Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2008)
171, *Part* 3, *pp.* 723–738

Good item or bad—can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions

Frauke Kreuter,

University of Maryland, College Park, USA

Ting Yan

National Opinion Research Center, Chicago, USA

and Roger Tourangeau

University of Michigan, Ann Arbor, and University of Maryland, College Park, USA

[Received November 2006. Final revision October 2007]

Summary. Latent class analysis has been used to model measurement error, to identify flawed survey questions and to estimate mode effects. Using data from a survey of University of Maryland alumni together with alumni records, we evaluate this technique to determine its usefulness for detecting bad questions in the survey context. Two sets of latent class analysis models are applied in this evaluation: latent class models with three indicators and latent class models with two indicators under different assumptions about prevalence and error rates. Our results indicated that the latent class analysis approach produced good qualitative results for the latent class models—the item that the model deemed the worst was the worst according to the true scores. However, the approach yielded weaker quantitative estimates of the error rates for a given item.

Keywords: Item development; Latent class analysis; Questionnaire design

1. Introduction

The development, testing and evaluation of questions in surveys remain a qualitative endeavour 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), for example). The tests that any survey item must ultimately pass involve quantitative standards such as low error variance and freedom from bias. Measures for these criteria are well established in the psychometric and survey literature. Unfortunately, the tools that are used most often for developing and testing survey questions yield information that is at best indirectly related to these quantitative standards.

There are a couple of reasons for this disconnection between the qualitative data that are actually collected during the questionnaire development and the quantitative standards that are the goals in theory. Ideally, we would measure validity and bias in responses to survey questions by comparing the answers with some external measure. If the external measure is regarded

Address for correspondence: Frauke Kreuter, Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742, USA.
E-mail: fkreuter@survey.umd.edu

 0964–1998/08/171723

as essentially error free (i.e. if it represents a 'gold standard' for the variable), the comparison between survey reports and external measures yields direct estimates of the error in the survey reports. Designs of this kind can sometimes be found for items of interest for which accurate administrative record data exist, e.g. annual earnings, recent visits to the doctor and children's immunizations.

In practice, there are problems with this design. The most prominent are as follows: no gold standard may exist for a variable; the gold standard may exist but collecting such external data can be very costly; external data may exist only for some specialized and possibly unrepresentative subpopulation (such as the members of a single health plan); external 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 non-response and missing data; external data may exist and be accessible but they may themselves be full of problems, such as missing or incorrect data; finally, even with accurate external data, there can be problems in matching the external records and the survey reports, biasing the estimates of the error in the latter. 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 because it does not rely on the availability of error-free measures. Instead of using error-free external data that are often not available, LCA makes use of multiple measures of the same construct or characteristics to estimate error. In addition, the multiple measures can be fallible or infallible, as long as the errors that are associated with each measure are independent conditional on the latent variable.

For a latent class model with two latent classes, three binary indicator variables are needed for the model 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 through a reinterview. One way to deal with this problem is to add a grouping variable to establish identifiability (Hui and Walter, 1980; Clogg and Goodman, 1984; Biemer and Witt, 1996). The selection of a grouping variable must satisfy certain assumptions regarding the prevalence of class membership and the error rates across the covariate groups (see Section 4.2 for details). Thus LCA relies on substantive knowledge of prevalence and error rates for subgroups of the population.

Our paper aims to assess systematically the potential of using latent class models in developing and testing survey questions. Two sets of LCA models are examined in this evaluation. Both sets of models assume two latent classes. One set examines three indicators. The other set uses two indicators and makes varying assumptions about prevalence and error rates to achieve identifiability. The first set of models directly 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. We seek to answer some specific questions about the applicability of LCA models as a tool for evaluating survey questions.

(a) Do latent class methods yield results that agree with accepted procedures for assessing questionnaire items? Specifically, we want to examine whether the LCA approach produces results that are comparable with those of analysis using external data.
(b) Can LCA identify the question with poorer performance when there are multiple survey questions, but no gold standard measure?
(c) Can LCA be used to help survey researchers to sort out the relatively better measures from the relatively poorer measures in the absence of a gold standard measure?

In the second set of models, we apply LCA with different pairs of indicators and varying

grouping variables. The grouping variables vary in actual prevalence and error rates. With the help of these analyses, we discuss the following questions.

(d) 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?
(e) Under what circumstances are LCA results robust over violations of these assumptions?

2. Latent class analysis in the context of question evaluation

We first briefly review LCA in Section 2.1 and discuss the use of two-indicator LCA in the context of question evaluation in Section 2.2 before we describe data and analyses for the current study.

2.1. Latent class analysis

The standard LCA 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 varies across classes. Within each of those mutually exclusive and exhaustive groups (latent classes) the observed variables are unrelated. (It is this 'local independence' (Lazarsfeld and Henry, 1968) assumption that allows inferences about the latent class variable.)

In such models, the probability for each item u_j (from a set of J observed items) is the product of the conditional probability of u_j, given membership in class k, summed over the latent classes. Since the responses to the items are conditionally independent given latent class membership, the marginal probability for the vector of responses $\mathbf{u}$ is given by

$$\mathbf{u} = \sum_{k=1}^{K} P(c=k) \prod_{j=1}^{J} P(u_j|c=k).$$

LCA produces unconditional probabilities $P(c=k)$, which are the probabilities that respondents belong to each class of the latent variable (Table 1). The unconditional probabilities estimate the prevalence of each class in the population (or the size of each latent class). In addition, LCA also allows us 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 1 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 2, $\rho_{1|2} = P(u_1 = 1|c=2)$ and $\rho_{2|2} = P(u_1 = 2|c=2)$.

Two of the conditional probabilities can be thought of as representing the probability of a false positive response ($P(u_1 = 1|c=2)$) and the probability of a false negative response

Table 1. Class probabilities

Item	*Probability with c = 1*	*Probability with c = 2*	*Unconditional probability*
$u_1 = 1$ (yes)	$P(u_1 = 1|c=1)$	$P(u_1 = 1|c=2)$	$P(u_1 = 1)$
$u_1 = 2$ (no)	$P(u_1 = 2|c=1)$	$P(u_1 = 2|c=2)$	$P(u_1 = 2)$
Unconditional probability	$P(c=1)$	$P(c=2)$	

($P(u_1 = 2|c = 1)$) for a question item given membership in the latent class c. These are sometimes referred to collectively as the 'error probabilities'. In our context of questionnaire development, a high false positive probability or a high false negative probability signals that there is a problem with a particular item. The primary purpose of applying LCA to questionnaire pretesting is to identify flawed questions that elicit unreliable or biased reports; such problem questions are identified via the estimated false positive and false negative probabilities.

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, 2004). For example, Biemer and Wiesen (2002) used three indicators to classify respondents regarding their use of marijuana. 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 had used marijuana or hashish in the past year (the frequency question). The third indicator was a composite that was derived from several questions related to drug intake, such as drug-related health problems or attempts to quit the taking of particular drugs: an affirmative answer to any of the questions was coded as a 'yes' on the third indicator. Biemer and Wiesen (2002) reported that the third indicator was prone to be inconsistent with the other two measurements of the use of marijuana in the past year. 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 multi-item composite (Biemer and Wiesen, 2002). In addition, LCA results showed a larger false negative rate for the recency question than for the frequency question. Combined with a more traditional analysis, Biemer and Wiesen (2002) concluded that this was because infrequent users responded falsely to the recency question but answered honestly to the frequency question.

2.2. Use of latent class analysis with two indicators

Three binary indicator variables are needed for a model of two latent classes to be just identified. However, in the survey context, three measurements are often difficult to come by. An extensive use of repeated measures can impose too much burden on respondents in a single interview and can be very expensive if multiple interviews were used.

When there are only two indicators, we can impose constraints on the parameters ρ to achieve identifiability. For instance, one possible assumption restricts the false negative probability to 0 or sets the latent class probabilities to be equal across subgroups. (For examples of various restrictive or equality assumptions, see McCutcheon (1987).) However, these assumptions are sometimes too stringent or theoretically implausible. Another way to free additional degrees of freedom is to include a group variable ($g_i = 1, 2, \ldots, G$) predicting latent class membership in the model. The grouping variable can be added to establish identifiability under various restrictions. Biemer and Witt (1996) referred to this as the *Hui–Walter* model (Hui and Walter, 1980). The grouping variable must satisfy two assumptions:

(a) the prevalence rates must be different across the levels of the grouping variable (the different prevalence assumption) and
(b) the false positive and false negative response probabilities must be equal across levels of the grouping variable (the equal error probabilities assumption).

We would, for example, assume

(i) different prevalence of each labour force category (employed or unemployed) for males and females, i.e. $P(c=1|g=1) \neq P(c=1|g=2)$ and
(ii) equal probabilities for males and females to endorse the item given their true state (class membership).

Biemer (2004) used the Hui–Walter model with gender as a grouping variable to estimate errors in the 1996 Current Population Survey labour force questions. Using data from the original survey and a reinterview, Biemer demonstrated that LCA results are consistent with both historical data on the reliability of the Current Population Survey data and theoretical expectations, suggesting that the concept of unemployment is difficult for many respondents (Biemer, 2004).

Because LCA relies on assumptions about subgroups to achieve identifiable models with only two indicators, a potential danger with LCA is that these assumptions will be invalid. Analysis under invalid assumptions can produce biased results. 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 even when the assumptions of the model are violated or their validity cannot be tested. The latter situation will be the typical one in applications of LCA for evaluating survey items. For example, when two survey questions are compared with external record data that are perceived as gold standard measures, 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.

2.3. Model fit

A popular method to compare LCA models is the use of information criteria such as the Bayesian information criterion BIC (Schwarz, 1978). BIC is defined as -2 log-likelihood $+ r \log(n)$ where r is the number of free model parameters and n is the sample size. BIC makes use of the likelihood ratio and rescales it so that a small value corresponds to a good model with large log-likelihood value and not too many parameters. Lower values of BIC are usually considered better. Sclove (1987) suggested a sample-size-adjusted BIC that, according to simulations from Yang (1998), is better suited for LCA. (For a discussion of model fit indices, see Yang (2006).)

Using the model estimates, researchers typically assign individuals to latent classes on the basis of the modal posterior probabilities. A summary measure of this classification is entropy. (On the basis of Ramaswamy *et al.* (1993), Muthén (2004) defined entropy as

$$E_K = 1 - \frac{\sum_i \sum_k -\hat{p}_{ik} \ln(\hat{p}_{ik})}{n \ln(K)}$$

with $\hat{p}_{ik}$ denoting the estimated conditional probability for individual i in class k. Entropy values range from 0 to 1, entropy values that are close to 1 indicating clear classifications.) We shall use both BIC and entropy to compare different LCA model results. The results of our model comparison do not change with the use of the sample-size-adjusted BIC.

3. Data

The data that are used in our study come from a survey that was done by the Joint Program in Survey Methodology practicum class. However, the fieldwork for the survey was done by a professional survey organization. The study was designed to examine the effects of the mode of data

collection on reports of sensitive questions. The study incorporated a check of administrative record data for some of the survey items. To study the robustness of LCA under 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 2005 practicum data are from a survey of alumni of the University of Maryland who graduated since 1989. The alumni survey has several important features that we can exploit to test the effectiveness of LCA in identifying flawed survey questions.

First, some 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 can check responses to survey questions against data at the Registrar's office with minimal risk of matching errors. The availability of multiple self-reports allows us to perform LCA and to compare the LCA results with conclusions that are based on the external measures from the Registrar's office. Therefore, we have a rare situation in which LCA results can be compared with the more traditional gold standard analysis (assuming that the record data are error free).

Second, the alumni survey included an experiment in which respondents were randomly assigned to different modes of data collection. All cases 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 one of three modes of data collection:

(a) computer-assisted telephone interviewing (CATI),
(b) 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, and
(c) a Web survey.

(For those without access to the Web, the random assignment was restricted to CATI or IVR.) We deliberately choose the mode of data collection as a grouping variable to test the Hui–Walter assumption. In this particular case neither of the two assumptions is met. Random assignment to modes of data collection should lead to equal prevalence rates across modes. At the same time we can expect different error rates across data collection modes on the basis of past findings in the survey literature (see Tourangeau *et al.* (2000)).

3.2. Survey design and data collection

The survey sample was drawn from the 55320 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 Registrar's records were used to select a random sample of 20000 graduates, stratified by graduation year. Of these 20000 graduates, 10325 could be matched to Alumni Association records containing telephone contact information. After excluding ineligible phone numbers and numbers that were used in the pretest, 7957 phone numbers were fielded for the survey. The survey interviewing was done by Schulman, Ronca, and Bucuvalas, Inc., in August and early September 2005. Call attempts were made to 7591 numbers. 24 alumni were deceased so the denominator for the response rate calculation was 7567. 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 the university. A total of 1501 alumni completed the screening and were randomly assigned to a mode of data

Table 2. Response rate and number of completers

Total	*Total*	*%*
Alumni eligible (number dialled)	7567	100.0
Screener completion	1501	19.8
Initially assigned	1501	100.0
Started main questionnaire	1094	72.9
Number of completers	1003	66.8
Response rate		13.2

Table 3. The three items that were 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?

collection; for 37 individuals without access to the Web, the random assignment was restricted to CATI *versus* IVR. A total of 1094 alumni started the main questionnaire and 1003 completed it. Multiplying the percentage of screener completion (19.8%) and the completion of the main questionnaire (66.8%), we obtain a response rate of 13.2% (Table 2). (The response rate that is reported here is equal to response rate 1 according to the standards and definitions that are set by the American Association for Public Opinion Research (2000).) Approximately a third of the respondents ($n = 320$) completed the interview via CATI. Another third ($n = 320$) completed via IVR. The final third ($n = 363$) answered the questions via the Internet.

3.3. Questionnaire and record data

The main questionnaire included 37 questions that were related to educational topics, current relationships to the university, community involvement and a final set of questions to capture perceptions of the sensitivity of some questions that were asked during the interview. We focus here on a subset of items for which we obtained multiple measures in the surveys as well as record data from the Registrar's office.

Three items were designed to tap essentially the same information about unsatisfactory or failing grades (Table 3). We deliberately designed the second of the three questions (Q18a) to be a vaguer version of the other two, hoping that it would yield higher overall classification errors. Responses to the third item were recoded according to whether the respondent reported a D or an F as his or her worst grade. (These are the two worst grades that are given out at the University of Maryland. According to the University of Maryland grading system, 'D' denotes borderline understanding of the subject. It denotes marginal performance, and it does not represent satisfactory progress towards a degree. 'F' denotes failure to understand the subject and unsatisfactory performance.)

4. Analyses and results

We first examine the three-indicator latent class model and compare the LCA estimates of the false positive and false negative probabilities for the three items with the estimates that are

obtained by direct comparison with the transcript data. These analyses assess 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 assumptions, we estimate several two-item LCA models that use one of four different covariates (grouping variables) to identify the model. The LCAs were done with Mplus version 4.1 (Muthén and Muthén, 2004). The analyses are unweighted, and only complete cases were used to keep the sample size comparable across models ($n = 954$). (Taking missing data into account changed the model results only slightly and did not affect the overall pattern of the results. The additional models treated missing data as missing at random, i.e. all available data were used to estimate the model by using full information maximum likelihood (Muthén and Muthén, 2004).)

4.1. Item analysis with external measures and latent class analysis

As a first step, we discuss the misclassification rates for each of the survey items compared with the gold standard—in this case the data from their official transcripts. We then estimate error rates for each of the survey items by using LCA and compare the gold standard and LCA estimates.

4.1.1. Misclassification rates for individual items

Of the 954 cases that were used in this analysis, 60% had received a D- or F-grade at some point during their undergraduate career. Not all of these respondents reported receiving such a grade in the survey. Table 4 presents the error rates for the three items, obtained from the comparison of survey data and academic transcripts. Of all the respondents who had received at least one D- or F-grade according to the Registrar's data, 26% failed to report this in response to question Q12 (see Table 4). They are referred to as false negative responses. Similarly, 59% failed to report their failing grades to question Q18a and 25% did not report a D- or F-grade to question Q18b.

False positive responses are cases reporting a D- or F-grade but who did not actually receive one. To continue with question Q12, of all respondents who did not receive a D- or F-grade according to the record data, roughly 3% did report having received one. It is noticeable that all three questions have much lower false positive rates than false negative rates, suggesting that the reports are distorted in a socially desirable direction.

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

It should be noted that an LCA including the true score indicator that is restricted to be a perfect indicator of the latent variable gives the same results as the comparison with the external

Table 4. Traditional gold standard analysis

Question	*% reporting having had D- or F-grade*	*False negative response rate (%)*	*False positive response rate (%)*
Q12	45.5	26.2	2.9
Q18a	25.4	59.0	1.8
Q18b	46.2	25.0	2.9

record data that was done here, i.e., under that LCA, the estimated error rates for each item are the same as those in Table 4.

4.1.2. Latent class analysis 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 seven free parameters and a BIC-value of 2392.1. Not surprisingly, this two-class model outperforms the one-class model, which tested the hypothesis that the responses to the three indicator questions are mutually independent. The one-class model had a much lower log-likelihood of −1856.2 and a BIC-value of 3733.0. Clearly, the responses to the three indicator questions are not mutually independent.

The probability of a false positive response (falsely reporting D- or F-grades) and a false negative response (falsely denying ever having received a D- or F-grade) are estimated for each of the items. Table 5 shows low false positive and false negative response rates for all three indicators with the exception of question Q18a, which was a false negative response rate of 45%.

4.1.3. Item evaluation through latent class analysis versus comparison with external data

Fig. 1 displays the results of the item evaluation by using LCA and results of the item comparison with the external record data. The broken lines represent the error rates when the survey answers are compared with the record data (true score/gold standard analysis). The full lines represent

Table 5. LCA

Question	*False negative response rate (%)*	*False positive response rate (%)*
Q12	2.1	2.5
Q18a	45.1	1.2
Q18b	1.2	3.2

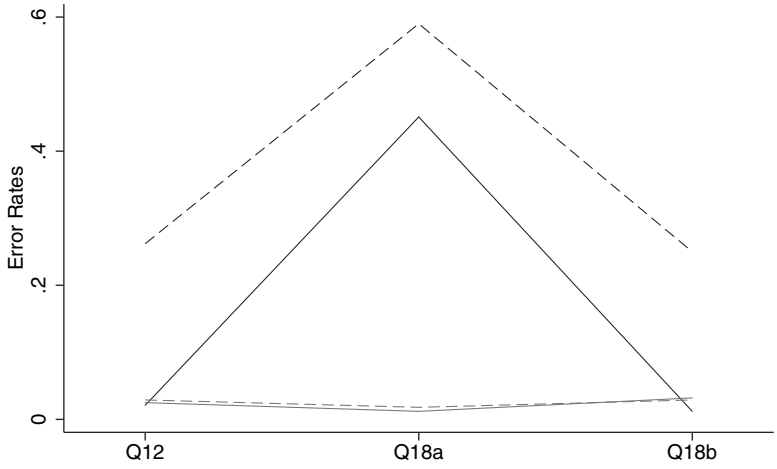

Fig. 1. False positive and false negative response rates for the LCA and true score comparison: – – –, true score, false negative; ———, LCA, false negative; — — —, true score, false positive; ———, LCA, false positive

the LCA results. The false negative response rates are displayed in black and both peak for item 18a. The false positive response rates are displayed in grey and both have their lowest value at item 18a. The graph shows that the LCA estimates of the false positive response probabilities are quite similar for all three items and also similar to those obtained from the direct comparison with academic transcripts. The LCA estimates of false negative response probabilities exhibit the same pattern as those from the gold standard analysis; question Q18a produces the highest false negative response probability, singling it out as a problematic item. Thus, the latent class approach successfully identified question Q18a as the bad item, a conclusion that is consistent with our original intention and with results from the analysis that was based on the record data.

However, even though the latent class estimates of the error rates lead to the same qualitative conclusion (that question Q18a is a flawed item), the LCA estimates of the false negative response 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 response rates for the other two items.

Although our primary interest is not estimating the proportion of alumni who received a D- or an F-grade, we nevertheless compared the prevalence rate that was obtained from LCA with the rate from analysing the record data. The LCA estimated that about 55% of survey respondents ever received a grade of D or F in college. This estimate is much closer to the true rate (60%) than the responses to any of the individual items would have indicated. The interesting question for questionnaire developers, however, is whether or not identifying assumptions can be made

(a) to detect flawed items and
(b) to reduce the number of multiple measures to 2 (and, thus, to reduce response burden).

The next section will examine the identifying assumptions in detail.

4.2. Examining identifying assumptions

When there are only two indicators, a two-class model is underidentified. As described in Section 2.2, assumptions can be imposed on the parameters to achieve identifiability. To test the robustness of these assumptions, we experimented with four potential grouping variables. The four grouping variables were gender of the respondents, respondents' grade point average (GPA), a random split of the sample and the mode of data collection. For all four grouping variables, we can use the academic transcript data to examine whether either or both of the identifying assumptions are fulfilled (see Section 4.2.1). A summary of this comparison is given in Table 6. For each of the four grouping variables, we indicate whether the levels of the grouping

Table 6. Grouping variables and identifying assumptions

Grouping variable	*Different prevalence assumption*	*Same error rate assumption*
Gender	✓	✓
GPA	✓	—
Random split	—	✓
Mode	—	—

variable differ in their prevalence rates (*assumption 1*) and whether their error rates are the same (*assumption 2*).

4.2.1. Prevalence and error rates for grouping variables

4.2.1.1. Both assumptions are satisfied with different prevalence and equal error rates for the grouping variable. Gender seems to satisfy both of the identifying assumptions. The proportion of alumni who received an unsatisfactory college grade is higher for males (68%) than for females (52%), which is a statistically significant difference ($\chi^2 = 27.27$; $p < 0.0001$). In addition, there is little reason to expect that males are more (or less) likely than females to underreport (or overreport) whether they have ever failed a class. Using the transcript data for the respondents, we see that the false positive response rates do not differ by gender for all three items. The false negative response rates for questions Q12 and Q18b do not differ by gender either. For question Q18a, however, the false negative response rate is significantly higher for females (0.65) than for males (0.54) ($\chi^2 = 6.7$; $p = 0.01$).

4.2.1.2. Satisfying the different prevalence assumption, but violating the equal error rate assumption. Respondents' GPA, by contrast, satisfies only the different prevalence assumption. About 91% of respondents whose GPA is equal to or lower than the median GPA received a D- or F-grade in college, which is a proportion that is significantly higher than the 30% for those whose GPA is higher than the median ($\chi^2 = 372.24$; $p < 0.0001$). But grouping by respondents' GPA violates the equal error probability assumption: the false positive response rates do not differ by GPA for all three items, but the false negative response rates do differ significantly by GPA group. Students with a higher GPA tend to have a lower false negative response rate than those with a lower GPA; the differences are statistically significant for all three items (for question Q12, $\chi^2 = 9.03$ and $p < 0.01$; for question Q18a, $\chi^2 = 13.54$ and $p < 0.001$; for question Q18b, $\chi^2 = 7.75$ and, $p < 0.01$).

4.2.1.3. Violating the different prevalence assumption, but satisfying equal error rate assumption. We divided the sample into two random halves and used those random half-samples as a third grouping variable. The random assignment ensures that the proportion who fail 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 that were generated for this analysis (58% *versus* 62%; $\chi^2 = 1.6$; $p = 0.21$). 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).

4.2.1.4. Violating both assumptions. Finally, using the mode of data collection as the grouping variable violates both assumptions. The respondent groups under each mode show equal rates of prevalence (because of the random mode assignment) but different error rates (because of differential mode effects in reporting). Of those respondents who were assigned to the CATI mode, 61.8% had a D- or F-grade according to the registrar's data. Of the Web respondents, 62.3% had a D- or F-grade and 58.8% of those respondents who were randomly assigned to IVR. Of those 573 respondents who had a D- or F-grade according to the transcript data, 31.5% denied having received such a grade in the CATI mode when asked question Q12, 28.8% in IVR and 20.0% in the Web mode ($\chi^2 = 7.7$; $p = 0.02$). The differences between the modes is equally pronounced for question Q18a, in particular comparing the Web mode with

Table 7. Item error rates and model fit statistic for three LCAs with 'gender' as the grouping variable and two indicator variables

Question	*False negative (%)*	*False positive (%)*	*Log-likelihood*	*BIC*	*Entropy*
Q12	2.6	17.8	−1659.5	3367.0	0.875
Q18a	28.9	1.0			
Q12	1.7	0.7	−1484.0	3016.0	0.980
Q18b	3.4	3.5			
Q18a	25.0	1.7	−1658.0	3364.0	0.889
Q18b	0.0	20.6			

IVR and CATI (the false negative response rate for CATI was 68.0%, for IVR 65.7% and for the Web mode 47.6%; $\chi^2 = 20.4$ and $p = 0.00$) and for question Q18b the results were, for CATI, 32.6%, IVR, 22.8% and Web mode 20.4%; $\chi^2 = 8.5$ and $p = 0.01$).

4.2.2. Application of the four grouping variables to latent class analysis

For each grouping variable, we compare three two-indicator LCA models. For each model, we examine false negative and false positive response rates for the individual items. We use the log-likelihood value as well as BIC and entropy to evaluate the fit of all models.

4.2.2.1. Unequal prevalence and equal error rates: gender as grouping variable. We first used gender as a grouping variable, applying the usual assumptions. We found that question Q18a consistently produced a higher false negative response rate than the other two survey items. Regardless of which two variables were entered into the latent class models, the LCA estimates of false negative response probabilities follow the same pattern as those from a gold standard analysis (Table 7). However, the LCA estimates are again consistently smaller than the true estimates. The differences between the estimated false positive response probabilities are even greater between the LCA and gold standard approaches. Furthermore, the quantitative differences between false positive and false negative response probabilities in the LCA results are also different from those which were obtained from direct analysis. It is noticeable that the model with the two 'better' indicators has a higher log-likelihood value, a lower BIC and a higher entropy value; all three indicate a better model fit for the model that does not include the item that is designed to produce higher levels of error.

4.2.2.2. Unequal prevalence and unequal error rates: grade point average as the grouping variable. By contrast, GPA fulfils the unequal prevalence rate assumption but violates the equal error rate assumption. Respondents having a GPA that is equal to or higher than the median GPA were classified into the high GPA group whereas those with a GPA that was lower than the median were grouped into the low GPA group.

The LCA estimates of error probabilities in Table 8 show that question Q18a has larger false negative response probabilities than the other two items, a conclusion that is supported by the analysis of the academic transcripts. Again, as when gender is used as a grouping variable, the LCA estimates of false negative response probabilities are consistently smaller than the true false negative response probabilities, but they follow the same trend across items. The models with GPA as the grouping variable seemed to be able to pick out the flawed question item.

Table 8. Item error rates and model fit statistic for three LCAs with GPA as grouping variable and two indicator variables

Question	*False negative (%)*	*False positive (%)*	*Log-likelihood*	*BIC*	*Entropy*
Q12	1.7	7.9	−1543.1	3134.3	0.920
Q18a	40.9	1.4			
Q12	4.4	0.7	−1361.6	2771.2	0.982
Q18b	3.4	1.2			
Q18a	40.8	1.5	−1540.6	3129.2	0.925
Q18b	0.5	8.6			

Table 9. Item error rates and model fit statistic for three LCAs with random halves as the grouping variable and two indicator variables

Question	*False negative (%)*	*False positive (%)*	*Log-likelihood*	*BIC*	*Entropy*
Q12	1.5	9.4	−1671.3	3381.5	0.974
Q18a	31.8	1.4			
Q12	0.0	2.3	−1492.4	3032.8	0.997
Q18b	1.4	4.8			
Q18a	25.6	0.6	−1669.8	3385.5	0.932
Q18b	2.5	16.6			

4.2.2.3. Equal prevalence and equal error rates: random halves as the grouping variable. 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 response probabilities again show question Q18a as the flawed question suffering from the largest false negative response probabilities (Table 9). Again, the model that does not include the deliberately vague question Q18a has better model fit statistics, with a higher log-likelihood value and a lower BIC. Note that the results that are presented in Table 9 are values that are averaged over three random splits. (We noticed some variations in the results as a function of splitting the sample in random halves. Averaging over several random splits helped to stabilize the estimates.)

Even though GPA and the random split both violated one of the two identifying assumptions, the LCA models using them as the grouping variable could still detect the flawed item Q18a.

4.2.2.4. Equal prevalence and unequal error rates: mode as grouping variable. Lastly, we used mode as a grouping variable and ran the same two-indicator models under the identifying assumptions, both of which were violated. The LCA estimates from the two-indicator models with the mode of data collection as the grouping variable are off. The LCA estimates lead to different conclusions from the analyses of academic transcripts: the LCA estimates indicate that question Q18a was a better performer than the other two items, with lower false negative and false positive response rates (Table 10). The estimated false negative response probabilities from the LCA models are consistently lower than the true false negative response rates. This is contradictory to the estimates from the gold standard analysis of the transcript data.

Table 10. Item error rates and model fit statistic for three LCAs with mode as grouping variable and two indicator variables

Question	*False negative (%)*	*False positive (%)*	*Log-likelihood*	*BIC*	*Entropy*
Q12	3.5	28.5	−2050.9	4156.6	0.992
Q18a	0.0	0.5			
Q12	2.3	3.5	−1874.2	3803.2	0.988
Q18b	0.0	3.0			
Q18a	18.3	0.0	−2048.9	4152.7	0.924
Q18b	3.7	23.7			

5. Discussion

This study examined the value of the LCA approach in evaluating survey questions. Survey researchers almost never have gold standard measures that are readily available. In the absence of true scores, LCA can be employed to assess the measurement properties of survey items. For the purpose of item evaluation, it is unrealistic to expect LCA to produce the same results as the gold standard analysis. Nevertheless, one can ask whether LCA passes various increasingly stringent tests.

(a) Can LCA identify a seriously flawed item, such as the question that we deliberately included?
(b) Does LCA reproduce other qualitative conclusions that can be gained from a gold standard analysis such as the differences in false negative and false positive response rates?
(c) Does LCA provide quantitative estimates of the error properties that are close to those derived from a gold standard analysis?

To answer these questions, we fitted a three-indicator latent model. Our results showed that the three-indicator LCA can successfully identify question Q18a as a flawed item. LCA also identified question Q18a as having the highest false negative response probabilities, thus reaching the same conclusion as a direct comparison of survey reports to the academic transcripts. It seems that the LCA can produce qualitative results that are consistent with those from the more traditional analysis with true scores. The quantitative estimates of the error probabilities from the LCA differ from those from the direct analysis (see Fig. 1). However, it should be noted that, although estimating a population percentage was not the purpose of this paper, the results show that, in the absence of a gold standard, the LCA model provides better estimates of the true proportion than any of the individual items would have, despite the presence of measurement error in all three indicators. These findings should encourage both questionnaire developers and substantive researchers to use LCA models.

In the survey context, three indicators are not always easily available. We therefore examined the robustness of the LCA when there are too few indicators to identify the models. We applied the Hui–Walter assumptions to achieve identifiability. These assumptions require prevalence rates to be different but the error probabilities to be equal across the levels of *some* grouping variable. We examined the validity of the LCA results when both Hui–Walter assumptions are 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 gold standard approach when both or only one of the assumptions is satisfied. For

instance, when gender or the GPA of the respondents are used as grouping variables, question Q18a is consistently shown to have the largest false negative response probabilities. (We deliberately wrote this item to be flawed.) However, when both assumptions are violated (when mode was the grouping variable), the LCA results are misleading; the LCA did not identify question Q18a as the item producing the largest error probabilities. In addition, we find that the quantitative estimates from LCA do not always agree with results from traditional gold standard 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 was violated.

Our results suggest that the two-indicator LCA models can tolerate violations of their identifying restrictions only to a limited extent. Survey designers who plan to use two-indicator LCA models to identify flawed items should have strong theory or prior data that helps them to identify good grouping variables and to assess how well the grouping variable(s) meet the identifying restrictions. In the future, we plan to carry out simulation studies to examine the relationship between the extent of violations and the validity of LCA results. In summary, the results suggest that the LCA models can pick out flawed items, but that the quantitative estimates of error rates cannot be taken literally. Still, in the absence of true scores, LCA can provide useful quantitative information about which items are working and which are not.

Acknowledgements

This study is based on work that was supported by the National Science Foundation under grant SES 0550002. We thank the Methodology, Measurement, and Statistics Program and Dr Cheryl Eavey for their support. Any opinions, findings and conclusions or recommendations that are 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. We thank Paul Biemer, Carolina Casas-Cordero, Elisabeth Coutts, Jill Dever, Stephanie Eckman, Ulrich Kohler, Michael Lemay, Katherine Masyn, Stanley Presser, Rainer Schnell and Elizabeth Stuart for critical comments and helpful suggestions. We are especially grateful to Katharine Abraham and Mirta Galesic who oversaw the data collection and allowed us to include three redundant items at the eleventh hour.

References

American Association for Public Opinion Research (2000) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Lenexa: American Association for Public Opinion Research.

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

Biemer, P. P. and Wiesen, C. (2002) Measurement error evaluation of self-reported drug use: a latent class analysis of the US National Household Survey on Drug Abuse. *J. R. Statist. Soc.* A, **165**, 97–119.

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

Clogg, C. C. and Goodman, L. A. (1984) Latent structure analysis of a set of multidimensional contingency tables. *J. Am. Statist. Ass.*, **79**, 762–771.

Conrad, F. and Blair, J. (2004) Data quality in cognitive interviews: the case of verbal reports. In *Methods for Testing and Evaluating Survey Questionnaires* (eds S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin and E. Singer), pp. 67–87. New York: Wiley.

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

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

McCutcheon, A. L. (1987) *Latent Class Analysis*. Beverly Hills: Sage.

Muthén, B. (2004) *Mplus: Statistical Analysis with Latent Variables*, technical appendices. Los Angeles: Muthén and Muthén.

Muthén, L. K. and Muthén, B. (2004) *Mplus User's Guide*. Los Angeles: Muthén and Muthén.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J. and Singer, E. (eds) (2004) *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.
Ramaswamy, V., DeSarbo, W., Reibstein, D. and Robinson, W. (1993) An empirical pooling approach for estimating marketing mix elasticities with pims data. *Marktng Sci.*, **12**, 103–124.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Sclove, L. S. (1987) Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, **52**, 333–343.
Tourangeau, R., Rips, L. and Rasinski, K. (2000) *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
Yang, C. C. (1998) Finite mixture model selection with psychometrics application. *PhD Dissertation*. University of California, Los Angeles.
Yang, C. C. (2006) Evaluating latent class analysis models in qualitative phenotype identification. *Computnl Statist. Data Anal.*, **50**, 1090–1104.