7

Assessing and Relaxing Assumptions in Quasi-Simplex Models

Alexandru Cernat, The University of Manchester, Peter Lugtig, Utrecht University, Nicole Watson, University of Melbourne, and S.C. Noah Uhrig, Ministry of Justice of England and Wales

Introduction

Measurement is one of the most important and complex aspects of research in the social sciences, as the presence of systematic or random error can render analyses invalid or unreliable. This issue is also of high importance in longitudinal studies, this context making measurement error even more complex by biasing estimates of change in unknown ways (Hagenaars, 1990; Plewis, 1985). To tackle some of these issues in longitudinal studies a number of statistical models have been proposed, one of which is the quasi-simplex model (QSM, Alwin, 2007; Heise, 1969; Jöreskog, 1970; Wiley and Wiley, 1970). The QSM has a number of characteristics that make it attractive. For example, it can be used to estimate test–retest reliability (Lord and Novick, 1968), that is, the proportion of variance due to the true score as opposed to random error, which can be used both to estimate data quality and to correct for random error. The model also has the advantage that it can be used for stand-alone items, which do not form part of a scale, implying that it can be applied to longitudinal data without extra data-collection costs.

This model, or related models, has been used in a number of different contexts. For example, the model has been used in developmental studies and compared to other longitudinal models, such as the latent growth curve model (Bast and Reitsma, 1997). Uhrig and Watson (2020) have applied it in order to correct for random error in wage decomposition models, thus showing the impact of measurement error on substantive analyses. Finally, the QSM has been used extensively in survey methodology to estimate the data quality of different types of questions (Alwin, 1989, 2007; Alwin and Krosnick, 1991; Saris and Van Den Putte, 1988) or to compare different survey designs (e.g. Cernat, 2014).

Although the utility of QSM to survey methodologists and users of longitudinal data is undisputed, the method also has a number of limitations. For example, parameter estimates are sometimes implausible and standard errors may be large (Alwin, 2007; Wiley and Wiley, 1970) or the model may fail to converge

Alexandru Cernat, Peter Lugtig, Nicole Watson, and S.C. Noah Uhrig, *Assessing and Relaxing Assumptions in Quasi-Simplex Models* In: *Measurement Error in Longitudinal Data*. Edited by: Alexandru Cernat and Joseph W. Sakshaug, Oxford University Press (2021). © Oxford University Press. DOI: 10.1093/oso/9780198859987.003.0007

altogether (Cernat, 2014; Coenders, et al., 1999; Hargens, Reskin, and Allison, 1976; Jagodzinski and Kuhnel, 1987). While a number of past studies have evaluated some properties of the QSM such as the appropriate time between two waves (Jagodzinski and Kuhnel, 1987), how ordinal data should be modelled (Alwin, 2007), and how means should be incorporated into the model (Mandys, Dolan, and Molenaar, 1994), no systematic presentation of the model assumptions has been made. Also, it is unclear at this moment how some of these assumptions could be freed in practice in order to make the model more plausible.

This chapter will fill this gap in the literature by presenting the main assumptions of the quasi-simplex model. It will also show how some assumptions can be relaxed when more than three waves of data are present. The chapter will also show empirical results from eleven items measured at two different time periods in the British Household Panel Survey. These examples are used to illustrate how assumptions of the QSM can be relaxed, leading to a better model.

7.1 The Quasi-Simplex Model

The basic quasi-simplex model, as shown in Figure 7.1, 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 and Novick, 1968), the observed score at each time point consists of the true score and measurement error (ε):

$$Y_t = \eta_t + \varepsilon_t. \tag{1}$$

The QSM uses repeated observations of the same variable to separate η from ε. Subsequent measurements are linked only by a stability coefficient between two true scores at times t and t−1 ($\beta_{t,t-1}$), and a random disturbance term (ζ_t) that represents the time-specific true score residual (or noise).[1] The true score

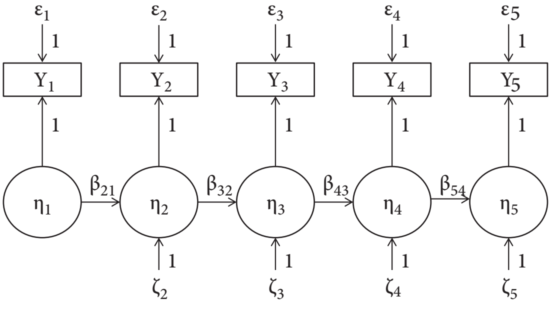

Figure 7.1 Quasi-simplex model with five measurements.

[1] To identify the true score at wave 1 this score is equivalent to the disturbance term, e.g. $\eta_1 = \zeta_1$.

stability coefficients mix between- and within-person stability, which can further be separated by imposing a random intercept (Hamaker, Kuiper and Grasman, 2015) or fixed-effects structure (Allison, Williams, and Mora-Benito, 2017) on the true scores. For now, we limit ourselves to discussing the classic stability model that is sufficient for estimating reliability coefficients:

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

The QSM is only empirically testable when several assumptions about the relations between the estimated parameters ε_t, η_t, and ζ_t are made.

7.1.1 Independence of Observations over Time

This implies that both the measurement error (ε_{t}) and disturbance (ζ_{t}) are uncorrelated over time. Jagodzinski, Kuhnel, and 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 correlations or effects between the measurement errors over time (Palmquist and Green 1992).

7.1.2 The Means of 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 QSM, as the model with centred variables is equivalent to the model with estimated means as long as no constraints are imposed on the means (see Blok and Saris, 1983), and so means can actually be ignored if one is not interested in them. Rogosa (1985) and Rogosa and Willett (1985) have criticized longitudinal models with a simplex structure that fail to include means, as any average development over time then remains hidden.

7.1.3 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. This assumption is called the lag-1 assumption. Rogosa (1985), for example, 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 likely to occur in addition to a lag-1 process when, for example, memory effects occur. A lag-2 (or more) process can further occur 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.

7.1.4 Equal $V(\varepsilon_t)$ or Reliabilities over Time

It is necessary to constrain some of the QSM parameters to be equal over time for the model to be identified. Heise (1969) favoured the idea that the reliabilities should be equal over time. As the reliabilities are calculated as the ratio between the true score variance ($V(\eta_t)$) and the observed variance ($V(Y_t)$), this means that the ratio between the two remains stable. Alternatively, Wiley and Wiley (1970) believed that the error variances $V(\varepsilon_t)$ should be constrained to be equal at every time t. When the variances of the observed scores 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 $V(\varepsilon_t)$. Conversely, the Wiley and Wiley specification will lead to slight differences in the reliability estimate over time.

The assumption that measurement errors or reliabilities are equal over time may be untenable under several situations. For example, the size of errors can 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, and 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, and 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 the error variances of the first and last measurement are constrained to ensure that the model is identified, the other error variances can be freely estimated (Werts, Jöreskog, and Linn, 1971). In a similar way, the assumption of equal reliabilities can also be relaxed.

7.1.5 Other Assumptions

Two other general structural equation modelling (SEM) assumptions have an impact on the QSM estimates. While we mention them here we do not show how to explicitly evaluate them in this chapter, as these are assumptions that are more generally made in SEM.

The first such assumption is that the errors ε_t and disturbances ζ_t are distributed normally with a mean of 0. Errors in SEM that are estimated using maximum likelihood algorithms are commonly assumed to follow a normal distribution. In the case of QSM this implies that the size of measurement errors in the positive direction (resulting in an overestimate of the observed score as compared to the true score) is equal to those in the negative direction. The distribution of the variance of the observed score does not necessarily follow the same distribution as the variance of the true score or measurement error, so transforming the observed scores does not resolve this if problems are encountered in the estimation of a variance term for the measurement error, for example when it approaches the boundary estimate of 0. One way to relax the assumption of normally distributed variances is to use Bayesian estimation (Cernat, 2014; Kaplan and Depaoli, 2012).

The second general assumption of SEM that applies to the QSM is that the covariances between the random error (ε_t), true score (η_t), and disturbance (ζ_t) are zero. The disturbance indicates the unexplained variance in the true score at time t and, as such, it should not be correlated to any other parameter in the model. True scores (η_t) and disturbances may, however, be correlated to measurement errors (ε_t) in specific research settings. For example, studies on income data have found that the amount of measurement error is higher for those with higher incomes (Bound, Brown, and 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, however, not aware of any study that has done this in the context of the QSM.

As shown above, all the assumptions of the QSM 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 (Coenders et al., 1999; Jagodzinski and Kuhnel, 1987). In many cases, the QSM will either not converge or will not fit the data, and/or the parameter estimates of the model will become biased because of violations of model assumptions (Hargens et al., 1976). This chapter shows how to assess and relax the assumptions of the QSM as long as data are assumed to be continuous.

7.2 Our Study

Our study examines a set of diverse variables taken from the British Household Panel Survey. 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 may wrongly assume stability in sample means over time (subjective health status).

First, a baseline QSM will be estimated using the five types of assumptions presented above. Then, for each variable, we will investigate to what extent relaxing some of the assumptions will improve the model fit and how this 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.

7.3 Data and Analytical Approach

7.3.1 Data

We examine data from the British Household Panel Survey (BHPS). 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 chapter, 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 and 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 in this example to generalize our findings to the UK population. We dealt with item and unit missing data using the default FIML-estimator in Mplus 7.11 (Muthén and Muthén, 2013).[3]

7.3.2 Instruments

We test the assumptions of the quasi-simplex model for eleven variables, which represent both facts and attitudes. Facts have been generally found to be more

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

[3] For syntax see the online appendix on the companion website at http://www.oup.co.uk/companion/LongitudinalData.

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 statement period (Taylor et al., 2010). The variable has been transformed using the log in order to normalize it.
2. **Hours worked**. 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 respondents take to 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?' [Interviewer provides response scale using a labelled 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] '1—living comfortably', '2—doing alright', '3—just about getting by', '4—finding it quite difficult', or '5—finding it very difficult?'
10. **Subjective health status**. This question asks respondents to evaluate their own subjective health against that of 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] '1—excellent', '2—good', '3—fair', '4—poor', or '5—very poor?'

11. **Respondent Cooperation**. This is the interviewer-evaluated respondent cooperation. 'In general, the respondent's cooperation during the interview was:' [answer categories] 'very good', 'good', 'fair', 'poor', or 'very poor'.

7.3.3 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**. Includes all the assumptions usually made when QSMs are estimated, using the Wiley and Wiley (1970) constraints on error variances.

Model 2: **Correlated errors**. Adds four lag-1 correlations between random errors to the baseline model. These are freely estimated.

Model 3: **Equal means in time**. 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 true scores by adding three lag-2 effects to the baseline model.

Model 5: **Unequal error 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 four chains with a thinning coefficient of 5, a convergence criterion of 0.01, and a minimum number of iterations of 5,000.

7.3.4 Estimation Problems

Estimation problems of QSM mentioned in other studies were also found during our analyses (Cernat, 2014; Coenders et al., 1999; Hargens et al., 1976; Jagodzinski and Kuhnel, 1987). For each of the problems, we have tried to resolve the issues

Table 7.1 Values for model fit (Bayesian information criterion) for six 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	**64,973**	64,987	65,213	64,981	65,087	64,973
	Hours worked	7,852	176,883	176,888	**176,876**	*No con*	177,240	176,893
	Minutes travelling to work	7,472	191,798	191,813	**191,767**	*No con*	197,200***	197,235
	General job satisfaction	7,580	**79,435**	79,454	79,469	79,459	79,465	79,437
	Satisfaction with wages	7,572	**89,056**	89,079	89,107	89,077	89,106	89,063
	Satisfaction with job security	7,516	88,290	88,312	**88,276**	88,317	88,463	88,291
	Satisfaction with actual work	7,578	**79,989**	80,010	80,025	80,008	80,122	79,990
	Satisfaction with work hours	7,578	**84,833**	84,864	84,856	84,859	84,863	84,833
	Subjective financial situation	12,466	**119,679**	119,697	119,743	119,707	119,707	119,680
	Subjective health	12,863	**111,456**	111,472	111,593	111,480	111,516	**111,456**
	Respondent cooperation	12,802	57,689	**57,603***	57,943	*No con*	57,717	57,689
Waves 11–15	Labour income	4,840	51,073	51,066	**51,046**	51,070	51,070	51,076
	Hours worked	4,485	116,445	116,477	**116,424**	*No con*	116,594	116,456
	Minutes travelling to work	4,156	131,329	**131,323**	131,305	*No con*	131,762	131,332
	General job satisfaction	4,332	52,609	52,633	**52,590**	52,631	53,318	52,616
	Satisfaction with wages	4,325	56,101	56,128	**56,079**	56,119	56,156	56,102
	Satisfaction with job security	4,318	55,464	55,487	**55,448**	55,484	55,762	55,464
	Satisfaction with actual work	4,330	52,961	52,979	**52,946**	52,972	52,978***	52,975
	Satisfaction with work hours	4,330	55,285	55,316	**55,260**	55,310	55,374	55,285
	Subjective financial situation	7,187	71,859	**71,852**	71,862	71,869	72,755	71,865
	Subjective health	7,386	75,057	75,058	**75,035**	75,083	75,304***	75,058
	Respondent cooperation	7,289	27,820	**27,770****	27,823	*No con*	28,052***	27,821
	No. of times best model		7	4	11	0	0	1

Notes: * Initial estimate of variance of T4 is −0.01. Model converges when variance of T4 is subsequently constrained to 0.01. ** Initial estimate of variance of T3 is −0.02. Model converges when variance of T3 is subsequently constrained to 0.01. *** Converges with Bayesian convergence criterion of 0.05, instead of 0.01. No con = failed to converge. Bold entry = lowest BIC.

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 (that 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 7.1 for details).

Second, we find that Model 4 (that with lag-2 parameters) produces inconsistent estimates. The standardized stability parameters are higher than 1 in the models for 'hours worked', interviewer-rated '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 (that with correlated errors). The interviewer-rated 'respondent cooperation' produced in this case a negative variance for the true score at wave four (unstandardized coefficient of -0.01). We have subsequently constrained this parameter to be 0.01 and proceed to interpret the other parameters of this model with caution.

Our results are structured as follows. 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 as compared to the baseline QSM (Model 1).

7.4 Results

Table 7.1 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, 'minutes travelling 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 appears as a correlated error in the model. For the other two variables that have correlated errors in waves 11–15 the reason is less obvious. However, if we look at the parameter estimates of the correlations for these variables in Table 7.2 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 'minutes travelling 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. Alternatively, it could be due to the respondent finding a new job in a new location or moving house so their commute to the same job is longer or shorter and the underlying trait has substantially changed but the change is being misclassified as measurement error by the model.

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 7.3 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 and 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

Table 7.2 Correlated measurement errors for variables where model with correlated errors fit the data best.

Coefficient	**Respondent cooperation**		**Subjective financial situation**	**Minutes travelling 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 7.1.

Table 7.3 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 travelling 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 travelling to work	11–15	Correlated errors	0.90	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 7.1.

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. First, 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. Second, we find that when the Model 2 (QSM with correlated errors) fits the data best, parameter 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 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.

7.5 Conclusions and Discussion

This chapter has shown 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 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 log-likelihood. In terms of parameter estimates, closer inspection of the results of Model 6 shows that for almost all our variables the variances in our model do follow a normal distribution. Only when either the reliability or stability estimate is close to 1 do 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 travelled 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 not only affects the interpretation of measurement errors but also 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 quasi-simplex model often fails to converge, the baseline QSM 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, 2014; Coenders et al., 1999; Hargens et al., 1976; Jagodzinski and Kuhnel, 1987). We still know relatively little about the causes of these convergence problems. Other models such as the latent state–trait model (Kenny and 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 the 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 chapter provides an overview of how to relax and test for this, and other assumptions of the quasi-simplex model. That being said, future research should also take into consideration how freeing multiple assumptions at the same time affects convergence and coefficients.

References

Allison, P.D., Williams, R., and Moral-Benito, E. (2017). 'Maximum Likelihood for Cross-Lagged Panel Models with Fixed Effects.' *Socius* 3, 2378023117710578.

Alwin, D.F. (1989). 'Problems in the Estimation and Interpretation of the Reliability of Survey Data.' *Quality and Quantity*, 23(3–4): 277–331. DOI: 10.1007/BF00172447.

Alwin, D.F. (2007). *The Margins of Error: A Study of Reliability in Survey Measurement*. Wiley-Blackwell.

Alwin, D.F. and Krosnick, J.A. (1991). 'The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes.' *Sociological Methods & Research*, 20(1): pp. 139–81. DOI: 10.1177/0049124191020001005.

Bast, J. and 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): pp. 135–67. DOI:10.1207/s15327906mbr3202_3.

Blok, H. and Saris, W.E. (1983). 'Using Longitudinal Data to Estimate Reliability.' *Applied Psychological Measurement*, 7(3): pp. 295–301.

Bound, J., Brown, C., and Mathiowetz, N. (2001). *Measurement Error in Survey Data* (PSC Research Report No. 00–450) (pp. 3705–843). Elsevier. Retrieved from http://ideas.repec.org/h/eee/ecochp/5–59.html.

Cernat, A. (2014). 'The Impact of Mixing Modes on Reliability in Longitudinal Studies'. *Sociological Methods & Research*, 0049124114553802. DOI: 10.1177/0049124114553802.

Coenders, G., Saris, W., Batista-Foguet, J., and Andreenkova, A. (1999). 'Stability of Three-Wave Simplex Estimates of Reliability.' *Structural Equation Modeling: A Multidisciplinary Journal*, 6(2): pp. 135–57. DOI: 10.1080/10705519909540125.

Hagenaars, J. (1990). *Categorical Longitudinal Data: Log-Linear Panel, Trend, and Cohort Analysis*. SAGE Publications.

Hamaker, E.L., Kuiper, R.M., and Grasman, R.P. (2015). 'A Critique of the Cross-Lagged Panel Model.' *Psychological Methods*, 20(1): p. 102.

Hargens, L.L., Reskin, B.F., and 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): pp. 439–58.

Heise, D.R. (1969). 'Separating Reliability and Stability in Test-Retest Correlation.' *American Sociological Review*, 34(1): pp. 93–101.

Jagodzinski, W. and Kuhnel, S. M. (1987). 'Estimation of Reliability and Stability in Single-Indicator Multiple-Wave Models.' *Sociological Methods & Research*, 15(3): pp. 219–58. DOI: 10.1177/0049124187015003003.

Jagodzinski, W., Kuhnel, S. M., and Schmidt, P. (1987). 'Is There a "Socratic Effect" in Nonexperimental Panel Studies? Consistency of an Attitude Toward Guestworkers.' *Sociological Methods & Research*, 15(3): pp. 259–302. DOI: 10.1177/0049124187015003004.

Jöreskog, K. G. (1970). 'Estimation and Testing of Simplex Models.' *British Journal of Mathematical and Statistical Psychology*, 23(2): pp. 121–45. DOI: 10.1111/j.2044–8317.1970.tb00439.x.

Kaplan, D., and Depaoli, S. (2012). 'Bayesian Structural Equation Modeling'. In R. H. Hoyle (ed.), *Handbook of Structural Equation Modeling*. New York: Guilford Press, pp. 650–73.

Kenny, D. and Zautra, A. (2001). 'Trait-State Models for Longitudinal Data'. In L. M. Collins and A. Sayer (eds.), *New Methods for the Analysis of Change*. Washington, DC: American Psychological Association, pp. 241–64.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Lugtig, P., Boeije, H. R., and Lensvelt-Mulders, G. J. L. M. (2012). 'Change? What Change?' *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 8(3): pp. 115–23. DOI: 10.1027/1614–2241/a000043.

Mandys, F., Dolan, C. V., and 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): pp. 201–15.

Muthén, L. and Muthén, B. (2013). *Mplus User's Guide* (7th edition). Los Angeles, CA: Muthén & Muthén.

Palmquist, B. and Green, D. P. (1992). 'Estimation of Models with Correlated Measurement Errors from Panel Data.' *Sociological Methodology*, pp. 119–46.

Plewis, I. (1985). *Analysing Change: Measurement and Explanation Using Longitudinal Data*. J. Wiley.

Rogosa, D. (1985). 'Myths and Methods: "Myths about Longitudinal Research" Plus Supplemental Questions'. In J. M. Gottman (ed.), *The Analysis of Change*. Mahwah, NJ: Erlbaum, pp. 3–66.

Rogosa, D. and Willett, J. B. (1985). 'Satisfying a Simplex Structure is Simpler than it Should Be.' *Journal of Educational Statistics*, 10(2): pp. 99–107. DOI: 10.2307/1164837.

Saris, W. and Gallhofer, I. (2007). 'Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions.' *Survey Research Methods*, 1(1): pp. 29–43.

Saris, W. and Van Den Putte, B. (1988). 'True Score or Factor Models: A Secondary Analysis of the ALLBUS Test–Retest Data.' *Sociological Methods & Research*, 17(2): pp. 123–57. DOI: 10.1177/0049124188017002001.

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., and Brunton-Smith, I. (2009). 'Attitudes over Time: The Psychology of Panel Conditioning'. In P. Lynn (ed.), *Methodology of Longitudinal Surveys*. Chichester: Wiley, pp. 113–26.

Taylor, M.F., Brice, J., Buck, N., and 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): pp. 120–36. DOI: 10.14301/llcs.v3i1.164.

Uhrig, N. and Watson, N. (2020). 'The Impact of Measurement Error on Wage Decompositions: Evidence from the British Household Panel Survey and the Household, Income and Labour Dynamics in Australia Survey.' *Sociological Methods and Research*, 49(1): pp. 43–78. DOI: 10.1177/0049124117701476.

Werts, C.E., Jöreskog, K.G., and Linn, R.L. (1971). 'Comment on "The Estimation of Measurement Error in Panel Data".' *American Sociological Review*, 36(1): pp. 110–13.

Wiley, D. and Wiley, J. (1970). 'The Estimation of Measurement Error in Panel Data.' *American Sociological Review*, 35(1): pp. 112–17.