Good Item or Bad – Can Latent Class Analysis Tell? Examining the Effectiveness of a Latent Class Analysis Approach to Item Evaluation

Frauke Kreuter[1], Ting Yan[2], Roger Tourangeau[1,2]
Joint Program in Survey Methodology, University of Maryland[1]
Survey Research Center, University of Michigan[2]

Abstract

Latent class analysis (LCA) has been used to model measurement error, to identify flawed survey questions, and to compare mode effects. Using data from a survey of University of Maryland alumni together with alumni records we evaluate this technique to determine its accuracy and effectiveness for detecting bad questions in the survey context. Our results showed good qualitative results for the latent class models – the items that the model deemed the worst were the worst according to the true scores – but weaker quantitative estimates of the error rates for a given item.

Keywords: Latent Class Analysis, Questionnaire Design, Item Development

1 Introduction

The development, testing, and evaluation of questions in surveys remains a qualitative endeavor that relies primarily on such tools as expert reviews of the questions, focus groups, and cognitive interviews (Presser et al., 2004). Many have questioned the effectiveness of these methods in identifying problem items (Conrad and Blair, 2004). The tests that any survey item must ultimately pass involve quantitative standards such as validity, reliability, low error variance, and, for factual items at least, freedom from bias - all of them are well-established in the psychometric and survey literatures. Unfortunately, the tools used most often for developing and testing survey questions yield information that is at best indirectly related to these quantitative criteria.

There are a couple of reasons for this disconnect between the qualitative data that are actually collected during the development of the typical questionnaire and the quantitative standards that are the ultimate goal. One common design for measuring the validity and bias in responses to survey questions is to compare the answers to some external measure, such as an administrative record, for the item of interest (annual earnings, recent doctor visits, children's immunizations). If the external measure is regarded as essentially error-free (that is, if it represents a "gold standard" for the variable), the comparison between survey reports and external measures yields direct estimates of the validity and bias in the survey reports.

In practice, there are problems with this common design. The most prominent ones are listed here: Collecting such external data can be very costly; no gold standard may exist at all for a variable; the gold standard may exist but only for some specialized and possibly unrepresentative subpopulation (such as the members of a single health plan); the data may exist in principle but permission to access them may be needed from the survey respondents, from whoever maintains the external data, or from both, producing high levels of nonresponse and missing data; the data may exist and be accessible but they may themselves be full of problems, such as missing or incorrect data; finally even with accurate record data, there may be problems in matching the records and the reports, biasing the estimates of the error in the survey reports. Thus, there is a real need for methods for assessing the measurement characteristics of survey items without collecting external validation data.

In recent years Latent Class Analysis (LCA) has become an attractive choice for assessing error in the absence of a gold standard because it doesn't rely on error-free measures. Biemer and colleagues have demonstrated how to apply LCA to identify flawed survey questions and to uncover the root causes of their problems (Biemer and Wiesen, 2002; Biemer, 2004b). For example, Biemer and Wiesen (2002) examined three indicators used to classify respondents regarding their marijuana use; the data were from the 1994, 1995, and 1996 National Household Survey on Drug Abuse (NHSDA). In another application of LCA to estimate error in survey items, Biemer (2004a) examined the labor force question in the 1996 Current Population Survey.

In LCA several measures are needed that can be treated as fallible indicators of the unseen states of being unemployment, using drugs etc. In a standard application of LCA three binary indicator variables are needed for a model with two latent classes to be just identified (McCutcheon, 1987). In a survey setting, it can be difficult to obtain three measurements of the same variable in a single questionnaire or a re-interview. However, instead of using three-indicator variables, one can use a covariate predicting latent class membership to inform the model and therefore to obtain additional degrees of freedom. One way to deal with this problem is adding a grouping variable to establish identifiability (Hui and Walter, 1980; Clogg and Goodman, 1984; Biemer and Witt, 1996). Biemer (2004b) demonstrated the application of LCA models where a grouping variable is used together with re-interview data that are collected routinely in several federal surveys. The application of LCA with only two observed indicators and a covariate relies on an assumption about the prevalence of class membership and error rates across the covariate groups (see section 4.2 for details).

Given LCA's reliance on assumptions to achieve identifiable models with only two indicators, one potential danger with LCA is that these assumptions will be invalid. Analysis under invalid assumptions can produce biased results and leads to unsound conclusions. Thus, LCA models will never be widely accepted within the survey research community unless researchers are confident that the LCA models give results that are consistent with more traditional procedures for testing survey items. For example, when two survey questions are compared to a gold standard measure, the LCA estimates of the false positive and false negative rates should be consistent with direct estimates based on a comparison between the survey responses and the gold standard. Similarly, when we administer two versions of a question but deliberately implant problems in one version, the LCA results should pick out the inferior version of the item.

Our paper aims to assess more systematically the potential of latent class models for use in developing and testing survey questions. We will apply LCA to evaluate three survey items asking respondents about their past academic difficulties; one of the three items was deliberately designed to be a flawed question. The true values for these items are based on the respondents' academic transcripts. With these data, we seek to answer some specific questions about the application of LCA models as a tool for evaluating survey questions:

1. Do latent class methods yield results that agree with accepted procedures for assessing questionnaire items? For example, when a gold standard is available, how well do the estimates from LCA models agree with those from more traditional analyses of item validity?

2. When the assumptions for the prevalence and error rates in covariate groups are not met, will the results of a two-indicator LCA still be valid? Under what circumstances are LCA results robust over violations of these assumptions?

2 Latent Class Analysis in the Context of Question Evaluation

The standard latent class analysis measures one or more unobserved (latent) categorical variables through a set of observed indicator variables. The basic idea of LCA is that the associations between the observed indicators arise because the population is composed of latent classes, and the distribution of the indicators vary across classes. Within each of those mutually exclusive and exhaustive groups (latent classes) the observed variables are unrelated. This 'local independence' (Lazarsfeld and Henry, 1968) assumption allows inferences about the latent class variable.

For a set of binary indicator variables, the relationship between the observed and the unobserved variables can be described with logistic regression equations. In such models, the probability for each observed item u (from a set of J observed items) is the product of the probability of being in class k of the latent class variable c and the probability of the observed response ($u_j = 1$), given class membership, summed across all of the latent classes:

$$P(u_j = 1) = \sum_{k=1}^{K} P(c = k) P(u_j = 1 | c = k)$$

Following the notation of Muthén (2001), the joint probability of all J observed variables (u_1, u_2, etc.) under the assumption of conditional independence is

$$P(U_i = u_i) = \sum_{k=1}^{K} \gamma_k \prod_{i=1}^{I} \rho_{i|k}.$$

LCA produces unconditional probabilities $\gamma_k = P(c = k)$, which represents the probability that respondents are assigned to each class of the latent variable. In a way, the unconditional probabilities estimate the prevalence of each class in the population (or the size of each latent class). In addition, LCA also allows one to obtain various probabilities conditional on class membership. For example in a two class model the probability of endorsing a binary item u_1 conditioned on being in class one will be estimated as $\rho_{1|1} = P(u_1 = 1|c = 1)$ and the probability of not endorsing this particular item as $\rho_{2|1} = P(u_1 = 2|c = 1)$; similarly for class two, $\rho_{1|2} = P(u_1 = 1|c = 2)$ and $\rho_{2|2} = P(u_1 = 2|c = 2)$. Two of the conditional probabilities represent the extent of misclassifications produced by the survey questions; they are the probability of a false positive response and the probability of a false negative response for a question item given membership in the latent class. These are sometimes referred to collectively as the "error probabilities". A high false positive probability or a high false negative probability usually signals that there is a problem with a particular survey question. Thus, LCA allows comparisons of question sensitivity and specificity to identify questions that best differentiate the classes. Here the primary purpose of applying LCA to questionnaire pretesting is to identify flawed questions that elicit unreliable or biased report; such problem questions are identified via the estimated false positive and false negative probabilities.

In the application of Biemer and Wiesen (2002) three indicators were used to classify respondents regarding their marijuana use. One indicator was the response to a question that asked about the length of time since the respondent last used marijuana or hashish (the recency question). The second indicator was the response to a question that asked how frequently the respondent has used marijuana or hashish in the past year (the frequency question). The final indicator was a composite of several questions on the drug answer sheet. An affirmative answer to any of the questions from which the composite is derived was coded as a 'yes' on the final indicator. Biemer and Wiesen (2002) report that the final indicator was prone to be inconsistent with the other two measurements of the use of marijuana in the past year in the 1994. However, it is not clear whether the problem was due to false positive errors in that indicator or false negative errors in responses to the recency and frequency items. The LCA estimates of the false positive and false negative error rates for the three indicators unequivocally identified the problem as false positives in the answer sheet composite (Biemer and Wiesen, 2002). In addition, LCA results showed a larger false negative rate for the recency question than for the set question. Combined with a more traditional analysis, Biemer and Wiesen (2002) concluded this was because infrequent users responded falsely to the recency question but answered honestly to the frequency question.

2.1 Hui-Walter model assumptions

Mathematically speaking, three binary indicator variables are needed for a model of two latent classes to be just identified. However, in survey context, three measurements aren't easy to come by. When there are only two indicators, one can impose assumptions on the parameters ρ to achieve identifiability. For instance, one possible assumption restricts the false negative probability to zero or sets the latent class sizes to be equal across subgroups[1]. However, sometimes these assumptions are either too stringent or theoretically implausible. Another way to free up additional degrees of freedom is to use a covariate $g_i = 1, 2, \ldots, G$ predicting latent class membership to inform the model. A grouping variable can be added to establish identifiability if restrictions on γ or ρ are imposed. Biemer and Witt (1996) refer to this as the *Hui-Walter* model (Hui and Walter, 1980). The grouping variable has to satisfy two assumptions.

1. The prevalence rates have to be different across the levels of the grouping variable (the different prevalence assumption).
2. The false positive and false negative probabilities have to be equal across levels of the grouping variable (the equal error probabilities assumption).

One would, for example, assume different prevalence of each labor force category (employed, unemployed) for males and females $P(c = 1|g = 1) \neq P(c = 1|g = 2)$, but assume that males and females have the same probabilities of endorsing the item given their true state (class membership) $\rho_{1|1,g=1} = \rho_{1|1,g=2}$.

Biemer (2004a) used the Hui-Walter model to estimate errors in in the 1996 Current Population Survey (CPS) labor force questions. Using data from the original survey and a reinterview, Biemer fit a LCA model without constraining the error probabilities for the interview and reinterview to be identical, as is usually done in using reinterview data. The estimated misclassification probabilities showed a high misclassification rate for the unemployed status; according to the results, about one third of unemployed persons were misclassified in the CPS and 80% of the misclassified cases were incorrectly classified as "not in labor force" (Biemer, 2004a). This finding was consistent with both historical data on the reliability of the CPS data and theoretical expectations suggesting that the concept of unemployment is difficult for many respondents (Biemer, 2004a,b).

3 Data

The data used for our study come from the Joint Program in Survey Methodology (JPSM) Practicum class. Each year JPSM carries out a survey designed largely by the students. However the field work for these surveys is done by a professional survey organization. In 2005, the study was designed to examine the effects of data collection mode on reports of sensitive questions. The study incorporated a record check of some of the survey items. To study the robustness of LCA when using the Hui-Walter model, we embedded multiple measures in the survey of an item that could be checked with the record data.

3.1 Description of the data

The Alumni Survey has several important features that we can exploit to test the effectiveness of the LCA in identifying flawed survey questions.

First, a number of the items on the survey questionnaire can be verified against university records. Because the University of Maryland Registrar's records were used to select the sample and because the item wording was designed to fit the information on the student transcripts, we are able to check responses to survey questions against data at the Registrar's office with minimal risk of matching errors. The availability of both record data and self-reports allow us to compare the LCA result to the more traditional 'gold-standard' analysis (assuming the record data to be error-free).

Second, the Alumni Survey included an experiment in which respondents were randomly assigned to different modes of data collection. Approximately one third of the respondents completed the interview via computer-assisted telephone interviewing (CATI). Another third completed it via interactive voice response (IVR), in which the computer played a recording of the questions over the telephone and the respondents indicated their answers by pressing the appropriate numbers on the telephone keypad. A third group answered the questions via the Internet. All three subgroups were initially contacted by telephone and administered a brief set of screening questions to verify that the interviewer had reached the correct person. Cases were then randomly assigned to a

[1] For examples of various restrictive or equality assumptions, see McCutcheon (1987)

Table 1: The Response Rate and the Number of Completes

Total	Total	Percent
Alumni eligible (number dialed)	7,567	100.0
Screener completion	1,501	19.8
Initially assigned	1,501	100.0
Started main questionnaire	1,094	72.9
Number of completes	1,003	66.8
Response Rate (AAPOR RR1)	66.8*19.8	13.3

mode of data collection[2].

3.2 Survey design and data collection

The survey sample was drawn from the 55,320 individuals who received undergraduate degrees from the University of Maryland from 1989 to 2002, as reflected in the records of the Office of the Registrar. The survey interviewing was done by Schulman, Ronca, and Bucuvalas, Inc., in August and early September 2005. The Registrar's records were used to select a random sample of 20,000 graduates, stratified by graduation year. Of these 20,000 graduates, 10,325 could be matched to Alumni Association records containing telephone contact information. After we dropped various ineligible numbers, including those used in the pretest, 7,957 phone numbers were fielded for the survey. Call attempts were made to 7,591 numbers. Twenty-four alumni were deceased so the denominator for the response rate calculation was 7,567. The alumni were initially contacted by telephone and administered a brief set of screening questions about the respondent's personal and household characteristics, access to the Internet, and affiliation with UMD. A total of 1,501 alumni completed the screener and were randomly assigned to a mode of data collection; for 37 individuals without access to the web, the random assignment was restricted to CATI versus IVR . A total of 1,094 alumni started the main questionnaire and 1,003 completed interviews were obtained. Taking into account the completion of the screener and the completion of the main questionnaire, the AAPOR response rate 1 was 13.2% (see Table 1).

Approximately a third of the respondents (n=320) completed the interview via computer-assisted telephone interviewing (CATI). Another third (n=320) completed via interactive voice response (IVR). The final third (n=363) answered the questions via the Internet. We restrict our analysis in this paper to the complete cases in each mode.

3.3 Questionnaire and record data

The main questionnaire included 37 questions related to educational topics, current relationships to the university, community involvement, and a final set of questions to capture perception of the sensitivity of some key questions asked during the interview. We focus here on the subset of items for which we obtained record data from the Registrar's office or the Alumni Association. The questionnaire included three items designed to tap essentially the same information (Table 2). We deliberately designed the second of the three questions (Q18a) to be a vaguer version of the other two, hoping it would yield higher overall error rates. Responses to the third item were recoded according to whether the respondent reported a D or an F as his or her worst grade.

Table 2: The Three Items Included in the Survey

12.	Did you ever receive a grade of 'D' or 'F' for a class?
18a.	Did you ever receive an unsatisfactory or failing grade?
18b.	What was the worst grade you ever received in a course as an undergraduate at the University of Maryland?

4 Analyses and Results

Our analysis first compares the LCA estimates of the false positive and false negative probabilities for the three items with the estimates obtained by direct comparison with the transcript data, to see whether latent class methods yield plausible results that agree with accepted procedures for assessing questionnaire items. Then, to evaluate the sensitivity of LCA methods that rely on the Hui-Walter model, we performed a set of analyses that compare estimates from four different two-item LCA models that use different covariates to identify the model. We selected four grouping variables as covariates in such a way that one satisfied both assumptions about prevalence and error rates, two violated one assumption but not the other, and one violated both assumptions. The latent class analyses were done with Mplus version 4.1[3]. The analyses are unweighted.

4.1 Agreement with accepted procedures

As a first step, we discuss the misclassification rates for each of the survey items using a 'gold standard' – here record data from academic transcripts. Subsequently, we will estimate error rates for each of the survey items using LCA. Those estimated error rates will then be compared to the error rates resulting from the 'gold standard' analysis.

4.1.1 *Misclassification rates from direct analysis with academic transcripts*

Of the 954 cases used in this analysis 60% had received a "D" or "F" at some point during their undergraduate career. Not all of these respondents reported receiving such a grade in the survey. Table 3 presents the misclassification rates for the three items, obtained from the

[2] For those without access to the web, the random assignment was restricted to CATI or IVR. More design details can be found in Kreuter et al. (2006).

[3] www.statmodel.com

Table 3: Traditional Gold-Standard Analysis (Error rates in %)

Question ID	False Negative	False Positive
Q12.	26.2	2.9
Q18a.	59.0	1.8
Q18b.	25.0	2.9

comparison of survey data and academic transcripts. Of all the respondents who had a received at least one “D” or “F” according to the Registrar’s data, 26% failed to indicated this in respondse to this to question Q12 (see Table 2). Those cases are referred to as false negatives.

False positives are cases reporting a “D” or “F” but who didn’t actually receive one. To continue the example with question Q12, of all respondents who did not receive a “D” or “F” according to the record data roughly 3% did report having got one. so. It is noticeable that all three questions have much lower false positive rates than false negative rates, evidence that the reports are distorted in a socially desirable direction.

The magnitude of false positive rates are similar for the three items, but Q18a has the highest false negative rate among the three, suggesting that this question is flawed and most prone to socially desirable responding. Question Q18a was deliberately designed to be flawed.

4.1.2 LCA estimates of misclassification

We ran a three-indicator latent class model with two latent classes. Using random starting values, the best log-likelihood for this model was -1172.0 with 7 free parameters and a BIC value of 2392.1. A two-class model outperforms a one-class model, which had a much lower log-likelihood of -1856.2 and a BIC value of 3733.0. Both the Lo-Mendel-Rubin Likelihood Ratio Test (Lo et al., 2001), as well as the Bootstrapped Parametric Likelihood Ratio Test (McLachlan and Peel, 2000; Nylund et al., 2006) show a significantly better fit of a two-class model compared to a one-class model.

The size of the latent classes based on the estimated model are 55% for those who received a grade of “D” or “F” in college (we labeled it “Failed”) and 45% for those who did not receive a grade of “D” or “F” (the label of this class is “Did Not Fail”). Part of the model results are estimated probabilities of endorsing each of the items given class membership. The probability of a false positive response (falsely reporting “D” or “F”) and a false negative response (not reporting “D” or “F”) are estimated for each of the items. Table 4 shows low false positive and false negative rates for all three indicators with the exception of Q18a, where the false negative rate is 45%.

Table 4: LCA Analysis (Error rates in %)

Question ID	False Negative	False Positive
Q12.	2.1	2.5
Q18a.	45.1	1.2
Q18b.	1.2	3.2

Table 5: LCA and True Score Comparison

	Record Data (Trues Score)		
	Fail	¬ Fail	Total
LCA: Fail	43.4%	1.1%	44.4%
LCA: ¬ Fail	16.7%	38.9%	55.6%
Total	60.1%	39.9%	100%

4.1.3 Comparing LCA Results to Gold-Standard Analysis

Figure 1 allows a direct comparison of the LCA results with the gold-standard analysis. It indicates that the LCA estimates of the false positive probabilities are quite similar for all three items and also similar to those obtained from the direct comparison to academic transcripts. The LCA estimates of false negative probabilities exhibits the same pattern as those from the direct analysis; Q18a shows the highest probability of false negative responses, singling it out as a problematic item. Thus, the latent class approach successfully identified Q18a as the bad item, a conclusion consistent with our original intention and with results from the direct analysis with record data.

However, even though the latent class estimates of the misclassification rates leads to the same qualitative conclusion (Q18a is a flawed item), the LCA estimates of false negative probabilities are consistently smaller than those from the gold-standard analysis. In addition, the LCA estimates failed to reveal the large quantitative differences between the false positive and false negative rates for the other two items.

The latent class model also assigns individuals to latent classes based on the modal class-membership probabilities to produce unconditional class probabilities (or prevalence rates of each class). We compared the LCA assignment of respondents to classes to the true values and display the results in Table 5. LCA correctly classified most of those who didn’t actually receive a grade of “D” or “F”, but wrongly classified almost one third of those who failed a course in college to the “Did Not Fail” class. Table 5 shows that the LCA estimates might not be effective at estimating the prevalence of students who failed at least one course in college.

4.2 Examining Underlying Assumptions

When there are only two indicators, a two-class model is under-identified. Hui-Walter assumptions are sometimes imposed on the parameters to achieve identifiability. In order to test the robustness of the Hui-Walter

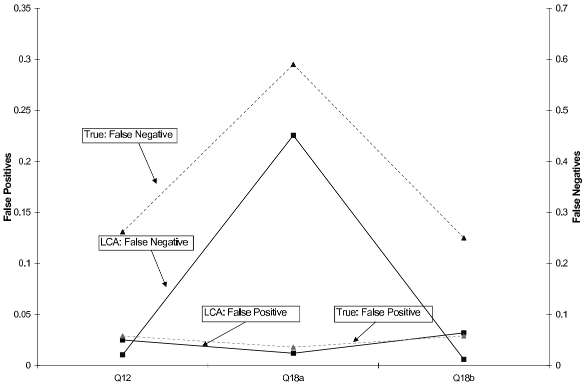

Figure 1: False Positive and False Negative Rates for LCA and True Score Comparison

assumptions, we experimented with four items as potential grouping variables, the gender of the respondents, respondents' GPA, a random split of the sample, and mode of data collection. For all four grouping variables, we can use the academic transcript data whether the different categories of the grouping variable differ in their prevalence rates (*Assumption 1*) and whether their error rates are the same (*Assumption 2*). The results are presented briefly in this summary. A detailed description can be found in the version of our paper available upon request.

1. *Both assumptions are satisfied*: Among the four grouping variables, gender satisfied both of the Hui-Walter assumptions. The proportion of alumni failing a class at college is higher for males (68%) than for females (52%), a statistically significant difference (χ^2=27.27, $p < .0001$). In addition, there is no reason to expect that males are more (or less) likely than females to under-report (or overreport) whether they have ever failed a class. Using the transcript data, we see that the false positive rates do not differ by gender of the respondents for all three items. The false negative rates of Q12 and Q18b do not differ by gender either. For Q18a, however, the false negative rate is significantly higher for females (0.65) than for males (0.54) (χ^2=6.7, p=.01).

2. *Satisfying assumption 1, while violating 2*: Respondents' GPA, by contrast, satisfies the different prevalence assumption; about 91% of respondents whose GPA is equal to or lower than the median GPA failed a class in college, a proportion significantly higher than the 30% for those whose GPA is higher than the median (χ^2=372.24, $p < .0001$). But grouping by respondents' GPA, violates the equal error probability assumption: the false positive rates do not differ by GPA for all three items, but the false negative rates do differ significantly by GPA group. Students with a higher GPA tend to have a lower false negative rate than those with a lower GPA; the differences are statistically significant for all three items (For Q12, χ^2=9.03, $p < .01$; For Q18a, χ^2=13.54, $p < .001$; For Q18b, χ^2=7.75, $p < .01$).

3. *Violating assumption 1, while satisfying 2*: We use two random half-samples of the respondents as a third grouping variable. The random assignment ensures that the proportion failing an undergraduate course is equal across the two groups in expectation and that the error probabilities are equal as well. Statistical tests show that the prevalence rates are not significantly different between the two random halves generated for this analysis (58% vs. 62%, χ^2=1.6, p=.21). Error rates do not differ by the random-half samples either. Therefore, the random split satisfies one assumption (the equal error probabilities assumption) but violates the other (the different prevalence assumption).

4. *Violating both assumptions*: Finally the mode of data collection as grouping variable violates both assumptions. The respondent groups under each mode show equal prevalence rates and different error rates. Of those respondents who were assigned to the CATI mode 61.8% had a "D" or "F" according to the registrar's data, the same is true for 62.3% of the Web respondents and 58.8% of those respondents randomly assigned to IVR. From previous mode analyses we know that the mode of administration has an impact on the error probabilities; thus, the mode of administration violates both assumptions of the Hui-Walter method.

4.2.1 Unequal Prevalence and Equal Error Rates

We first used gender of the survey respondents as a grouping variable, applying the Hui-Walter assumptions. We found that Q18a consistently produced a higher false negative rate across gender groups than the other two survey items. Regardless of which two variables were entered into the latent class models, the LCA estimates of false negative probabilities follow the same pattern as those from direct analysis. However, the LCA estimates are again consistently smaller than the true estimates. The differences between the estimated false positive probabilities are greater between the two approaches. Furthermore, the quantitative differences between false positive and false negative probabilities in the LCA results are also different from those obtained from direct analysis.

4.2.2 Unequal Prevalence and Unequal Error Rates

By contrast, GPA fulfills the different-prevalence-rate assumption of the Hui-Walters, but violates the equal-error-rate assumption. Respondents having a GPA equal or higher than the median GPA are classified into the high GPA group while those with a GPA lower than the median is grouped into the low GPA group.

The estimates of error probabilities from the LCA model show that Q18a has larger false negative probabilities than the other two items across the two GPA groups, a conclusion supported by both direct analysis of academic transcripts and the LCA models. Again, similar to the case when gender is used as a grouping variable, the LCA estimates of false negative probabilities are consistently smaller than the true false negative probabilities, but they follow the same trend. The LCA estimates of false positive probabilities are inconsistent with the true positive probabilities.

Models with GPA as a grouping variables seemed to be able to pick out the flawed question item, despite that the true false negative rates are higher for respondents with a higher GPA than for those with a lower GPA. Even though GPA and the random split both violated one of the two Hui-Walter assumptions, the LCA estimates resulting from them as the grouping variable vary substantially. It seems that the violation of the equal error rate assumption is less serious than the violation of the different prevalence assumption. Future research should do a simulation study to investigate the severity of violating the two assumptions.

4.2.3 Equal Prevalence and Equal Error Rates

We next split the sample randomly into two equal halves. The random split was included in the LCA models as the grouping variable. The LCA estimates of false negative probabilities didn't show Q18a as the flawed question suffering from the largest false negative probabilities. Instead, Q18a is shown to be a better indicator than the other two with lower error rates. In addition the LCA estimates of false positive and false negative probabilities are also off for the other two questions. Recall that, under the Hui-Walter method, the grouping variable is required to have different prevalence rates across levels of the grouping variable and the random split failed this assumption. The biased LCA estimates could be a result of this violation.

4.2.4 Equal Prevalence and Unequal Error Rates

Lastly, we used mode as a grouping variable and ran the same 2-indicator models under the Hui-Walter assumptions. The LCA estimates from the 2-indicator models with the mode of data collection as the grouping variable are off as well. The LCA estimates lead to different conclusions from the analysis of academic transcripts: the LCA estimates indicate that Q18a was a better performer than the other two items, with lower false negative and false positive rates. The estimated false negative probabilities from the LCA models are consistently lower than the true false negative rates. This is contradictory to the estimates from the direct analysis of the transcript data. Given that the mode variable also violates the different-prevalence-rate assumption (as the random split grouping variable), it seems that the violation of this key assumption could lead to seriously misleading conclusions.

5 Discussion

This study examined the effectiveness and robustness of the latent class analysis approach in evaluating survey questions. In the absence of true scores or gold standards, the LCA can be employed to assess measurement properties of survey items. We fitted a three-indicator latent model and various two-indicator latent models. Our results showed that the 3-indicator LCA is able to successfully identify Q18a as a flawed item having the highest false negative probabilities, thus reaching the same qualitative conclusion as when directly compare the reports to the to the academic transcripts. Thus, in the absence of gold standards, it seems that the LCA can produce qualitative results consistent with those from the more traditional analysis with true scores. However, the quantitative estimates of the error probabilities from the LCA differ from those from the direct analysis.

We also examined the robustness of the LCA when there are not enough indicators to identify the models. We applied the Hui-Walter assumptions to achieve identifiability. These assumptions require prevalence rates to be different but constrain the error probabilities to be equal across the levels of a grouping variable. We examined the validity of the LCA results when Hui-Walter assumptions are both satisfied, when only one of the assumptions is satisfied, and when both are clearly violated. Our analysis showed that the LCA can produce quite reasonable results that are consistent with the direct approach when both or only the equal-error-rate assumption is satisfied. For instance, when gender and the GPA of the respondents are used as a grouping variable, Q18a is shown to have the largest false negative probabilities. However, when the different-prevalence-rate assumption or both assumptions are violated, the LCA results are quite misleading; the LCA did not identify Q18a as the item producing the largest false negative probabilities when the sample was randomly split or when the mode of data collection was used as the grouping variable. Therefore, our results suggest that the LCA models can tolerate violations of their assumptions only to a limited extent.

In addition, we find that the quantitative estimates from LCA do not always agree with results from direct analysis even when the qualitative conclusions of the LCA are in agreement. The discrepancy in quantitative estimates was largest when the different-prevalence-rate assumption is violated. Our next step is to quantify the fit of LCA results to the results of direct analysis. We plan to carry out simulation studies to examine the relation between the extent of violations and the validity of LCA results.

5.1 Acknowledgement

This study is based upon work supported by the National Science Foundation under Grant No.0550002. We like to thank the Methodology, Measurement, and Statistics

(MMS) Program and Dr. Cheryl Eavey for their support. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors contributed equally to this paper. A first set of the results was presented by Ting Yan at the WAPOR/AAPOR meeting in Montreal in 2006. We thank Paul Biemer, Elisabeth Coutts, Jill Dever, Michael Lemay, and Stanley Presser for critical comments and helpful suggestions.

References

Biemer, P. and C. Wiesen (2002). Measurement error evaluation of self-reported drug use: A latent class analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society, Series A 165*, 97–119.

Biemer, P. P. (2004a). An analysis of classification error for the revised Current Population Survey employment questions. *Survey Methodology 30*, 127–140.

Biemer, P. P. (2004b). Modeling measurement error to identify flawed questions. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*, pp. 225–246. New York: Wiley.

Biemer, P. P. and M. Witt (1996). Estimation of measurement bias in self-reports of drug use with applications to the national household survey on drug abuse. *Journal of Official Statistics 12*, 275–300.

Clogg, C. C. and L. A. Goodman (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association 79*, 762–771.

Conrad, F. and J. Blair (2004). Data quality in cognitive interviews: The case of verbal reports. In S. e. a. Presser (Ed.), *Methods for Testing and Evaluating Survey Questionnaires*, pp. 67–87. New York: Wiley.

Hui, S. L. and S. D. Walter (1980). Estimating the error rates of diagnostic tests. *Biometrics 36*, 167–171.

Kreuter, F., S. Presser, and R. Tourangeau (2006). Comparing social desirability bias in cati, ivr, and web surveys. *Paper presented at the Annual Conference of the American Association for Public Opinion Research.*

Lazarsfeld, P. F. and N. W. Henry (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lo, Y., N. Mendel, and D. Rubin (2001). Testing the number of components in a normal mixture. *Biometrika 88*, 767–778.

McCutcheon, A. L. (1987). *Latent class analysis.* Beverly Hills: Sage Publications.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models.* New York: Wiley.

Muthén, B. (2001). Second-generation structural equation modelling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modelling. In A. Sayer and L. Collins (Eds.), *New methods for the analysis of change*, pp. 291–322. Washington, DC: American Psychological Association.

Nylund, K., T. Asparouhov, and B. Muthén (2006). Deciding on the number of classes in latent class analysis and growth mixture modeling. a monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal Accepted for publication.*

Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (2004). *Methods for Testing and Evaluating Survey Questionnaires.* New York: Wiley.