Local dependence in latent class models: application to voting in elections

DL Oberski

Abstract Demonstrates methods of detecting local dependence in binary data latent class models. Latent class models are applied to five repeated measures of voter turnout in the Dutch Parliamentary elections of 2006 and 2010 obtained from a probability sample of 9510 citizens. Modeling substantive local dependence as separate discrete latent variables while modeling nuisance dependencies as direct effects yields an interpretable model, giving insight into the classification errors present in survey questions about voting. The procedure followed stands in contrast to the "standard" procedure of increasing the number of latent classes until information criteria are satisfactory.

Key words: local independence, bivariate residual, score test, misclassification, measurement error, latent variables, turnout.

1 Introduction

Latent class models for binary variables are finite mixtures of binomials, and applied in a broad range of fields including the social sciences (Hagenaars and McCutcheon, 2002; Savage et al., 2013), machine learning (Hastie et al., 2008), psychological measurement (Heinen, 1996), public health and epidemiology (Collins and Lanza, 2010), and the biomedical sciences (Walter et al., 2013).

The key assumption of latent class models is conditional independence of the observed variables given the latent class or mixture component. Violations of this assumption may occur when there are unmodeled latent classes – a common reac-

DL Oberski
Department of Methodology and Statistics, Tilburg University
Room P1105, PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: doberski@uvt.nl

tion to detected misfit is therefore to increase the number of classes, based on criteria such as L^2, χ^2, (C)AIC, BIC, CVIC or ICL (McLachlan and Peel, 2000). However, the local dependence and pursuant additional latent classes may not be substantively interesting. For example, Hagenaars and McCutcheon (2002) suggested that local dependence between items in a questionnaire or psychological test can occur because respondents attempt to make their responses consistent; and Oberski and Vermunt (frth) found that ethnicity measurements discussed by Johnson (1990) were locally dependent due to the fact that some were measured on the same occasion. In these instances, additional classes do not yield substantively useful results. Local dependence, is however, still of importance in such cases, because unmodeled local dependencies may bias model parameters of interest as well as posterior classifications (Vacek, 1985; Qu et al., 1996; Hadgu et al., 2005).

This paper demonstrates an alternative to increasing the number of classes: modeling additional discrete latent variables when the dependence between items is substantively interesting, while modeling local dependencies directly when dependence is considered a nuisance. Such an approach will not be fruitful in all applications of latent class models, since a strong theory is needed regarding the type of dependencies deemed "substantively interesting". We discuss an application in which such judgements can be made, and the approach is therefore useful.

It is not generally desirable (or feasible) to model all possible local dependencies. Which dependencies should be modeled can be monitored with the "bivariate residual" (BVR). The BVR for a pair of observed variables is the Pearson residual in their bivariate cross-table (Vermunt and Magidson, 2005, pp. 72-3). It is equivalent to an uncorrected score test for freeing an additional local dependence parameter (Oberski et al., frth). When this measure is "large", it will indicate a potential omitted local dependence. However, Oberski et al. (frth) showed that the raw BVR does not follow a known distribution, and recommended using either the score test or parametrically bootstrapped p-values for the BVR, both of which have been implemented in the standard latent class software Latent Gold 5 (Vermunt and Magidson, 2005). Here we demonstrate the use of bootstrapped p-values for the BVR.

Application of the BVR is demonstrated on an analysis of voting in elections. What drives citizens to turn out in elections is a topic that receives intense interest in political science (e.g. Campbell, 1960; Verba and Brady, 1995; Franklin, 2004; Gallego and Oberski, 2012). Even so, citizens' turnout decisions are often not directly observed but rather the answer to the survey question "did you vote in the last election?" is observed. That these two things are not the same is well-known from validation studies (see Ansolabehere and Hersh, 2012, for a recent overview), and means that misclassification should be estimated so that parameter estimates may be corrected (Vermunt, 2010). Obtaining such gold standard validation data, is, however, sometimes prohibited by law and always costly. As an alternative to validation data, repeated measurements may be available, which raise the question "Do repeated measurements without a gold standard allow for the estimation of misclassification rates in voting?". This paper analyses such data obtained from a representative sample of Dutch citizens by applying latent class models with mul-

tiple discrete latent variables and local dependence parameters, demonstrating the use of local dependence measures.

2 Data

The LISS panel is a probability sample of 5000 Dutch households whose members are interviewed regularly over the web. All individuals aged 16 or more in the selected household are invited to participate. In this study we only consider respondents who are eligible to vote. These participants were asked whether they had voted in the Parliamentary elections held in the Netherlands in November 2006 (official turnout 80.4%) and June 2010 (turnout 75.4%). For more information on the design of the study, response rates, and recruitment efforts, please see Scherpenzeel (2011).

Participants were asked whether they had voted in these two elections on five occasions: in 2008, 2009, and 2010 (for the 2006 election), and 2011 and 2012 (for the 2010 election). The percentages of respondents who claimed to have voted were 87%, 84%, 81%, 87%, and 84% respectively. All of these differences are statistically significant ($p < 0.01$)[1]. Initially reported turnout thus exceeded actual turnout, but, even though the same respondents were asked whether they had voted in the same elections, over time the claimed turnout rate declined toward the actual turnout rates.

In total there were 9510 respondents for whom at least one answer was recorded; if all rows containing at least one missing value were deleted, only 2424 would remain, however. We therefore estimate our model using full-information maximum likelihood, which assumes the missing values are MAR given the observed values.

3 Model

3.1 Latent class model with possible local dependencies

Suppose an i.i.d. sample of size N is obtained on J observed binary variables, aggregated by the R response patterns into $\mathbf{Y}$. Let $\mathbf{n}$ be the R-vector of observed response pattern counts. We also postulate K discrete latent variables ξ_k, collected in a vector ξ, whose distribution is to be estimated. The K-way cross-table of ξ yields T unobserved patterns. In the case of latent structure analysis, there is only one discrete latent variable and T will equal the number of latent classes. The log-likelihood for the latent class model is then the discrete mixture (e.g. Formann, 1992)

[1] The percentages shown here were calculated using pairwise deletion but the pattern and statistical significance persist when using listwise deletion. The differences are therefore not likely due to panel attrition (nonresponse).

$$\ell(\theta) = \mathbf{n}' \log \Pr(\mathbf{Y}) = \mathbf{n}' \log \left[\sum_T \Pr(\mathbf{Y}|\xi) \Pr(\xi) \right], \tag{1}$$

where log and exp denote elementwise operations,

$$\Pr(\mathbf{Y}|\xi) = \frac{\exp(\eta_{\mathbf{Y}|\xi})}{\mathbf{1}_R' \exp(\eta_{\mathbf{Y}|\xi})}, \qquad \text{and} \qquad \Pr(\xi) = \frac{\exp(\eta_\xi)}{\mathbf{1}_T' \exp(\eta_\xi)}. \tag{2}$$

The GLM linear predictors $\eta_{\mathbf{Y}|\xi}$ and η_ξ are parameterized using effect-coded design matrices (Evers and Namboodiri, 1979):

$$\eta_{\mathbf{Y}|\xi} = \mathbf{X}_{(Y)}\tau + \mathbf{X}_{(YY)}\psi + \mathbf{X}_{(Y\xi)}\lambda, \qquad \text{and} \qquad \eta_\xi = \mathbf{X}_{(\xi)}\alpha + \mathbf{X}_{(\xi\xi)}\beta, \tag{3}$$

where $\mathbf{X}_{(Y)}$, $\mathbf{X}_{(YY)}$ and $\mathbf{X}_{(Y\xi)}$ are design matrices for the observed variables' main effects τ, bivariate associations ψ, and associations with the latent discrete variables λ ("slopes"), respectively. Similarly, $\mathbf{X}_{(\xi)}$ and $\mathbf{X}_{(\xi\xi)}$ are design matrices for the discrete unobserved variables' main effects α and associations β. This parameterization of the local dependence latent class model is similar to that adopted by Hagenaars (1988) and Formann (1992, section 4.3), except that we additionally allow for explicit modeling of multiple discrete latent variables and their interrelations (Magidson and Vermunt, 2001; Vermunt and Magidson, 2005).

The q-vector of parameters θ can be defined as $\theta' := (\alpha', \beta', \tau', \lambda', \psi')$. There are thus $q \leq T(J+1) - 1 + \binom{J}{2}$ (possible) parameters. The standard local independence latent class model, however, has as its key assumption that $\psi = \mathbf{0}$. In addition, the slopes λ are typically restricted such that, given exactly one unobserved discrete variable, each indicator is conditionally independent from all other latent variables.

Maximum likelihood estimates of the parameters of the model are usually obtained as $\hat{\theta} = \arg\max_{\theta \in \mathbb{R}^q} \ell(\theta)$ by expectation-maximization (see Formann, 1992), quasi-Newton methods, or a combination of both (Vermunt and Magidson, 2005). Goodman (1974) showed that the parameters of the model are locally identifiable when the Jacobian $\mathbf{S} := \partial \Pr(\mathbf{Y}) / \partial \theta$ is of full column rank. A necessary but not sufficient condition for this is that $R > q$. In practice, local identifiability can be evaluated empirically by examining the rank of the information matrix at the maximum likelihood solution, or by randomly sampling many parameter values in the parameter space and evaluating the information matrix at each point (Forcina, 2008). For a general discussion of identification in latent class models, we refer to Huang and Bandeen-Roche (2004); for a discussion of identifiability of the local dependence parameters, see Oberski and Vermunt (frth, appendix).

3.2 Model misfit and local dependence

After estimation, for each response pattern expected frequencies $\hat{\mu}_r := N \cdot \Pr(\mathbf{Y}_r | \theta = \hat{\theta})$ are obtained, which can be compared with the observed frequencies n_r. Overall goodness of fit measures based on this comparison such as the chi-square and likeli-

hood ratio (L^2), as well as information criteria such as BIC, AIC, CAIC, CVIC, and ICL are often used to evaluate whether the latent class model adequately describes the observed data (see McLachlan and Peel, 2000, chapter 6).

Since the key assumption is that of local independence ($\psi = \mathbf{0}$), a major source of misfit will be locally dependent item pairs. In our example, local dependence may, for instance, arise because respondents remember their answer on the first measurement occasion and try to remain consistent on later occasions (Hagenaars and McCutcheon, 2002). Assuming the model is overidentified, such local dependence will be picked up by the overall fit statistics and information criteria. When these indicate a problem, additional latent classes are then included in the model to account for the dependence. This will lead to a latent class model in which some of the classes represent, for instance, "consistent answering".

However, local dependencies and the pursuant additional classes are not necessarily of scientific interest. For theoretical reasons, one may prefer a model with fewer classes in the voting data application: we know that respondents have either voted or not and that the measurements pertain to two separate elections. Two classes are also preferred when evaluating diagnostic tests for disease/non-diseased status (Qu et al., 1996).

When a specific number of classes is desired or local dependence is not substantively meaningful, it may be preferable to model local dependencies by freeing elements of ψ. Freeing all local dependencies is, however, usually not desirable for reasons of model stability and (sometimes) identifiability (Oberski and Vermunt, frth). We therefore use the "bivariate residual" (BVR) between item pairs to monitor whether it might be necessary to free local dependencies (Vermunt and Magidson, 2005). The bivariate residual is an intuitively attractive fit index measuring the degree to which the bivariate cross-table between a pair of observed variables fits the model:

$$\text{BVR}_{jj'} := \sum_{k\in\{0,1\}} \sum_{l\in\{0,1\}} \frac{(n_{kl} - \hat{\mu}_{kl})^2}{\hat{\mu}_{kl}} = r_{11}^2 \sum_{k\in\{0,1\}} \sum_{l\in\{0,1\}} \frac{1}{\hat{\mu}_{kl}}, \tag{4}$$

where the raw residuals $r_{kl} := n_{kl} - \hat{\mu}_{kl}$, and n_{kl} and $\hat{\mu}_{kl}$ now indicate observed and expected frequencies in the bivariate 2×2 cross-table of the observed variables y_j and $y_{j'}$ ($j \neq j'$). The last step follows because the marginals are perfectly reproduced (Oberski et al., frth). A BVR can be obtained for each of the $\binom{J}{2}$ pairs of observed variables; in this way, for each pair it can be investigated whether the cross-table between this pair appears to fit the hypothesis of local independence.

The BVR has the same form as a Pearson residual and is often treated in applied research as though its asymptotic distribution converged to a chi-square distribution. Oberski et al. (frth) showed that this is not a good practice; the BVR is a score test uncorrected for cell interdependencies and far from chi-square distributed. Instead, p-values for the BVR very close to Rao (1948)'s classic efficient score test can be obtained by a parametric bootstrap (Efron, 1982; Langeheine et al., 1996). The software Latent Gold 5.0 (Vermunt and Magidson, 2005) implements this procedure.

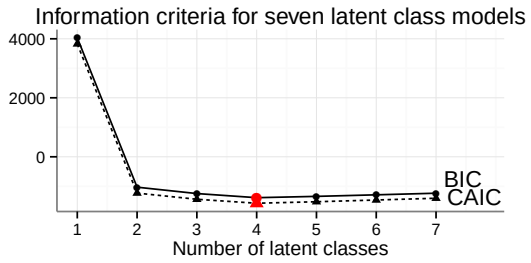


Fig. 1 Model selection increasing the number of classes. Using both BIC and CAIC the four-class solution would be selected.

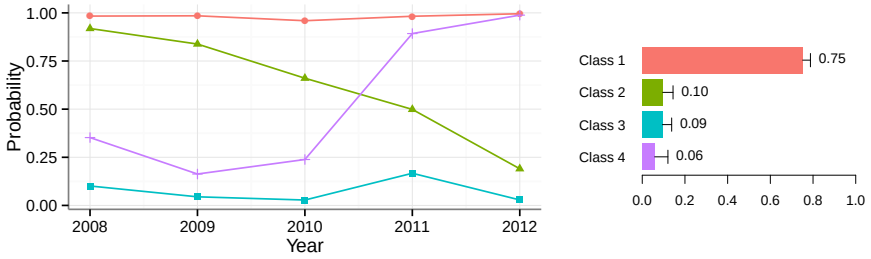
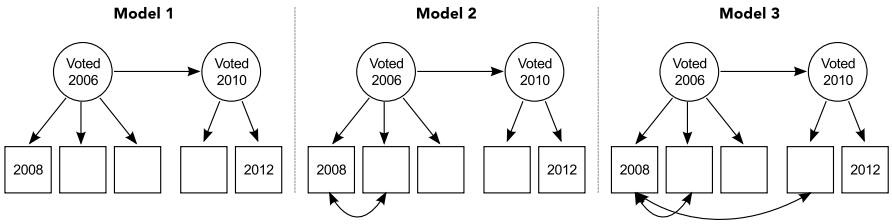


Fig. 2 Left: probability profile plot for the four-class solution. Right: legend with estimated class sizes and 2 s.e. error bars.

4 Results

We now follow two procedures for data analysis of the Dutch voting example. The first procedure is a standard single nominal latent class model, which is fitted to the data with an increasing number of classes. BIC and CAIC are used to select the number of classes, after which these are interpreted. We compare this standard procedure with one in which two discrete latent variables are modeled jointly, one for voting in each of the 2006 and 2010 elections, and the bivariate residuals are inspected to decide which local dependencies should be freed.

Figure 1 shows criteria used to select the number of classes. Both BIC and CAIC select the four-class model. When this model is fit to the five claims of having voted, the conditional probabilities shown in Figure 2 result. The left-hand side of Figure 2 shows the probability of claiming to have voted on each of the five measurement occasions given the four latent classes, indicated by the different lines (colors, point shapes). The right-hand side of Figure 2 provides a legend and shows class size estimates with 2 s.e. error bars. Figure 2 shows that class 1 is the class of people who voted in both elections, while class 3 is voting in neither election. Class 4 appears to represent voting in 2010 but not in 2006, although the probability of claiming to have voted in 2006 in this class is still around 0.25. Class 2, containing 10% of observations, is the most difficult to explain; it contains people who initially claim to have voted, but, as time goes by, become more likely to admit that they did not.



	2008	*2009*	*2010*	*2011*
2009	0.000			
2010	0.586	0.116		
2011	0.000	0.808	0.004	
2012	0.460	0.148	0.000	0.688

L^2 p-value: 0.000.

	2008	*2009*	*2010*	*2011*
	0.736			
	0.138	0.038		
	0.000	0.060	0.358	
	0.022	0.004	0.036	0.714

L^2 p-value: 0.000.

	2008	*2009*	*2010*	*2011*
	0.622			
	0.120	0.272		
	0.554	0.130	0.222	
	0.296	0.030	0.234	0.824

L^2 p-value: 0.052.

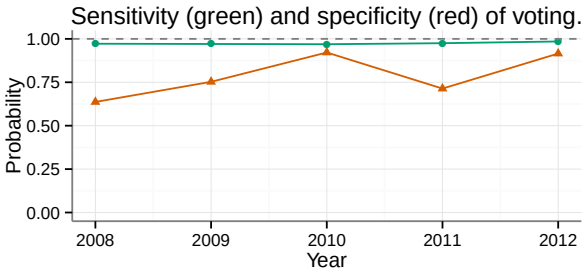
Fig. 3 Top: three sequential models, starting from the conditional independence model (Model 1). Bottom: bootstrapped p-values for the bivariate residuals and L^2 under each model.

The standard latent class model procedure applied to these data is somewhat unsatisfactory. Considering that there are only two actual elections, one would expect the four classes to represent the $2 \times 2 = 4$ cells in the cross-table of voting or not in 2006 and 2010. Four classes are indeed selected, but instead of a class "voting in 2006 and not in 2010", the difficult-to-interpret class 2 results, which partially also represents artefacts that are not of interest to political scientists.

An alternative procedure is to fit a model with two discrete latent variables, one for each election, each with two classes (voted/did not vote). The first three answers, being about the 2006 elections, are related to the first latent variable and the last two answers, about the 2010 elections, to the second latent variable. Conditional probabilities then represent misclassification rates with respect to true turnout in the 2006 and 2010 elections, which is the question of scientific interest.

Initially a model is fit in which all $\binom{5}{2} = 10$ possible local dependencies are set to zero. This "Model 1" is shown as a graph in Figure 3. The table under "Model 1" in Figure 3 provides p-values for the 10 bivariate residuals obtained by parametric bootstrapping. The BVR's of the dependence between answers in 2008 and in other years correspond to Hagenaars and McCutcheon (2002)'s suggestion that respondents sometimes attempt to make their answers consistent with the first occasion. Based on these and the values of the BVR's (not shown for conciseness), we free the local dependence between the answer in 2008 and in 2009 and re-fit the model to obtain the model and BVR p-values shown under "Model 2". One p-value is then still < 0.01 and in line with the memory effect theory: the corresponding dependence is therefore freed. The final model ("Model 3") does not have any BVR with a bootstrapped p-value < 0.01. The overall bootstrapped likelihood ratio test L^2 indicates a good fit as well.

This final model has several advantages over the four-class model. First, it explicitly models true turnout in the two elections so that the conditional probabilities



	Voted in 2006		
	No	Yes	
2010			
No	0.713	0.051	18%
Yes	0.287	0.949	82%
	19%	81%	

Fig. 4 Left: probability profile (i.e. sensitivity and specificity) plot for the two-class, two-variable solution. Right: turnover table of "true vote" classes.

may be interpreted as misclassification rates ("specificity" and "sensitivity"). These misclassification rates are of interest to political scientists. Second, the two-variable classification allows researchers to relate voting in these two elections to external variables (Vermunt, 2010). Third, nuisance local dependencies such as memory effects are not part of the classification but are accounted for by local dependence parameters.

Sensitivity and specificity (misclassification rates) are shown on the left-hand side of Figure 4. The Figure shows that the probability of a respondent claiming to have voted when they have not decreases as the election period becomes more distant. This finding corresponds to the idea that false positives are due to social desirability, since the "norm" of voting will be less salient three years after the election than during election season[2]. This pattern explains the overall pattern that claimed turnout rates approached the actual turnout rates as time goes by.

The right-hand side of Figure 4 shows the estimated turnover table of true turnout from 2006 to 2010 with class sizes in the margins. The class prevalences of 81 and 82 percent are higher than the actual turnout rates 80.4 and 75.4 percent, although they are much closer to true turnout than the raw reported rates (around 87%). The turnover table suggests that voters mostly remained voters whereas non-voters in 2006 had a chance of 0.287 of voting in the 2010 election. If such a pattern were to be predicted for future elections, it would suggest that efforts to encourage citizens to vote would be best focused on non-voters in previous elections.

5 Summary

This paper applied latent class modeling with multiple latent variables and local dependencies to voter turnout data. It illustrated the use of such models, particularly the role that bivariate residuals and other measures of residual local dependence such

[2] Note that 2010 was not known in advance to be an election season by respondents since elections were called due to the sudden collapse of the government.

as the score test may play in practical data analysis. The approach of modeling substantive dependencies as discrete latent variables while modeling non-substantive dependencies as local dependence parameters yielded a useful model that gave insight into questions of interest to political scientists. In our opinion this approach was more fruitful than the four-class nominal latent class solution.

References

Ansolabehere, S. and Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*.

Campbell, A. e. a. (1960). *The American Voter*. Wiley, New York.

Collins, L. M. and Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, volume 718. Wiley.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics*, 38.

Evers, M. and Namboodiri, N. (1979). On the design matrix strategy in the analysis of categorical data. *Sociological methodology*, 10:86–111.

Forcina, A. (2008). Identifiability of extended latent class models with individual covariates. *Computational Statistics & Data Analysis*, 52(12):5263–5268.

Formann, A. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418):476–486.

Franklin, M. (2004). *Voter turnout and the dynamics of electoral competition in established democracies since 1945*. Cambridge University Press, New York.

Gallego, A. and Oberski, D. (2012). Personality and political participation: The mediation hypothesis. *Political Behavior*, 34:424–451.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215.

Hadgu, A., Dendukuri, N., and Hilden, J. (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology*, 16(5):604–612.

Hagenaars, J. A. P. (1988). Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research*, 16(3):379–405.

Hagenaars, J. A. P. and McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge University Press Cambridge United Kingdom:.

Hastie, ., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage, Thousand Oaks, CA.

Huang, G. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Johnson, R. (1990). Measurement of hispanic ethnicity in the us census: An evaluation based on latent-class analysis. *Journal of the American Statistical Association*, 85(409):58–65.

Langeheine, R., Pannekoek, J., and Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24(4):492–516.

Magidson, J. and Vermunt, J. K. (2001). Latent class factor and cluster models, biplots, and related graphical displays. *Sociological Methodology*, pages 223–264.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*, volume 299. Wiley-Interscience.

Oberski, D., Van Kollenburg, G., and Vermunt, J. (frth). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models.

Oberski, D. and Vermunt, J. (frth). The Expected Parameter Change (EPC) for local dependence assessment in binary data latent class models.

Qu, Y., Tan, M., and Kutner, M. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, pages 797–810.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge Univ Press.

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., Friedman, S., and Miles, A. (2013). A new model of social class? findings from the bbc's great british class survey experiment. *Sociology*, 47(2):219–250.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: How the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 109(1):56–61.

Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, pages 959–968.

Verba, S., S. K. and Brady, H. (1995). *Voice and equality: Civic voluntarism in American politics*. Harvard University Press, Cambridge, MA.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18:450–469.

Vermunt, J. K. and Magidson, J. (2005). Technical guide for latent GOLD 4.0: Basic and advanced. *Belmont Massachusetts: Statistical Innovations Inc*.

Walter, S. D., Riddell, C. A., Rabachini, T., Villa, L. L., and Franco, E. L. (2013). Accuracy of p53 codon 72 polymorphism status determined by multiple laboratory methods: A latent class model analysis. *PloS one*, 8(2):e56430.