context of self-reported measurements at individual and aggregate levels. Even though we proposed a novel application, the method is not new, and has been widely used in economics and operational research for over three decades. The procedure is easily adoptable by researchers, since it is already supported by several statistical packages including Stata (StataCorp, 2009) and Limdep (Econometrica Software, 2009).

Invariance of measurement has long been a key issue in psychometrics. However, almost all attempts to address the issue have been confined to using SEM to test for ME/ at the aggregate level. As noted in the introduction, our approach has three significant advantages over SEM techniques that try to measure response bias. SEM requires more data – multiple time periods and multiple measures, and measures bias only in the aggregate. SFE can identify bias with a single observation (although multiple observations are needed to identify bias shift) and identifies response biases across individuals. Perhaps the biggest advantage over SEM approaches is that because SFE identifies bias it provides information about the root causes of the bias. SFE allows simultaneously analysis about treatment effectiveness, causal factors of outcomes, and covariates to the bias, improving the statistical efficiency of the analysis over traditional SEM which often cannot identify causal factors and covariates to bias, and when it can requires two-step procedures. Perhaps most important, SFE allows the researcher to identify bias and causal factors at the individual level, expanding our ability to identify, understand, explain, and potentially correct for, response shift bias. Of course, bias at the individual level can be aggregated to measures comparable to what is learned through SEM approaches.

.

References

Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics, 6*(1), 21-37. doi: http://www.sciencedirect.com/science/journal/03044076

Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics, 20*(2), 325-332. doi: http://www.springerlink.com/link.asp?id=102505

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods, 13*(2), 75-98. doi: 10.1037/1082-989x.13.2.75

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational and Psychological Measurement, 44*(4), 781-804. doi: 10.1177/0013164484444002

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton, Mifflin and Company.

Duncan, T. E., Duncan, S. C., & Stoolmiller, M. (1994). Modeling developmental processes using latent growth structural equation methodology. *Applied Psychological Measurement, 18*(4), 343-354. doi: 10.1177/014662169401800405

Econometrica Software, I. (2009). LIMDEP Version 9.0 [Computer Software]. Plainview, NY: Econometrica Software, Inc.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*, 153-161.

Hill, L. G., & Betz, D. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26*, 501-517.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research. Special Issue: Quantitative topics in research on aging, 18*(3-4), 117-144.

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review, 4*(1), 93-106. doi: 10.1177/0193841x8000400105

Howard, G. S.,& Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144-150.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*(2), 221-232. doi: 10.1037/0022-3514.77.2.221

Kumpfer, K. L., Molgaard, V., & Spoth, R. (1996). The Strengthening Families Program for the prevention of delinquency and drug use. In R. DeV. Peters & R. J. McMahon (Eds.) *Preventing childhood disorders, substance abuse, and delinquency*. Banff International Behavioral Science Series, Vol. 3. (pp. 241-267). Thousand Oaks, CA, US: Sage Publications, Inc.

Masunaga, H., & Horn, J. (2000). Characterizing mature human intelligence: Expertise development. *Learning and Individual Differences, 12*(1), 5-33. doi: 10.1016/s1041-6080(00)00038-8

Meeusen, W., & van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review, 18*(2), 435-444.

Mellenbergh, G. J. (1989). *Empirical specification of utility functions*. Department of Psychology, University of Amsterdam, Amsterdam.

Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies. *Quality of Life Research*, *12*(3), 239-249.

Oort, F. J. (1994). Potentiale schenders van de eendimensionaliteit van psychologische meetinstrumenten. / Potential violators of the single dimensionality of psychological measurement instruments. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 49*(1), 35-46.

Oort, F.J. (1991), Theory of violators: assessing unidimensionality of psychological measures. In: Steyer, R.,Wender, K.F., Widaman, K.F. (eds.) *Psychometric Methodology,* 377–381. Fischer, Stuttgart (1991)

Oort, F.J. (2005) Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, 14(3), 587–598.

Oort, F.J., Mechteld R. M. Visser and Mirjam A. G. Sprangers (2005), An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, 14(3), 599–609.

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology, 62*(11), 1126-1137. doi: S0895-4356(09)00096-1.3

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation*, *21*, 341-349.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. [Article]. *Human Resource Management Review, 18*(4), 210-222. doi: 10.1016/j.hrmr.2008.03.003

Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine, 48*.

Spoth, R. (2007). Opportunities to meet challenges in rural prevention research: Findings from an evolving community-university partnership model. *The Journal of Rural Health, 23*(Suppl1), 42-54. doi: 10.1111/j.1748-0361.2007.00123.x

Spoth, R., Redmond, C., Haggerty, K., & Ward, T. (1995). A controlled parenting skills outcome study examining individual difference and attendance effects. *Journal of Marriage &amp; the Family, 57*(2), 449-464. doi: 10.2307/353698

Spoth, R., Redmond, C., & Shin, C. (1998). Direct and indirect latent-variable parenting outcomes of two universal family-focused preventive interventions: Extending a public health-oriented research base. *Journal of Consulting and Clinical Psychology, 66*(2), 385-399. doi: 10.1037/0022-006x.66.2.385

Sprangers, M. (1989). Subject bias and the retrospective pretest in retrospect. *Bulletin of the Psychonomic Society*, *27*(1), 11-14.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*(2), 265-272. doi: 10.1037/0021-9010.74.2.265

StataCorp. (2009). Stata Statistical Software: Release 11 [Computer Software]. College Station, TX: StataCorp LP.

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158. doi: 10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature : Suggestions , Practices , and Recommendations for Organizational Research. *Organizational Research Methods, 3*.

Visser, M. R. M., Oort, F. J., & Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: A convergent validity study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care &amp; Rehabilitation, 14*(3), 629-639. doi: 10.1007/s11136-004-2577-x

Table 1

Variable Names, Descriptions and Summary Statistics

Name	Description	*M*	*SD*
Pre-test functioning	Semi-continuous (0-5)	3.979	0.546
Post-test functioning	Semi-continuous (0-5)	4.273	0.461
Male	If Male=1	0.250	0.433
Gender missing	If gender not reported=1	0.030	0.170
White	If White=1	0.601	0.490
Black	If Black=1	0.023	0.150
Latino/Hispanic	If Latino/Hispanic=1	0.269	0.443
Native American	If Native American=1	0.040	0.195
Other	If Other race/ethnicity =1	0.034	0.182
Race missing	If race not reported=1	0.034	0.182
Age	Integer (17-73)	38.822	7.846
Partner or spouse	If Partner or spouse in family=1	0.736	0.441
Partner or spouse missing	If Partner or spouse in family not reported=1	0.077	0.266
Partner or spouse attends	If Partner or spouse attended SFP=1	0.499	0.500
Washington State	If family lives in Washington State=1	0.622	0.485

Table 2

Stochastic Frontier Estimation -- Total Effects Model

Variable	*β*	*SE*	*Z*	*p* < *Z*
Functioning				
Treatment	0.156	0.027	5.87	0.000
Male	-0.119	0.020	-6.03	0.000
Gender missing	-0.018	0.058	-0.30	0.760
Black	0.167	0.054	3.11	0.002
Latino/Hispanic	0.256	0.029	8.86	0.000
Native American	0.090	0.043	2.08	0.038
Other	0.174	0.045	3.83	0.000
Race missing	0.113	0.054	2.08	0.038
Age	-0.005	0.001	-3.92	0.000
Partner or spouse	-0.026	0.022	-1.18	0.237
Partner or spouse missing	-0.062	0.037	-1.70	0.090
Washington State	0.023	0.018	1.31	0.189
Constant	4.605	0.054	85.63	0.000
μ				
Treatment	-1.195	0.407	-2.94	0.003
Hispanic	1.100	0.383	2.87	0.004
Age	-0.052	0.028	-1.88	0.061

lnsigma2	0.291	0.201	1.00	0.317
inlgtgamma	2.559	0.263	9.72	0.000
σ^2	1.338	0.389		
γ	0.928	0.018		
σ_u^2	1.242	0.383		
σ_ε^2	0.096	0.010		

Wald χ^2(15)=331.46

Prob>χ^2 = 0.000

Table 3

Averages of Bias and Change

Variable	*M*	*SD*
Estimated u, pre-treatment	0.469	0.368
Estimated u, post-treatment	0.335	0.273
Change in u, post minus pre	-0.133	0.346

Table 4

Regression of Bias Change

Dependent variable: Change in Bias	*β*	*SE*	*t*	*p < t*
Male	0.050	0.023	2.19	0.029
Gender missing	0.100	0.064	1.55	0.122
Black	0.114	0.062	1.84	0.066
Latino/Hispanic	0.015	0.022	0.68	0.496
Native American	0.048	0.047	1.02	0.308
Other	0.078	0.051	1.54	0.125
Race/ethnicity missing	-0.147	0.061	-2.42	0.016
Age	0.003	0.001	2.74	0.006
Partner or spouse	0.032	0.028	1.13	0.258
Partner or spouse information missing	0.051	0.040	1.27	0.203
Washington State	-0.002	0.020	-0.11	0.912
Partner or spouse attended	-0.009	0.024	-0.36	0.721
Constant	-0.303	0.054	-5.65	0.000

Source	Sum of Square errors	df	F(12,1424)=2.4
Model	3.408042	12	Prob>F = 0.0044
Residual	168.2181	1424	R-squared=0.019
Total	171.6262	1436	

Table 5

Stochastic Frontier Estimation with Heteroscedasticity

Variable	β	*SE*	*Z*	*p*< Z
Functioning				
Treatment	0.222	0.032	6.94	0.000
Male	-0.098	0.019	-5.11	0.000
Gender missing	0.002	0.057	0.04	0.970
African Americans	0.159	0.054	2.95	0.003
Hispanic	0.344	0.035	9.95	0.000
Native American	0.096	0.042	2.27	0.023
Other	0.158	0.044	3.63	0.000
Race missing	0.090	0.053	1.69	0.091
Age	-0.001	0.002	-0.65	0.516
Partner or spouse	-0.027	0.021	-1.29	0.199
Partner or spouse missing	-0.044	0.035	-1.25	0.213
Washington State	0.017	0.017	0.98	0.325
Constant	4.532	0.088	51.55	0.000
Ln $(\sigma_{\varepsilon}^{2})$				
Treatment	-0.715	0.187	-3.81	0.000
Hispanic	-1.132	0.288	-3.94	0.000
Age	-0.007	0.010	-0.66	0.512
Constant	-1.906	0.434	-4.39	0.000

ln (σ_u^2)				
Treatment	-0.247	0.116	-2.13	0.033
Hispanic	0.913	0.123	7.42	0.000
Age	-0.005	0.007	-0.67	0.504
Constant	-0.761	0.319	-2.39	0.017

Wald $\chi^2(12)=253.60$

Prob>χ^2 = 0.000