ISER Working Paper Series

Assessing and relaxing assumptions in quasi-simplex models

Alexandru Cernat
Peter Lugtig
S.C. Noah Uhrig

Institute for Social and Economic Research
University of Essex

Nicole Watson

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne

No. 2014-09
February 2014

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non-Technical Summary

Panel data (repeated measures of the same individuals) has become more and more popular in research as it has a number of unique advantages such as enabling researchers to answer questions about individual change and help deal (partially) with the issues linked to causality. But this type of data has some special limitations as well, such as the training effect of respondents and gradual drop-out from the survey (i.e., attrition).

In this context an approach that evaluates data quality using reliability (the amount of the true value as opposed to random noise) in panel data has been proposed in previous research. This approach, named the quasi-simplex model, brings a number of innovations but also makes a number of strong assumptions about the data, such as: the absence of memory effects of respondents or equal error over time. This paper aims to assess the validity and impact of these assumptions as these have largely not previously been examined

Our research shows that most of the previously made assumptions hold and more often than not the model can be even more restrictive. But, even if this is true, four out of the 22 circumstances analysed here presented violations of an assumption that lead to different results. Our research shows that when processes such as the respondent memory effect are present in the data it can lead to overestimation of reliability and underestimation of stability in time.

Assessing and relaxing assumptions in quasi-simplex models

Alexandru Cernat[a] (acerna@essex.ac.uk)

Peter Lugtig[a b] (plugtig@essex.ac.uk)

S.C. Noah Uhrig[a] (scnuhrig@essex.ac.uk)

Nicole Watson[c] (n.watson@unimelb.edu.au)

[a] Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK.
[b] Department of Methods and Statistics, Utrecht University, Padualaan 14, 3508 TC, Utrecht, the Netherlands.
[c] Melbourne Institute of Applied Economic and Social Research, 111 Barry Street, The University of Melbourne, Victoria, 3010, Australia.

Abstract:

The quasi-simplex model makes use of at least three repeated measures of the same variable to estimate its reliability. The model has rather strict assumptions about how various parameters in the model are related to each other. Previous studies have outlined how several of the assumptions of the quasi-simplex model may be relaxed using more than 3 waves of data. It is unclear however whether the assumptions of the quasi-simplex model are *overly* strict. In other words, it is not known whether relaxing the assumptions results in better models or different substantive conclusions with regard to the reliability of survey measures. Using data from the British Household Panel Survey this paper shows how the assumptions of the quasi-simplex model can be relaxed. We conclude that relaxing the assumptions in practice seldom leads to a better model or different conclusions than the traditional quasi-simplex model.

Keywords: quasi-simplex model, reliability, panel data, BHPS, measurement error

JEL Codes: C330, C520

INTRODUCTION

Measurement is one of the most important and complex aspects of research in the social sciences. And, while good measurement is essential for valid scientific inference, it is marred by unknowns regarding validity (systematic error) and reliability (random error). In this paper we are especially interested in the latter. Stemming from the initial theoretical development made by Lord and Novick (1968), two main approaches to modelling this type of error have developed. The first of them uses multiple items that measure the same dimension in order to parse out random or unique variances from common variance (Alwin, 2007; Bollen, 1989). The other uses multiple measures in time of the same item to reach the same goal (Alwin, 2007; Heise, 1969; Wiley & Wiley, 1970). The most widely used model for the first approach is Confirmatory Factor Analysis (CFA) (and equivalent approaches such as Item Response Theory or Latent Class Analysis) while for the second, researchers use the quasi-simplex model (QSM) (or the Latent Markov Chain in the case of the categorical variables).

The QSM (Heise, 1969; Wiley & Wiley, 1970) is used to estimate the reliability of a single variable that is measured repeatedly at least three times. The model has been used in research on attitude stability and attitude formation (Alwin & Krosnick, 1991; Alwin, 1989), development studies (e.g. Bast & Reitsma, 1997) or to test the quality of survey questions (Alwin, 2007; Saris & Van Den Putte, 1988). And although the multiple items approach to estimate reliability has been more popular in recent decades, the multiple measures design has a number of characteristics that make it attractive in certain contexts. Firstly, the QSM results in reliabilities that are closer to the definition initially put forward by Lord and Novick (1968), i.e., the percentage of variance due to the true score as opposed to random error. Using multiple items will almost always result in data that contain common variances, item specific variances, and measurement error, and separating the three is impossible. The QSM results in a different estimation of reliability from the multiple items model in the sense that all variances in the model are either due to common variance or measurement error (Alwin, 2007). The QSM also has the advantage that it can be used for standalone items, not part of a scale.

Although the QSM has some benefits it also has a number of limitations. The first one is the need for at least three repeated measures of the same item. Although the model is just-identified in this case, the resulting parameter estimates are sometimes implausible and standard errors may be large (Alwin, 2007; Palmquist & Green, 1992; Wiley & Wiley, 1970).

Also, the model may fail to converge altogether (Cernat, 2013; Coenders, Saris, Batista-Foguet, & Andreenkova, 1999; Hargens, Reskin, & Allison, 1976; Jagodzinski & Kuhnel, 1987).

Past studies on the properties of QSM to measure the reliability of survey questions have centered on the appropriate time between two waves (Jagodzinski & Kuhnel, 1987), how ordinal data should be modelled (Alwin, 2007), and how means should be incorporated into the model (Mandys, Dolan, & Molenaar, 1994). These have increased our understanding of the model and its possible limitations. But, some of the most important assumptions of the QSM have been ignored so far. The assumptions include a diverse set of preconditions and model convergence issues and implausible parameter estimates are sometimes linked to their violation (Jagodzinski & Kuhnel, 1987).

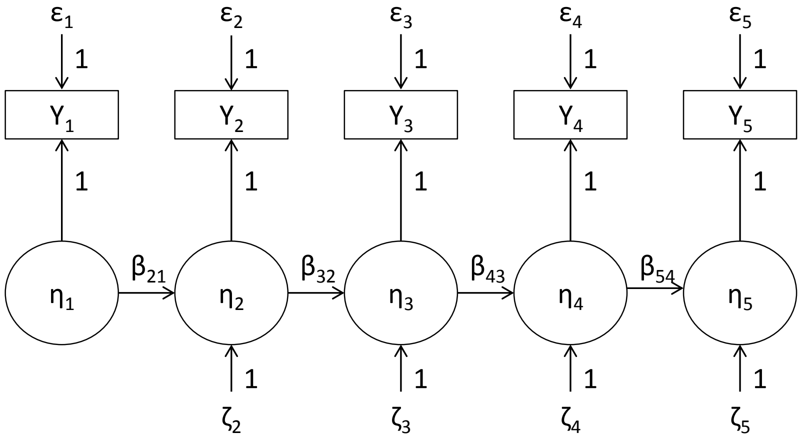
In this paper we illustrate how several of the strict assumptions of QSM can be relaxed when more than three waves of data are used. Using examples from the British Household Panel Study (BHPS) we demonstrate how relaxing those assumptions sometimes leads to better fitting models as compared to the traditional QSM. Also, we show that the substantive conclusions drawn from the reliability and stability parameters sometimes change when specific assumptions are relaxed. We conclude with a discussion of implications and recommendations for testing QSM assumptions when more than three waves of data are available.

THE QUASI SIMPLEX MODEL

The basic quasi-simplex model as shown in ***Figure 1*** can be summarized in two related equations (equations 1 and 2 below) that link the observed responses (Y) to a latent true score (η) at every time t (i.e., the measurement part of the model). Following the theory of the true-score model (Lord & Novick, 1968), the observed score at each time point consists of the true score and measurement error (ε).

$$Y_t = \eta_t + \varepsilon_{t,} \tag{1}$$

Figure 1: Quasi-simplex model with five measurements.



The quasi-simplex model uses repeated observation of the same variables to separate η from ε. Subsequent measurements are linked only by a stability coefficient between two true scores at times *t* and *t-1* ($ß_{t,t-1}$), and a random disturbance term (ζ_t) that represents the time specific true score (or noise)[1].

$$\eta_t = ß_{t,t-1}\eta_{t-1} + \zeta_t \tag{2}$$

The quasi-simplex model is only empirically testable when several assumptions about the relations between the estimated parameters ε_t, η_t and ζ_t are made:

1. Independence of observations over time. This implies that both ε_t and ζ_t are uncorrelated over time.

2. The mean Y_t and η_t are 0. This implies that all variables are normalized and that one is not interested in the development of means over time. This assumption usually remains implicit in the quasi-simplex model, as the model with centered variables is equivalent to the model with estimated means as long as no constraints are imposed on the means (see Blok & Saris, 1983).

3. Every true score (η_t) is only explained by the true score at the previous wave ($\eta_{t,t-1}$), leading to a lag-1 process of change.

[1] To identify the true score at wave one this score is equivalent to the disturbance term:

$$\eta_1 = \zeta_1 \tag{3}$$

4. Equality of measurement error (ε_t) over time. It is necessary to constrain some of the parameters to be equal over time in order to allow the model to be identified. Heise (1969) favoured the idea that the reliabilities should be equal over time. As the reliabilities can be calculated as the ratio between the true score variance (V(η_t)) and the observed variance (V(Y_t)), this means that the ratio between the two remains stable, and that any difference in the variance of the observed variables is reflected in the true score and both variances. Alternatively, Wiley & Wiley (1970) believed that the error variances V(ε_t) should be constrained to be equal at every time t. When the observed variances over time do not differ, both of these assumptions lead to identical results. When the observed variances do differ over time, restricting the reliability to be equal will lead to differences in the estimated error variances V(ε_t). Conversely, the Wiley and Wiley specification will lead to slight differences in the reliability estimate over time.

5. The variances of the error term V(ε_t) and disturbance V(ζ_t) follow a normal distribution, with a mean of 0.

6. The covariances between the true scores (η_t), errors (ε_t), and disturbances (ζ_t) are zero. (E(ε_t, η_t)=0, E(ε_t, ζ_t)=0 and E(η_t, ζ_t)=0).

For a more elaborate discussion and closed-form derivation of all model parameters, we refer to Alwin (2007). With three waves of data, a quasi-simplex model that implements the restrictions outlined above will be just-identified. Any more waves of data will lead to a degrees of freedom larger than 0, and thus enable a test of the model fit.

HOW ASSUMPTIONS MAY BE VIOLATED IN THE QUASI-SIMPLEX MODEL

As we will show below, all six assumptions of the quasi-simplex model are likely to be violated under certain circumstances, depending on the study's population, the variable of interest, and the measurement procedure used for those variables. Violations of the assumptions may, but do not necessarily, lead to estimation problems in the quasi-simplex model (Coenders et al., 1999; Jagodzinski & Kuhnel, 1987). In many cases, either the quasi-simplex model will not fit the data (if degrees of freedom are larger than 0), fit the data badly, and/or the parameter estimates of the model will become biased because of violations of model assumptions (Cernat, 2013; Hargens et al., 1976). Under what circumstances violations lead to estimation problems, bad model fit, or implausible parameters still remains unclear. In this paper, we only discuss models where data are assumed to be continuous. If ordinal data are used, polychoric correlations should be used to arrive at consistent estimates (Alwin, 2007; Jagodzinski & Kuhnel, 1987), or a Latent-Markov model should be used (Cernat, 2013; van de Pol & Langeheine, 1990).

1. *Independence of observations over time*

Jagodzinski, Kuhnel & Schmidt (1987) believe that if the time between measures is short, there might be memory effects, and stability parameters (or reliabilities) are then overestimated. If memory effects are to be accounted for, they are typically included as correlated errors or effects between the measurement errors over time. This, in turn, relaxes the assumption of independence of observations over time.

A quasi-simplex model with correlated errors is normally not identified. Wiley & Wiley (1974) first showed that such a model can be fitted with additional assumptions for three waves of data, and Palmquist & Green (1992) formalised the situations under which these models are identified. They recommend the use of at least 4 waves of data, if errors are to be correlated in the quasi-simplex model. In the present paper this assumption will be tested by including lag-1 correlated errors.

2. *The mean Y_t and η_t are 0*

Rogosa (1985) and Rogosa & Willett (1985) have criticized longitudinal models with a simplex structure that aim to study change over time. Because the QSM focuses on the covariances between variables, and ignores changes in means, any average development over time remains hidden. Data that resemble overall growth among sample members will fit the quasi-simplex model just as well as a model that assumes there is no growth over time. Mandys, Dolan & Molenaar (1994) argue that means should therefore be investigated and modelled if one suspects a process of growth or decline in the sample as a whole. Mean structures are nowadays easily modeled in SEM software packages. Subsequently, means can be constrained over time (Hamaker, Kuiper, & Grasman, *in press*) to evaluate whether the quasi-simplex model with unconstrained means should be abandoned in favour of a model which explicitly constrains means to be equal over time.

The fact that it is uncommon in the QSM literature to model means is probably due to historical reasons more than anything else. Although for methodologically motivated research this is not necessary, substantive evaluation of change over time is improved with this information. In this paper we constrain the intercept of the observed items to be equal over time in order to test this assumption.

3. The change process is lag-1

Methodologists often criticize the QSM assumption that any true score is only determined by the true score at the previous measurement occasion. For example, Rogosa (1985) notes that the lag-1 assumption is often too easily made in the social sciences. Coenders et al. (1999) show that if the lag-1 assumption does not hold, the reliability coefficient will be severely biased. Using life satisfaction as an example, they argue that a lag-2 process is more likely to occur in addition to a lag-1 process than, for example, memory effects (Coenders et al., 1999). A lag-2 (or more) effect is possible in case of a temporary change in the situation of a group of individuals, such as the impact of a temporary economic downturn on employment and income variables. Testing the autoregressive assumption of the quasi-simplex model is straightforward when more than three waves of data are used. In this paper, the lag-2 assumption is tested by adding the three extra parameters to the QSM.

4. Equivalence of $V(\varepsilon_t)$ or reliabilities over time

The assumption that measurement errors are equal over time may be untenable under several situations. Under the first assumption presented we discussed how memory effects may lead to correlated measurement errors, but the size of the measurement errors under that scenario are still assumed to be equal.

The size of errors can however also change with time due to the measurement process itself. Repeated measurements may lead to attitude or behaviour changes in respondents, or panel respondents may simply learn how to complete surveys in a consistent way (Sturgis, Allum, & Brunton-Smith, 2009; Uhrig, 2012). Under both processes measurement errors may decrease and reliabilities increase over time. Alternatively, the size of measurement errors may change over time if the population of interest is undergoing a period of change. For example, students' attitudes towards studying may crystallize over the course of university, leading to lower measurement errors at later waves (Lugtig, Boeije, & Lensvelt-Mulders, 2012).

Fortunately, the assumption of equal error variances can be easily relaxed when data from more than three waves are used. As long as two variables, of which one if the first, and the other the last, are constrained, the other error variances can be freely estimated (Werts, Jöreskog, & Linn, 1971). In a similar way, the assumption of equal reliabilities can also be relaxed. The two assumptions lead to slightly different results in case the true score variances change over time (Alwin, 2007). In this paper we will constrain the error variances to be equal in waves one and five while the rest are freely estimated.

5. The variances of the error term $V(\varepsilon_t)$ and disturbance $V(\zeta_t)$ follow a normal distribution with a mean of 0

Variances in Structural Equation Models that are estimated using Maximum Likelihood algorithms are commonly assumed to follow a normal distribution. That implies for quasi-simplex models that the size of measurement errors in the positive direction (resulting in an overestimate of the true score) are equal to those in the negative direction (resulting in underestimating the true score). The size of the variance of measurement errors determines the reliability. Because of sampling variance in both the estimate of the true score variance and the observed variance, the sampling variance of a reliability estimate can be very large, especially when stability is low between waves (Coenders et al., 1999; Palmquist & Green, 1992). Additionally, it is sometimes not reasonable to assume that measurement errors are normally distributed. This is especially the case when measurement errors are small and approach the boundary estimate of 0. Such boundary estimates easily lead to non-convergence. Sometimes, problems of nonnormal variance distribution can be solved by transforming the observed scores. The distribution of the variance term does not necessarily follow the same distribution as the variance of the true score, so transforming the observed scores does not always solve this problem. A further way to relax this assumption is by using Bayesian estimation (Kaplan & Depaoli, 2012). In this paper, we will use Bayesian estimation to estimate a traditional QSM.

6. Covariances between ε_t, η_t and ζ_t are zero

Coefficients for the true scores (η_t) and disturbances (ζ_t) are directly linked. The disturbance indicates the unexplained variance in the true score at time *t* and, as such, they cannot be correlated. True scores (η_t) and disturbances may, however, be correlated to measurement errors (ε_t) in specific research settings. For example, research on attitude formation and attitude strength has shown that specific subgroups within populations have "strong", "weak", or "non-" attitudes (Converse, 1964; Judd & Milburn, 1980; Zaller, 1992). "Weak" or "non-" attitudes in quasi-simplex models appear either as a low stability coefficient between waves, or a low reliability coefficient. People with strong attitudes tend to be more consistent responders thus the chance of random error is lower amongst this subgroup (see Prislin, 1996). In such a situation, the error (ε_t) could be related to the true score (η_t). Similarly, studies on income data have found that the amount of measurement error is higher for those with higher incomes (Bound, Brown, & Mathiowetz, 2001). The assumption of zero covariances between ε_t, η_t and ζ_t cannot be relaxed easily. When one has validation data about the variable of interest at multiple time points the association between the two can be investigated. We are not aware of any study that has done this in the context of the QSM. Given that no such validation data exists for our data, we will not test this assumption in this paper.

We must make a distinction between two different types of assumptions within the six discussed above. The first type includes assumptions that are specific to the QSM and have been used in practice without previous testing of their plausibility. The first four assumptions fall in this category. A second type of assumption is a more general one which is specific to all Structural Equation Models that use Maximum Likelihood estimation. Thus, assumptions five and six apply both to the multiple items approach of testing reliability, such as CFA, and to the multiple measures approach examined here.

OUR STUDY

Our study examines a set of diverse variables taken from British Household Panel Survey data. Because we study variables that span different substantive concepts, we expect different violations of the assumptions for each of the variables. For some variables we may expect correlated measurement errors (when the same interviewer records his subjective feelings about the same respondent over time), for some other variables we may expect a lag-2 effect (job hours that may be lower or higher than usual at a particular wave due to special circumstances), while for others we assume stability in sample means over time (subjective health status).

First, a traditional QSM will be estimated using the six strict assumptions presented above. Then, for each variable, we will investigate to what extent relaxing some of the assumptions will improve the model fit and affects the substantive results of the quasi-simplex. Our main parameters of interest are the stability and reliability coefficients.

We will use the five-wave QSM for the different types of variables. For each variable we have chosen two five-wave periods of the BHPS in order to take into account factors such as attrition and panel conditioning. So, for each variable, we use data from waves 1-5 and waves 11-15 to test the QSM assumptions.

DATA AND ANALYTICAL APPROACH

Data

We examine data from the British Household Panel Survey. The BHPS is an interviewer-administered panel survey of the UK population that started in 1991 with an address-based sample of 5,500 households. All household members aged 16 and older are interviewed annually and followed as long as they remain resident in the UK. The BHPS is a general purpose panel survey covering such topics as household composition, housing conditions, work, health, income, spending and socio-economic attitudes.

In this paper, we use data only from BHPS waves 1-5 and waves 11-15[2]. Earlier studies about the assumptions of the quasi-simplex models have recommended using at least four waves of data (Palmquist & Green, 1992; Werts et al., 1971), but have often used five waves as well. It would be possible to estimate the model with more than five waves, but with every wave that is added to the model, it is more likely that some assumptions of the QSM are violated. Analyses are of unweighted data, as it is not our goal to generalise our findings to the UK population. We dealt with item and unit missing data using the default FIML-estimator in MPLUS (Muthén & Muthén, 2013).

Instruments

We test the assumptions of the quasi-simplex for eleven variables, which represent both facts and attitudes. Facts have been generally found to be more reliably measured than attitudes, and this may affect how the QSM assumptions are met (Alwin, 2007).

1) *Labour income:* This variable is derived from survey responses concerning: 1) employment earnings and pay periods, and 2) profit and loss from self employment. The derivation yields monthly total income, regardless of pay period or self employment earnings

[2] We have restricted the sample only to the original sample members for waves 11-15 in order to avoid confounding with other effects possible with refreshment and booster samples.

statement period (Taylor, Brice, Buck, & Prentice-Lane, 2010). The variable has been transformed using the log in order to normalize it.

2) *Job hours:* A continuous measure of the regular weekly work hours amongst employees. The questions reads: “Thinking about your (main) job, how many hours, excluding overtime and meal breaks, are you expected to work in a normal week?”

3) *Minutes traveling to work:* A continuous measure of the minutes employed respondents travel to their job: “About how much time does it usually take for you to get to work each day, door to door?”

4) *General job satisfaction:* A categorical evaluation of a respondent’s job satisfaction: “ All things considered, how satisfied or dissatisfied are you with your present job overall?” with a response scale using a labeled midpoint and endpoints: “1 - not satisfied at all”,“2”, “3”, “4 - neither satisfied, nor dissatisfied”, “5”, “6”,”7 - completely satisfied”.

Aspects of job satisfaction. The next set of questions (items 5-8) asks respondents about their satisfaction with several aspects of their job. “I'm going to read out a list of various aspects of jobs, and after each one I'd like you to tell me from this card which number best describes how satisfied or dissatisfied you are with that particular aspect of your own present job.” Each aspect is evaluated using the same response scale as the question for general job satisfaction (see above).

5) Satisfaction with wages: “The total pay, including any overtime or bonuses”.

6) *Satisfaction with job security*: “Your job security”.

7) *Satisfaction with actual work:* “The actual work itself”.

8) *Satisfaction with work hours*: “The hours you work”.

9) *Subjective financial situation:* This is the respondent's self-evaluated financial situation: “How well would you say you yourself are managing financially these days? Would you say you are [Interviewer reads out answer categories]” with answer categories: “1 - living comfortably”, “2 - doing alright”, “3 - just about getting by”, “4 - finding it quite difficult” and “5 - finding it very difficult”.

10) *Subjective health status:* This question asks respondents to evaluate their own subjective health against other people of the same age.: “Please think back over the last 12 months about

how your health has been. Compared to people of your own age, would you say that your health has on the whole been [Interviewer reads out answer categories]" with answer categories "1 - excellent", "2 - good", "3- fair", "4- poor", and "5 - very poor".

11) *Respondent Cooperation:* This is the interviewer evaluated respondent cooperation. "In general, the respondent's cooperation during the interview was…" with answer categories "very good" "good" "fair" "poor" and "very poor".

Analytical approach

To test QSM assumptions, we estimated six models for each of these eleven variables. We relied on the Bayesian Information Criterion (BIC) to evaluate which models best fit the data. This goodness of fit indicator takes into account both overall fit and model complexity and can be used even when models are not nested. After selecting the best fitting models we compare the estimated reliabilities and stabilities of those models against the baseline QSM to see if freeing these assumptions changes estimates of data quality and stability.

The six models tested are:

- ***Model 1 - The baseline QSM***. This model includes all the assumptions usually made when QSMs are estimated.

- ***Model 2 - Correlated errors.*** This models adds four lag-1 correlations between random errors to the baseline model. They are freely estimated.

- ***Model 3 - Equal means in time.*** This model adds the means to the baseline model by estimating the intercept of the observed scores. We assume the intercepts to be equal over time.

- ***Model 4 - Lag-2 of true scores.*** We relax the assumption of solely a lag-1 relationship between the trues scores by adding three lag-2 effects to the baseline model.

- ***Model 5 - Unequal variances in time.*** We relax the assumption of equal variances in time by constraining the variance of the measurement errors to be equal only at waves one and five. The other measurement error variances are freely estimated.

- ***Model 6 - Baseline model with Bayesian estimation.*** We use *Model 1* but change the estimation method from ML to Bayesian with non-informative priors in order to free the assumption that the disturbance and measurement error terms are normally distributed. In this case we used 4 chains with a thinning coefficient of 5, a convergence criterion of 0.01, and a minimum number of iterations of 5000.

Estimation problems

Estimation problems of QSM mentioned in other studies were also found during our analyses (Cernat, 2013; Coenders et al., 1999; Hargens et al., 1976; Jagodzinski & Kuhnel, 1987). For each of the problems, we have tried to resolve the issues by 1) outlier removal, 2) transforming the variables or 3) using Bayesian estimation where this is not explicitly done to test normality assumptions in the QSM. Three problems stand out.

First, we find that some models fail to converge. This is especially the case for *Model 5*; the model where we allow the measurement error variances to be unequal over time. To overcome this problem we have used Bayesian estimation, often with a more liberal convergence criterion (see ***Table 1*** for details).

Secondly, we find that *Model 4* - the model with lag-2 parameters - produces inconsistent estimates. The standardized stability parameters are higher than 1 in the models for 'job hours worked', 'interviewer rating of respondent cooperation' and 'minutes travelled to work'. We have not been able to resolve this issue, and so deemed these models 'failed to converge'.

The third issue we encountered was for one variable in *Model 2* - the model with correlated errors. The interviewer-rated 'respondent cooperation' produced in this case a negative variance for the true score at wave four (unstandardised coefficient of -.01). We have subsequently constrained this parameter to be .01 and proceed to interpret the other model parameters of this model with caution.

Our results are structured as followed. For all models we compare the BIC coefficient to evaluate the relative model fit of each model. Then, we compare the parameter estimates for the best fitting models out of the six models we estimate, to evaluate whether any relaxation of the assumptions of the QSM affects our substantive estimates on the stability and reliability coefficients. All models were estimated using MPLUS 7.11 (Muthén & Muthén, 2013).

RESULTS

Table 1 shows BIC values after running the six versions of the QSM on the 11 variables and two time periods. The BIC values shown in bold represent the best models in terms of model fit. Despite the fact that the baseline QSM has rather strict assumptions, we find that for seven out of the 22 situations this model is the best fitting model. *Model 3*, which has even stricter assumptions than *Model 1*, is the best model for eleven variables while *Model 2* - the model with correlated errors - is the best for the remaining four variables. This implies that for only four out of 22 situations, we conclude that the strict assumptions of the QSM do not hold, and should be relaxed. *Models 4, 5* and *6* never produce the best model fit.

The four variables for which we find that the strict assumptions of the QSM should be relaxed to include correlated errors are ‘Respondent cooperation’ at both waves 1-5 and 11-15, and ‘minutes traveling to work’ and the ‘subjective financial situation’ of the respondent at waves 11-15. In the case of ‘Respondent cooperation’, we can find a reasonable post-hoc reason for our finding. Typically, some, but not all respondents are interviewed by the same interviewer over time. Respondents interviewed by the same interviewer are more likely to have highly consistent ratings over time, and therefore, this shows up as a correlated error in the model. For the other two variables that have correlated errors in wave 11-15 the reason is less obvious. However, if we look at the parameter estimates of the correlations for these variables in ***Table 2*** it becomes clear that many correlations over time are quite small. Even for the variable respondent cooperation we find that the correlated errors are mostly smaller than 0.1, apart from the correlated error between wave 4 and 5, this being 0.32. The only variable for which correlations are substantial is for the variables ‘minutes traveling to work’. This could be due to respondents consistently over- or underreporting their travel duration in two subsequent waves, while at the same time, not doing so over all five waves.

Table 1: Values for model fit (Bayesian Information Criterion) for 6 versions of the quasi-simplex model.

	Variable	Sample size All models	Model 1 Baseline QSM	Model 2 Correlated errors	Model 3 Equal means	Model 4 Lag-2 parameter	Model 5 Unequal error variances (Bayesian)	Model 6 Bayesian estimation
Waves 1-5	Labour income	8,702	**64973**	64987	65213	64981	65087	64973
	Hours worked	7,852	176883	176888	**176876**	No con	177240	176893
	Minutes traveling to work	7,472	191798	191813	**191767**	No con	197200***	197235
	General job satisfaction	7,580	**79435**	79454	79469	79459	79465	79437
	Satisfaction with wages	7,572	**89056**	89079	89107	89077	89106	89063
	Satisfaction with job security	7,516	88290	88312	**88276**	88317	88463	88291
	Satisfaction with actual work	7,578	**79989**	80010	80025	80008	80122	79990
	Satisfaction with work hours	7,578	**84833**	84864	84856	84859	84863	84833
	Subjective financial situation	12,466	**119679**	119697	119743	119707	119707	119680
	Subjective health	12,863	**111456**	111472	111593	111480	111516	**111456**
	Respondent cooperation	12802	57689	**57603***	57943	No con	57717	57689
Waves 11-15	Labour income	4,840	51073	51066	**51046**	51070	51070	51076
	Hours worked	4,485	116445	116477	**116424**	No con	116594	116456
	Minutes traveling to work	4156	131329	**131323**	131305	No con	131762	131332
	General job satisfaction	4,332	52609	52633	**52590**	52631	53318	52616
	Satisfaction with wages	4,325	56101	56128	**56079**	56119	56156	56102
	Satisfaction with job security	4,318	55464	55487	**55448**	55484	55762	55464
	Satisfaction with actual work	4,330	52961	52979	**52946**	52972	52978***	52975
	Satisfaction with work hours	4,330	55285	55316	**55260**	55310	55374	55285
	Subjective financial situation	7,187	71859	**71852**	71862	71869	72755	71865
	Subjective health	7,386	75057	75058	**75035**	75083	75304***	75058
	Respondent cooperation	7,289	27820	**27770****	27823	No con	28052***	27821
	# times Best model		7	4	11	0	0	1

Notes: * Initial estimate of variance of T4 is -.01. Model converges when variance of T4 is subsequently constrained to .01. ** Initial estimate of variance of T3 is -.02. Model converges when variance of T3 is subsequently constrained to .01. *** Converges with Bayesian convergence criterion of .05, instead of .01. No con=failed to converge. Bold entry=lowest BIC.

Table 2: correlated measurement errors for variables where model with correlated errors fit the data best.

Coefficient	**Respondent cooperation**		**Subjective financial situation**	**Minutes traveling to work**
Wave	**1-5**	**11-15**	**11-15**	**11-15**
x1 <-> x2	0.001	0.346	0.142	0.344
x2 <-> x3	0.072	0.041	0.099	0.231
x3 <-> x4	0.045	0.109	0.102	0.112
x4 <-> x5	0.318	0.126	0.142	0.159

Note: for sample sizes, see ***Table 1.***

Apart from looking at the fit of each model, the parameter estimates themselves are the second heuristic we use to assess the assumptions of the QSM. ***Table 3*** shows the mean reliability and stability for the baseline QSM and estimates for the best fitting model, as long as that is not the baseline QSM, for each variable. Overall we observe the expected levels of reliability and stability for facts and attitudes (Alwin, 2007; Saris & Gallhofer, 2007). In the baseline QSM model, the three variables asking about facts have reliabilities between 0.81 (log of 'labour income' waves 11-15) and 0.93 ('hours worked' waves 1-5). The attitudinal variables have much lower reliabilities. Here, the lowest reliability is found for general job satisfaction in waves 11-15 (0.51), and the highest for subjective health in waves 11-15 (0.68). Overall, the average estimate across all variables for the reliability coefficient is somewhat higher in waves 1-5 (0.69) than in waves 11-15 (0.66).

The stability for all variables is relatively high. The lowest average stability parameter is 0.61 for satisfaction with job security in waves 1-5, and the highest stability is found for subjective health in waves 11-15 (0.88). Where the reliability was higher in waves 1-5 as compared to waves 11-15, we now find the opposite effect for stabilities. The average stability across all variables is 0.74 for waves 1-5 and 0.78 for waves 11-15.

When we compare the parameter estimates that were obtained using the baseline QSM to the model that fits best for each variable two things stand out. Firstly, we find negligible differences between the estimates of *Models 1* and *3*. This is to be expected as the models only differ in the means, not in the covariances. Secondly, we find that when the *Model 2* - QSM with correlated errors - fits the data best, parameters estimates do differ. Adding correlated errors results in lower estimates for the reliability. The changes range from a minimum of 0.04 for 'Respondent cooperation' in waves 11-15 to a maximum of 0.11 for 'Respondent cooperation' in waves 1-5. While reliabilities always decrease (i.e. are over-estimated if errors are assumed uncorrelated), the stabilities increase in these models (i.e. are underestimated if errors are assumed uncorrelated). Here the minimum increase is 0.03 for respondent cooperation in

Table 3: mean reliability and stability parameter for the baseline QSM, best fitting model, and difference in parameter estimates of the two models.

			Model 1 - baseline QSM		Best model		Difference	
Variables	**Wave**	**Best model fit**	**Mean reliability**	**Mean stability**	**Mean reliability**	**Mean stability**	**Mean reliability**	**Mean stability**
Labour income	1-5	Baseline QSM	0.92	0.82	-	-	-	-
Hours worked	1-5	Equal means	0.93	0.84	0.93	0.83	0.002	-0.005
Minutes traveling to work	1-5	Equal means	0.83	0.74	0.83	0.74	0.001	-0.001
General job satisfaction	1-5	Baseline QSM	0.61	0.66	-	-	-	-
Satisfaction with wages	1-5	Baseline QSM	0.65	0.68	-	-	-	-
Satisfaction with job security	1-5	Equal means	0.66	0.61	0.66	0.61	0.002	-0.002
Satisfaction with actual work	1-5	Baseline QSM	0.60	0.70	-	-	-	-
Satisfaction with work hours	1-5	Baseline QSM	0.60	0.70	-	-	-	-
Subjective financial situation	1-5	Baseline QSM	0.68	0.81	-	-	-	-
Subjective health	1-5	Baseline QSM	0.67	0.84	-	-	-	-
Respondent cooperation	1-5	Correlated errors	0.52	0.66	0.41	0.71	-0.106	0.047
Labour income	11-15	Correlated errors	0.81	0.82	0.81	0.82	0.000	0.000
Hours worked	11-15	Equal means	0.93	0.88	0.93	0.88	0.000	0.000
Minutes traveling to work	11-15	Correlated errors	0.9	0.79	0.84	0.83	-0.060	0.040
General job satisfaction	11-15	Equal means	0.51	0.72	0.51	0.72	0.000	0.000
Satisfaction with wages	11-15	Equal means	0.59	0.76	0.59	0.76	0.000	0.000
Satisfaction with job security	11-15	Equal means	0.55	0.75	0.55	0.75	0.000	0.000
Satisfaction with actual work	11-15	Equal means	0.54	0.74	0.54	0.74	0.000	0.000
Satisfaction with work hours	11-15	Equal means	0.56	0.77	0.56	0.77	0.000	0.000
Subjective financial situation	11-15	Correlated errors	0.66	0.85	0.59	0.91	-0.070	0.060
Subjective health	11-15	Equal means	0.68	0.88	0.68	0.88	0.000	0.000
Respondent cooperation	11-15	Correlated errors	0.58	0.83	0.54	0.86	-0.040	0.030

Note: for sample sizes, see ***Table 1***.

waves 11-15 to 0.06 for subjective financial situation in waves 11-15. Thus we observe that increases in the reliability are mirrored by a decrease in stability that is about equal in size.

CONCLUSIONS AND DISCUSSION

This paper showed how to relax and assess five of the most important assumptions of the quasi-simplex model. We find that freeing the assumptions of the QSM does not improve model fit for most of our variables. For about half the variables, we find that the QSM can actually be more restricted by adding an equality constraint on the means of the variables over time. In addition, we see that relaxing the assumptions by adding a lag-2 parameters to the true scores (*Model 4*), or allowing unequal measurement error variances (*Model 5*) never leads to a better model fit. This implies that for the variables we tested, we can conclude that these crucial assumptions of the quasi-simplex model hold.

Using Bayesian estimation (*Model 6*) instead of maximum Likelihood does not lead to a better model either. However, we do find that Bayesian estimation can be instrumental to test some of the assumptions of the quasi-simplex model, as we found the model with unequal error variances converged with Bayesian estimation even when most of the ML models had problems. The BIC values of *Model 1* and *Model 6* are almost equivalent and any difference is probably caused by the fact that Bayesian estimation approximates the maximized value of the Loglikelihood. In terms of parameter estimates, closer inspection of the results of *Model 6* show that for almost all our variables the variances in our model do follow a normal distribution. Only when either the reliability or stability estimate approximates 1 we find that the posterior distribution of the measurement error (ε_t) and disturbances (ζ_t) are skewed. Even for those variables however, we find no differences in stability and reliability coefficients.

These findings have to be interpreted with some caution. For four out of 22 situations, including correlated errors (*Model 2*) leads to a more appropriate model than the baseline QSM. In our study, this is the case for interviewer ratings, subjective financial situation and minutes traveled to work. When correlated measurement errors are included in the model, reliabilities decrease and stabilities increase. This is likely due to the fact that the model allows for a more flexible estimation of the error variance (ε_t). For that reason, error variances increase, while the disturbances of the true scores decrease. In other words, when correlated measurement errors are present in the data and allowed in the model, the estimates of measurement errors are no longer biased negatively, and reliabilities decrease. Adding correlated measurement errors does not only affect the interpretation of measurement errors but also affects the stability and reliability parameters substantially. This implies that when four or more waves of data are available correlated measurement errors should be added to the model to test whether this improves the model and/or affects the parameters of interest.

We find small differences in the stability and reliability parameters depending on whether we use data from waves 1-5 or waves 11-15. Reliabilities are higher when data from waves 1-5 are used while stabilities are higher for waves 11-15. The reasons for this may be related to attrition and panel conditioning. When attrition is related to undergoing change, the stability coefficients of the people that are continuing sample members will become higher. However, this does not explain why the reliabilities of the variables should become lower at later waves.

Although earlier studies have reported that the QSM often fails to converge, the baseline quasi-simplex model converges and provides credible parameter estimates for all our variables. Nevertheless some of the other models have shown that the QSM still presents convergence issues that have been reported in the literature previously (Cernat, 2013; Coenders et al., 1999; Hargens et al., 1976; Jagodzinski & Kuhnel, 1987). We still know relatively little about the causes of these convergence problems. Other models such as the Latent State-Trait Model (Kenny & Zautra, 2001) or MTMM models (Scherpenzeel, 1995) are known to have convergence problems too, and all three models bear some similarities in terms of model complexity and model assumptions. We have seen that Bayesian estimation may prove to be a solution for some of the issues but more research is needed to understand why Maximum Likelihood estimation results in convergence problems and why or when the Bayesian estimation performs better. For this, a more formal simulation study is necessary.

A limitation of this study is that we used only 11 variables across two time windows that were all measured in British Household Panel Survey. Other variables may need some of the model modifications we examined here. For example, theoretically, one may expect a lag-2 parameter between true scores when a respondent's situation has temporarily changed at the time of the interview. If one suspects this to be the case, this article provides an overview of how to relax and test for this, and other assumptions of the quasi-simplex model.

REFERENCES:

Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality and Quantity*, *23*(3-4), 277–331.

Alwin, D. F. (2007). *The margins of error: a study of reliability in survey measurement*. Wiley-Blackwell.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement the influence of question and respondent attributes. *Sociological Methods & Research*, *20*(1), 139–181.

Bast, J., & Reitsma, P. (1997). Mathew effects in reading: a comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research*, *32*(2), 135–167. doi:10.1207/s15327906mbr3202_3

Blok, H., & Saris, W. E. (1983). Using longitudinal data to estimate reliability. *Applied Psychological Measurement*, *7*(3), 295–301.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley-Interscience Publication.

Bound, J., Brown, C., & Mathiowetz, N. (2001). *Measurement error in survey data* (PSC Research Report No. 00-450) (pp. 3705–3843). Elsevier.

Cernat, A. (2013). The impact of mixing modes on reliability in longitudinal studies. *ISER Working Paper*, (09), 1–27.

Coenders, G., Saris, W., Batista-Foguet, J., & Andreenkova, A. (1999). Stability of three-wave simplex estimates of reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(2), 135–157.

Converse, P. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and Discontent* (1st ed.). The Free Press of Glencoe.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (n.d.). A critique of the cross-lagged panel model. *In Press*.

Hargens, L. L., Reskin, B. F., & Allison, P. D. (1976). Problems in estimating measurement error from panel data an example involving the measurement of scientific productivity. *Sociological Methods & Research*, *4*(4), 439–458.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, *34*(1), 93–101.

Jagodzinski, W., & Kuhnel, S. M. (1987). Estimation of reliability and stability in single-indicator multiple-wave models. *Sociological Methods & Research*, *15*(3), 219–258.

Jagodzinski, W., Kuhnel, S. M., & Schmidt, P. (1987). Is there a “socratic effect” in nonexperimental panel studies? Consistency of an attitude toward guestworkers. *Sociological Methods & Research*, *15*(3), 259–302.

Judd, C. M., & Milburn, M. A. (1980). The structure of attitude systems in the general public: comparisons of a structural equation model. *American Sociological Review*, *45*(4), 627.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 650–673). New York: Guilford Press.

Kenny, D., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. M. Collins & A. Sayer (Eds.), *New Methods for the Analysis of Change* (pp. 241–264). Washington, DC: American Psychological Association.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company, Inc.

Lugtig, P., Boeije, H. R., & Lensvelt-Mulders, G. J. L. M. (2012). Change? What change? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*(3), 115–123.

Mandys, F., Dolan, C. V., & Molenaar, P. C. M. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics*, *19*(3), 201–215.

Muthén, L., & Muthén, B. (2013). *Mplus user's guide. seventh edition* (Seventh Edition.). Los Angeles, CA: Muthén & Muthén.

Palmquist, B., & Green, D. P. (1992). Estimation of models with correlated measurement errors from panel data. *Sociological Methodology*, 119–146.

Prislin, R. (1996). Attitude stability and attitude strength: One is enough to make it stable. *European Journal of Social Psychology*, *26*(3), 447–477.

Rogosa, D. (1985). Myths and methods: “Myths about longitudinal research” plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, N.J: L. Erlbaum.

Rogosa, D., & Willett, J. B. (1985). Satisfying a simplex structure is simpler than it should be. *Journal of Educational Statistics*, *10*(2), 99–107.

Saris, W., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*(1), 29–43.

Saris, W., & Van Den Putte, B. (1988). True score or factor models: A secondary analysis of the ALLBUS-test-retest data. *Sociological Methods & Research*, *17*(2), 123–157.

Scherpenzeel, A. C. (1995). *A question of quality: evaluating survey questions by multi trait - multi method studies*. Doctoral Dissertation, Royal PTT, Amsterdam, Netherlands.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 113–126). Chichester: Wiley.

Taylor, M. F., Brice, J., Buck, N., & Prentice-Lane, E. (Eds.). (2010). *British Household Panel Survey user manual. Volume A: Introduction, technical report and appendices*. Colchester: University of Essex.

Uhrig, S. N. (2012). Understanding panel conditioning: an examination of social desirability bias in self-reported height and weight in panel surveys using experimental data. *Longitudinal and Life Course Studies*, *3*(1), 120 – 136.

Van de Pol, F., & Langeheine, R. (1990). Mixed markov latent class models. In C. C. Clogg (Ed.), *Sociological methodology* (Vol. 20, pp. 213–247). Oxford: Blackwell.

Werts, C. E., Jöreskog, K. G., & Linn, R. L. (1971). Comment on “The estimation of measurement error in panel data.” *American Sociological Review*, *36*(1), 110–113.

Wiley, D., & Wiley, J. (1970). The estimation of measurement error in panel data. *American Sociological Review*, *35*(1), 112–117.

Wiley, J., & Wiley, M. (1974). A note on correlated errors in repeated measurements. *Sociological Methods & Research*, *3*(2), 172–188.

Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.