Latent Class Analysis of Measurement Error in the Consumer Expenditure Survey

Clyde Tucker[1] Brian Meekins[1] Paul Biemer[2]
1 Bureau of Labor Statistics, 2 Massachusetts Ave, NE, Washington DC 20212
2 RTI International, PO Box 12194, Research Triangle Park, NC 27709

Abstract
Previous research by Tucker et al. (2010), working with the Consumer Expenditure Survey (CE), explores the factor structure of measurement error indicators such as: interview length, extent and type of records used, the monthly patterns of reporting, reporting of income, attempt history information, and response behavior across multiple interviews in a latent class model. Findings from this research, using approximately 30,000 cases from 2005 to 2009, yielded models with slightly poorer fit and less efficacy in predicting household expenditure than models using data from 1996 to 2001 (Tucker et al., 2008). While a number of revisions have been made to the CE in the years since 2001 that may have resulted in less measurement error overall, the differences in model fit and efficacy of the latent construct are worthy of further investigation. In current research we add the use of the information booklet as a possible indicator of measurement error. In addition, we examine the context of the model in much greater detail, examining subgroups where superior model fit and greater efficacy of the latent construct is observed. The description of the context extends beyond characteristics of the responding household to include the mode of administration of the survey (telephone, in-person, likely cell phone), as well as other process variables and their interactions with each other and the latent class measurement error construct.

Key Words: latent class model, measurement error, consumer expenditure

1. Introduction

This work is part of a continuing effort to identify sources of measurement error in the Consumer Expenditure Interview Survey (CEIS), a household survey of expenditure reports of a variety of different commodity categories (e.g. furniture, clothing, utilities, etc.). Previous efforts have used Markov Latent Class Models to analyze patterns of item missing (where respondents do not report an expenditure in that category), and latent class models to identify characteristics of poor reporting consumer units (CUs). This work extends previous research by further refining models, examining indicators not previously tested and validating these models with previous research using Markov Latent Class Analysis. Finally, a new model is proposed the fuses these two different types of models.

2. Consumer Expenditure Survey

The data used in this study consist of interviews collected in six years of the CEIS: 2005 through 2009. Each survey was designed to collect information on up to 95 percent of total CU expenditures. We define a CU as the members of a household who are related and/or pool their incomes to make joint expenditure decisions. In the CEIS, CUs are interviewed once every three months for five consecutive quarters to obtain the expenditures for 12 consecutive months. The initial interview for a CU is used as a bounding interview and these data are not used in the estimation. The survey is designed to collect data on major items of expense which respondents can be expected to recall for three months or longer. New panels are initiated every quarter of the year so that each quarter, 20 percent of the CUs are being interviewed for the first time. Only CUs completing and reporting an expense in wave 2 are used in this analysis, for a total of 29,347 respondents.

3. Previous Work

For panel surveys such as the CEIS, a related statistical method referred to as Markov latent class analysis (MLCA) is available, which essentially relaxes the requirement that the replicate measurements pertain to the same point. Thus, this method of analysis is feasible for analyzing repeated measurements of the same units at different time points available in panel surveys. MLCA requires a minimum of three measurements of the same units, as would be the case for a panel survey where units are interviewed on three occasions. The MLCA model then specifies parameters for both the period-to-period changes in the status of the item as well as the measurement error associated with measuring those changes.

Previous work by the authors used MLCA to make aggregate estimates of underreporting in a category only by respondents reporting no expenditures in that category. Biemer (2000) applied the MLCA methodology to the CEIS in order to determine whether useful information on the magnitudes and correlates of screening question reporting error can be extracted directly from the CEIS panel data. Biemer and Tucker (2001) extended the earlier analysis using data from four consecutive quarters of the CEIS by considering CUs that were interviewed four consecutive times beginning in the first quarter of 1996 and ending in the last quarter of 1998. This allowed the authors to consider a wider-range of models including second-order Markov models. First order Markov models assume that a purchase or non-purchase at quarter q is affected only by quarter q-1 purchases or non-purchases. A second order Markov model assumes that both quarters q-1 and q-2 affect purchasing behavior at quarter q. Their analysis provided evidence of second-order Markov effects and recommended that second-order terms be included in the models.

In Tucker, Biemer, and Vermunt (2002), model estimates with both unweighted and weighted data were compared. The results indicated that few differences were found between the two; therefore, given the ease of use, unweighted data were used in these analyses. A thorough examination of all explanatory variables considered in the previous studies was undertaken, and a reduced set of the most powerful ones was identified. A new diagnostic technique was developed and used to evaluate the validity of the models. In 2003, Tucker, Biemer, and Meekins developed methodology for estimating the amount of the missing expenditures.

Unlike the previous work, a micro-level approach incorporating measures specific to a given interview was used by Tucker, Biemer, Meekins, and Shields (2004) to examine underreporting for total expenditures. A latent variable that adequately accounted for the shared variance among a set of observed response error indicators was created. The observed indicators for this latent construct were based on information collected from each CU during the interview. The latent variable was believed to be a better measure of underreporting than any of the observed indicators taken individually. Each CU then was assigned to a particular class of the latent variable representing its hypothesized level of expenditure underreporting based on the CUs values on the observed variables. See Tucker (1992) for an earlier empirical example.

For this analysis the authors used only second interview data and examined reporters of expenditures while ignoring nonreporters. They wished to develop a model separate from covariates with only indicators of the quality of response. The authors began with the simplest identifiable model composed of three indicators (each with three classes) and a latent variable with three classes. From this point they ran all possible combinations of three indicators for a three class latent variable. The analysis was further extended by examining restricted models based on the hypothetical relationship of some of the indicators with the latent variable, thus ordering the latent classes in what we believed to be an interpretable manner. These "restricted" models were compared to the unrestricted models to aid in interpretability and choices of model fit. Some of the indicators are dichotomous. These were entered into the best three variable models along with other combinations to create four-indicator models. The goal was to develop a latent variable (preferably ordered) that indicated the quality of responses, such that poor reporters could be easily identified.

Using both objective and subjective measures of goodness of fit a "best" model was selected. Latent classes aligned with expenditure means as expected. Those with lower expenditure means had higher levels of underreporting. For example, those in the low underreporting class had a total expenditure mean of $10,625, while those in the high underreporting class had a mean of $6,948

In Tucker, Biemer, and Meekins (2005), the authors continued with a more in-depth exploration of micro-level measures of underreporting. In this analysis, only second wave data are used from those respondents actually reporting expenditures in the commodity classes under study (57,184 families interviewed in 1996 through 2001). Thus, we first were interested in the response errors for those respondents reporting expenditures and not those who said they had no expenditures in these categories. Again, the authors assumed response errors come largely in the form of underreports. In this case, a refined set of manifest indicators of response error were created.

For each of seven expenditure categories: children's clothing, women's clothing, men's clothing, furniture, electricity, minor vehicle expenses, and kitchen accessories, we began with the simplest identifiable model composed of three indicators and a latent variable with three classes. Models were again estimated using lEM. Only three manifest variables were used to maximize cell sizes in the manifest tables. We ran all possible combinations of three indicators for each expenditure class. The analysis involved both "restricted" and "unrestricted" models. Restricted models forced a hypothesized ordering of the manifest indicators to the latent response error (ordering the latent classes in what we believed to be an interpretable manner), while unrestricted models did not. Based on comparisons of the results from restricted and unrestricted models, it was decided to proceed with only restricted models from that point. Combinations of four and five manifest indicators were examined, but all models with more than four variables were of little value. Again, we ran models with several different sets of starting values to avoid reaching only a local solution.

The selection of the best model for each expenditure category was based primarily on the BIC and the Dissimilarity Index. The same set of manifest indicators were not used for the best model in each case, but the statistical diagnostics confirm a good fit for all final models chosen.

The authors also extended the use of substantive diagnostics used in earlier work. For each model they examined both conditional probabilities of the latent variable given each value of each indicator and the conditional probabilities of each indicator given each value of the latent variable. In addition, they also examined the actual probabilities of a case being in a particular latent class given its manifest cell location, as well as the proportion of cases assigned to each manifest cell by the latent class model. To gain a further understanding of the models, the authors again turned to the expenditure means for the three latent classes. The results, while not completely disconfirming, were not that promising. Across all seven categories of expenditures we analyzed, we found that the three classes of the latent variable failed to distinguish CUs based on their expenditures. However, for kid's clothing, women's clothing, and kitchen accessories, two separate groups could be identified that met our expectations. By including CUs that reported no expenditures in our analysis, the authors found that, for most commodities, mean expenditure amounts increased monotonically across the latent scale, and the three means were significantly different from one another.

Research by Tucker, Biemer, Meekins, Kesselman (2006) advanced the effort by examining a much larger number of commodity categories (29) and more rigorously examining and validating the results of the latent class models. The “final” model for each of the 29 commodities and overall were selected in a similar manner to past research, using both objective statistics and subjective diagnostic tools. Based on the results of the models CUs were then assigned to certain classes of reporting in the same way as previous research. The classification variable, corresponding to poor, fair, and good reporting quality was then regressed on a number of demographics in order to assess the content validity of the latent variable. After finding similar patterns across all commodity categories and verifying the results of the latent class modeling, the authors regressed the expenditure mean for each commodity category and overall on the latent classification controlling for key demographics, examining the contribution of the latent variable in predicting expenditure, controlling for demographic variables (such as one would use in weighting or nonresponse adjustment). Consistent with previous research the results of this research provided validation for the latent class approach to modeling measurement error, but a model that could differentiate levels of underreporting (given a report) remained elusive, while models classifying CU’s by whether they erroneously omit a report altogether were more successful.

Other research by Meekins, Tucker, and Biemer (2008) used the latent construct developed in Tucker et al. (2006) to examine the relationship of measurement error to subsequent wave nonresponse and bias. It was found that those in the poorest category of reporting were somewhat less likely to respond in subsequent interviews, volunteered expenditure reports in fewer categories, and had more sharply declining overall expenditure amounts in subsequent interviews than their counterparts in the fair and good reporting categories. Using these results Tucker, Biemer, Meekins (2009), included indicators that characterized the experience of the CU throughout the course of the entire panel (including panel nonresponse).

The three best fitting models were selected. Two models utilized three indicators, while one model used four indicators to examine the quality of expenditure reports. The three indicator models used the indicators: missing on income, length of interview, and average number of commodity categories to differentiate three and four latent classes of reporting quality. The four indicator model used all of those used in the three indicator models combined with the number of good interviews in the panel. Following the same process as prior research, these models were validated with demographic and process variables. The models showed good differentiation of expenditure estimates, even when controlling for demographics and process variables. Of particular note is the contribution of the interaction between income level and the latent class variable to this model. At very high or very low incomes the relationship between level of reporting and reported expenditure is significantly stronger. Indeed, the contribution of the interaction term is much higher than the direct effect of the latent construct. When examining bias (second quarter reported expenditure – fourth quarter reported expenditure), we find similar results. The variable derived from the latent class analysis showed good differentiation in the expected direction. The authors concluded that the latent construct was indeed measuring the quality of reporting, but either lacked the sensitivity needed to adequately predict underreporting, or measurement error is not a strong predictor of the average expenditure reported by the CU or the amount of bias, as measured by the coarse measure

In 2010, Tucker, Biemer, Meekins applied previous models and developed additional models using CE data from the second quarter of 2005 to the second quarter of 2009. The more current data also incorporate information collected in the Contact History Instrument (CHI). These data were not collected prior to the second quarter of 2005. These data capture a number of attributes of contact attempts made by field interviewers, including the number of contacts, mode of contact (phone or in-person), and reasons for refusal or noncontact that were recorded by the interviewer. The contact history data were refined into indicators for each wave and for the overall panel for each second quarter respondent. Reasons for refusal were grouped into factors relating to: privacy concerns, reluctance, and hostility. Reasons for noncontact were grouped into only two factors: gatekeepers or barriers and “other”.

Overall, the latent constructs did not perform as well as they did in previous research. While the model fit was still good, the CHI variables grouped more closely with themselves than with any other indicators and seemed to contribute little to the efficacy of the latent construct in predicting reporting error. Overall the strongest indicators across models were income missing, record use combined with interview length, number of completed interviews, reluctance due to time constraints, and average number of attempts. These indicators were also strong in previous research. As in previous research, the latent constructs lack the sensitivity needed to adequately predict underreporting, or measurement error, and are not strong predictors of the average expenditure reported by the CU and the amount of bias, as measured by total expenditure, or the proportional difference in Wave 2 and Wave 4 expenditure reports.

Unlike previous research analysis on the data from 2005 to 2009 did not consistently show differences in expenditure amounts in the expected direction across the three classes. For many commodity categories and for overall expenditure we can only differentiate between two classes of reporting quality. The current latent constructs appear to be relatively blunt instruments (although the only instruments we have), and are probably not useful for adjustment as they do not explain much of the variation in expenditure or change in expenditure.

4. Current Research

In an attempt to understand the difference in the results obtained during research conducted in 2010 on current data (2005 – 2009), compared to that which was conducted in prior years, the authors decided to reexamine the models and validate the results using the previous mention Markov Latent Class Analysis. In addition, the authors refined the models slightly incorporating one more indicator - the respondent’s use of the information booklet. Including that variable then, the indicators that are tested in the current latent variable factor models are as follows (with the coding scheme used for each):

1. Number of contacts the interviewer made to complete the interview (1=0-2; 2=3-5; 3=6+)
2. The ratio of respondents to total number of household members (1=<.5; 2>.5)
3. Whether the household was missing a response on the income question (1=present; 2=missing)
4. The type and frequency of records used. This variable indicates whether a respondent used bills or their checkbook to answer questions, and how often they did so. (1=never; 2=single type or sometimes; 3=multiple types or always)
5. The length of the interview (1<45min; 2=45-90; 3>90)
6. A ratio of expenditures reported for the last month of the 3 month reporting period to the total expenditures for the 3 months (1<.25; 2=.25-.5; 3=>.5)
7. A combination of type of record used and the length of the interview. (1=poor; 2=fair; 3=good) as shown below for the combined variable.

8. Number of expenditure questions within commodity category for which a response was imputed or allocated.
9. Use of the information booklet.
10. Number of completed interviews across the panel (1-4).
11. Pattern of attrition combined with the number of completed interviews (those with a pattern of attrition as opposed to a sporadic nonresponse pattern were further penalized).
12. Average number of commodity categories for which CU had at least one expenditure report.
13. The number of interviews in which the third month's expenditure to the quarter was between 0.25 and 0.5.
14. Panel averages of some of the interview level indicators.
15. Indication of respondent reluctance based on privacy concerns.
16. Indication of respondent reluctance based on time concerns.
17. Indication of an especially hostile refusal.
18. Any of the above indications of reluctance.
19. Incidence of noncontact based on gatekeepers or other barriers.
20. Incidence of noncontact based on other problems.
21. Any incidence of noncontact.
22. Proportion of attempts made in-person.
23. Proportion of completed interviews that were completed by phone.

Models were estimated using lEM, LCA software developed by Vermunt (1997). Model selection was based on a number of objective and subjective measures. The authors primarily used the Bayesian Information Criteria (BIC), the L^2 test statistic, and the dissimilarity index. However, for each model the authors examined the conditional probabilities of the latent variable given the value of each indicator. In this way we assessed the relative influence of each indicator and the degree to which an indicator effectively differentiated the respondents with respect to the classes of the latent variable.

The following six models from previous and current research were examined in greater detail in this work:

1. ***Old***: Income missing; record use/ interview length; average commodity categories; number of completed interviews
2. ***CHI***: Reluctance due to time constraints; tally of noncontact problems; average number of attempts
3. ***Combo***: All the "old" model indicators and average number of attempts
4. ***Reluctance***: Income missing; record use/interview length; number of completed interviews; reluctance due to time constraints
5. ***Noncontact***: Record use/ interview length; tally of noncontact problems; average number of attempts
6. ***New Combo***: All the "old" model indicators and the information book use.

For each of these final models the authors assigned survey participants to a given latent class based on probability of being in that class as given the value of the indicators. Table 1 shows the proportion of the CEIS Consumer Units in each latent class for each of the six final models, as well as the proportion of Consumer Units in each class of the previous "best" model (as tested on data from 1996 to 2001). For the most part (with the exception of the Reluctance model), the largest class in all models is the middle class.

Table 1: Proportion of Consumer Units in Each Latent Class by Model

	Poor	Fair	Good
Previous Best*	.203	.232	.565
Old	.222	.556	.222
CHI	.240	.428	.332
Combo	.149	.576	.275
Reluctance	.086	.298	.617
Noncontact	.243	.425	.332
New Combo	.223	.597	.180

*Estimated using data from 1996 - 2001

After assigning consumer units to latent classes, expenditure means were estimated for each class, for each model. Table 2 shows the overall quarterly expenditure by latent class, while Table 3, shows expenditure for a selected group of commodity categories for the New Combo model. This model, based on these expenditure means and goodness of fit statistics was determined to be the best performing model out of the six estimated with current data.

Table 2: Overall Expenditure Means by Latent Class

	Poor	Fair	Good
Previous Best	6,946.84	8,920.20	11,985.71
Old	10,683.82	10,797.76	16,832.02
CHI	12,916.41	12,666.16	10,840.80
Combo	11,772.09	10,720.57	15,803.90
Reluctance	10,803.81	10,857.06	12,911.35
Noncontact	12,211.46	13,065.22	10,842.66
New Combo	10,684.51	11,405.51	16,221.98

Table 3: Expenditure Means for Selected Commodities by Latent Class: New Combo Model

	Poor	Fair	Good
All expenditure	10,684.51	11,405.51	16,221.98
Electricity	275.55	326.24	341.25
TV and other electronics	93.70	83.84	172.55
Furniture	85.58	85.96	155.09
Kitchen Accessories	20.15	20.00	55.32
Men's apparel	72.08	63.90	92.58
Women's apparel	112.47	99.80	174.78
Kid's apparel	45.24	46.00	63.49
Minor vehicle repairs	56.32	59.22	97.09
Major vehicle repairs	37.76	47.08	81.16
Dental	44.49	59.01	124.70
Drugs	40.75	57.79	111.48
Pets	21.75	26.82	53.04

In addition, for the best performing models, the latent class variable was regressed on demographic variables in order to understand the nature of the variable. Table 4 shows the results of this analysis for the New Combo model. The majority of effects are quite small, although statistically significant. Of special note is the relationship between income and the latent class variable, where lower income CUs are less likely to be good reporters. Renters and those that are likely to have completed the interview over a cell phone (we can't ascertain this with certainty) are also less likely to be in the good reporting latent class.

Table 4: Latent Class Variable Regressed on Demographic Covariates

	Exp(b)	PR(X^2)
Famsize 1: one member	.936	.0119
Famsize 2: two members	1.035	.1171
Age	1.012	<.0001
Education	.927	<.0001

Inc rank1: Lowest 25%	.765	<.0001
Inc rank2: Middle 50%	.967	.1097
Race: White	1.234	<.0001
Renter	.835	<.0001
Urban	.990	.7664
Likely cell phone completion	.801	<.0001
Max-rescaled R^2	.054	

Table 5 summarizes the results of regressing the overall quarterly expenditure on the latent class variable as determined by the New Combo model together with demographic variable. An interaction of the latent class variable and income is also introduced. Note the decrease in the F statistic when the latent class is included in the model and the near incremental increase in the R^2 statistic. Also note that the mean expenditure for each latent class (LS Means) is no longer clearly distinguishable with the Poor and Fair categories not statistically significantly different, with the value of the mean expenditure for the Fair class is actually lower than that of the Poor class. Table 6 shows similar analysis for all commodity categories.

Table 5: Total Quarterly Expenditure Regressed on Latent Construct Controlling for Demographics

	Baseline model		With LV	
	Estimate	p-value	Estimate	p-value
F	1569.60	<.0001	878.40	<.0001
R^2	.41		.43	
F[Contribution of Latent variable]:			137.72	<.0001
F[Contribution of Interaction Term]:			16.81	<.0001

Total Expenditure: Least Squared Means Controlling for All Other Variables in the Model*

		p-values for differences in LSMean		
Class	LS Mean	Poor	Fair	Good
Poor	10,340.05		0.10	<.01
Fair	9,937.05	0.10		<.01
Good	12,361.58	<.01	<.01	

*Scheffe adjustment for multiple comparisons

Table 6: Least Squared Means by Latent Class Controlling for Demographics

	Poor	Fair	Good
All expenditure	10,340.47	9,937.05	12,361.58
Electricity	271.09	280.01	259.18
TV and other electronics	92.27	77.55	142.26
Furniture	88.49	73.81	116.22
Kitchen accessories	15.13	11.39	37.54
Men's clothing	54.65	52.33	64.43
Women's clothing	87.64	64.97	101.21
Kid's clothing	23.63	24.54	33.53
Minor vehicle repairs	53.57	49.82	73.95
Major vehicle repairs	29.25	40.53	60.70
Dental	41.06	32.62	79.60
Drugs/Pharm	43.70	47.86	83.49
Pets	20.26	16.03	31.86
Trash	17.64	13.81	21.33
Gas (HH)	92.73	93.57	101.52
Sports equipment	17.17	18.60	37.96
Eye care	18.19	18.60	32.00
Oil changes	16.75	17.03	21.52
Housekeeping	6.94	5.17	12.09
Major appliances	25.22	15.26	20.69

Very few of the mean expenditures are distributed across the latent classes in the expected direction. Yet the demographic analysis seem to show that the latent construct does capture measurement error. That is, the relationships that are observed between the latent class variable and the demographic variables are consistent with a variable that is capturing measurement error. In order to further assess the efficacy of the latent construct and validate the results of the models, the authors decided to incorporate the current factorial measurement error latent model into a Markov Latent Class Analysis.

Recall that the MLCA model then specifies parameters for both the period-to-period changes in the status of the item as well as the measurement error associated with measuring those changes. Therefore we would expect that those that are determined to be Poor reporters using the factorial latent class model to have low accuracy as determined by the Markov Latent Class Model. Table 7 confirms this result.

Table 7: Validation of Latent Construct with Markov Latent Class Analysis

P(A=1|W=1) Accuracy	**Category of Latent Class Reporter**		
	Poor	Fair	Good
Electricity	.9701 (.0034)	.9873 (.0028)	.9910 (.0016)
Gas (HH)	.9125 (.0080)	.9734 (.0032)	.9774 (.0053)
Trash	.9354 (.1861)	.9454 (.0053)	.9463 (.0060)
Housekeeping	.4823 (.0225)	.8241 (.0080)	.8536 (.0097)
Dental	.1825 (.0084)	.4622 (.0159)	.7865 (.0250)
Drug/Pharm	.2881 (.0074)	.6130 (.0114)	.8765 (.0139)
Eye care	.1131 (.0062)	.2767 (.0116)	.5457 (.0221)
Oil Changes	.5475 (.0158)	.7268 (.0168)	.7675 (.0222)
Minor Vehicle Repair	.1620 (.0062)	.3409 (.0108)	.5191 (.0169)
Major Vehicle Repairs	.1238 (.0125)	.2691 (.0254)	.4640 (.0446)
Computer	.2938 (.0288)	.3850 (.0274)	.5585 (.0432)
Television and other electronics	.5020 (.0189)	.7313 (.0273)	.7554 (.0218)
Sports equipment	.1840 (.0080)	.3439 (.0107)	.6416 (.0180)
Pet supplies	.2432 (.0100)	.5900 (.0118)	.7308 (.0144)
Major appliances	.1721 (.1181)	.3429 (.3745)	.5513 (.5586)
Kitchen items	.2167 (.0083)	.3886 (.0118)	.7044 (.0206)
Furniture	.1317 (.0114)	.2423 (.0198)	.4954 (.0376)
Men’s clothing	.5362 (.0787)	.7166 (.0180)	.7315 (.0243)

Women's Clothing	.8550	.7738	.8564
	(.3445)	(.0110)	(.0129)
Kid's clothing	.3366	.7019	.8017
	(.0099)	(.0078)	(.0097)

4. Discussion

The current factorial model then is very likely capturing measurement error in consumer expenditure reports. However, the ability of the variable to predict expenditure amounts is quietly small. Indeed, relative to other sources of error, and valid predictors of expenditure, the type of measurement error that is captured in this variable may be small. Measurement error due to reluctance seems to be captured by this latent construct, but measurement error due to cognitive process or misreporting of expenditure amount does not appear to be captured. In addition, the effects of noncontact, in particular, and nonresponse, in general, appear to be unintentionally combined. Future research will utilize more complex models like those in Figures 1 and 2, where W,X, Y, and Z are latent expenditures, A,B,C and D are indicators and ξ are measurement error constructs that may vary with time or may be consistent across time. In Figure 2, noncontact is modeled simultaneously further disambiguating the effects of measurement error on expenditure.

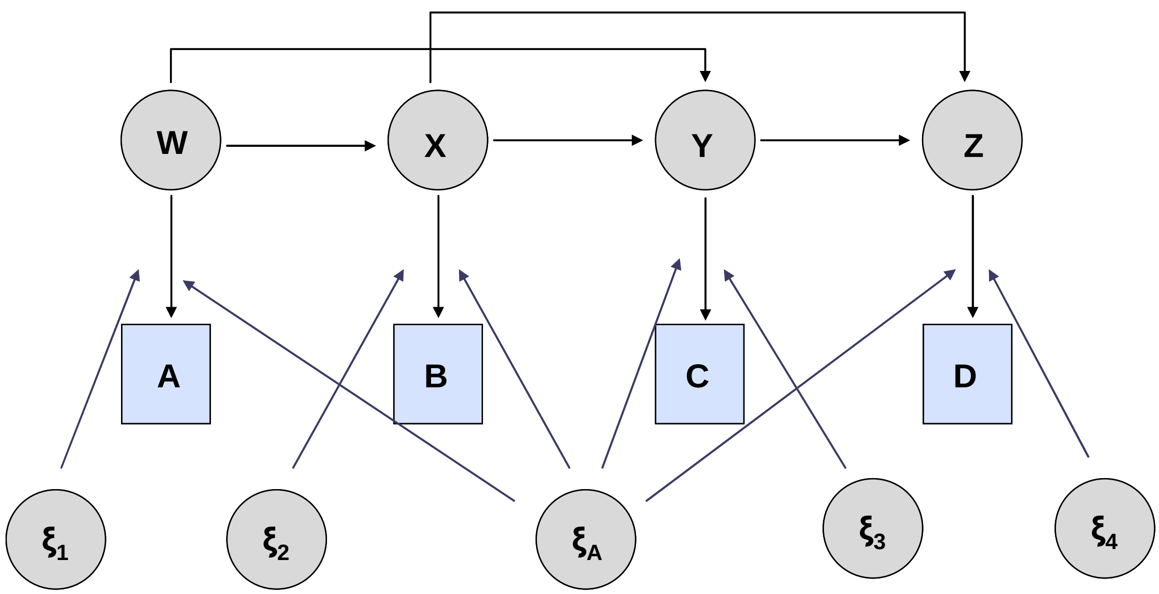

Figure 1 Combined Measurement Error Model

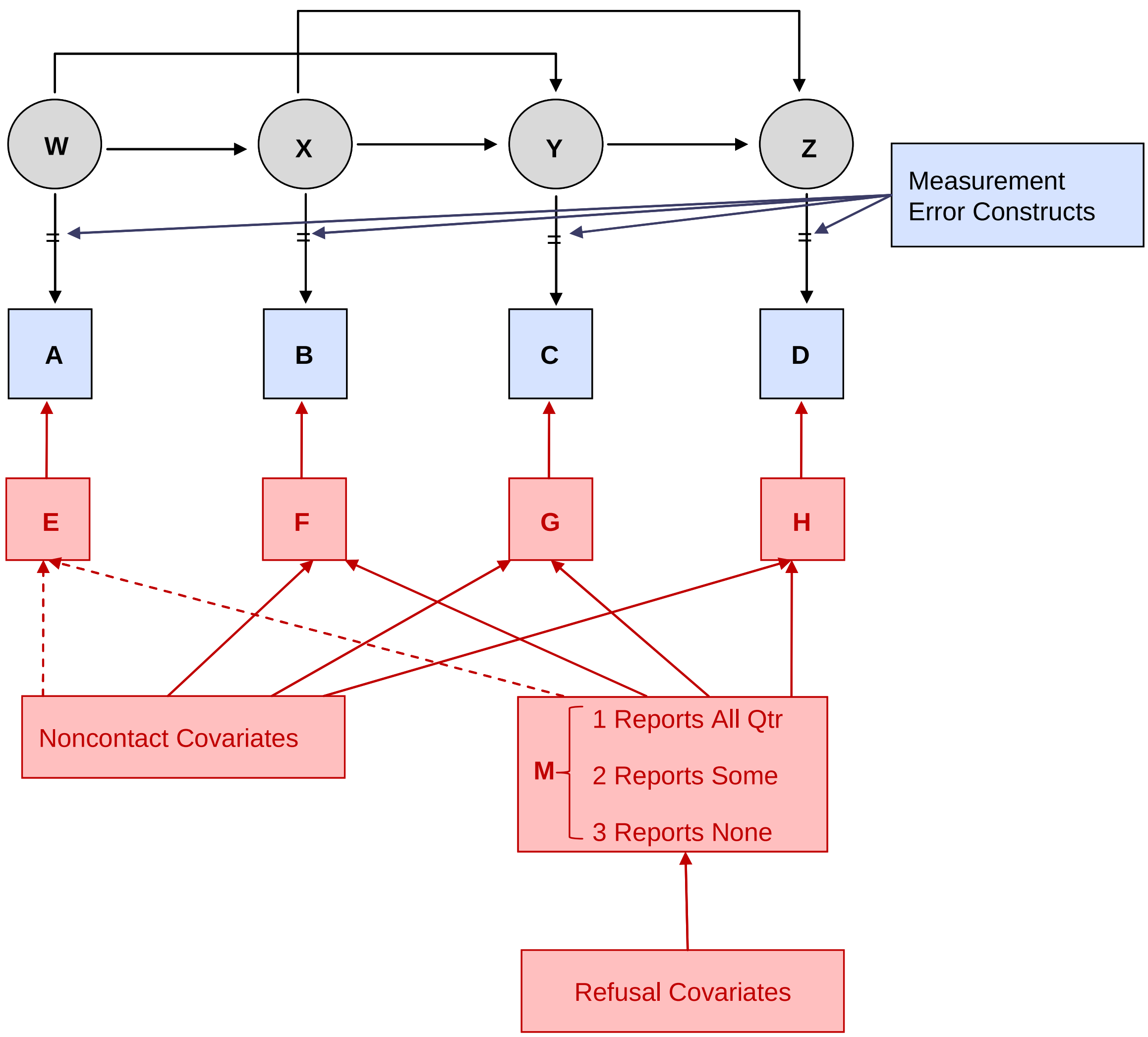

Figure 2 Combined Measurment Error and Noncontact Model