City Research Online

City St George's, University of London

Citation: Schnell, R. & Noack, M. (2016). Reliability and Stability of the Standard Fear of Crime Indicator in a National Panel Over 14 Years. Survey Research Methods, 10(3), pp. 253-264. doi: 10.18148/srm/2016.v10i3.6714

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/15962/

Link to published version: https://doi.org/10.18148/srm/2016.v10i3.6714

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Reliability and Stability of the Standard Fear of Crime Indicator in a National Panel Over 14 Years

Rainer Schnell
City University London
Rainer.Schnell@city.ac.uk

Marcel Noack
University of Duisburg-Essen
Marcel.Noack@uni-due.de

September 21, 2016

Abstract

Reducing citizens' fear of crime is a salient policy topic in Western societies. For the measurement of fear of crime (FoC), variations of a single item measure ("Is there any area near where you live – that is, within a mile – where you would be afraid to walk alone at night?") have been used in hundreds of surveys for half a century. Despite a long and ongoing discussion on the doubtful usefulness of this standard indicator, no estimate for the reliability of this item is available. Using panel data from the British Household Panel Survey in combination with the UK Household Longitudinal Survey, reliability and stability of the standard fear of crime indicator are estimated using quasi-Markov simplex models for the first time. The model shows estimated reliabilities of about .67 for the standard indicator. Estimated reliability for FoC is smaller than other reliability estimates for single item measures in the same data set but close to the median reliability of single item measures reported in the literature.

Keywords: Panel, fear of crime, Wiley and Wiley model, reliability, stability

1 Introduction

Personal safety seems to be among the most important human goals. For example, Maslow (1943) lists safety as a second motivation after physiological needs. Accordingly, personal safety has become an important policy

area since about 1970 (Farrall et al., 2009). Since the relationship between objective victimization risks and subjective perceptions of safety appears to be moderated by mechanisms only little understood (Jackson, 2011), direct measurements of perceived safety seem to be needed for policy evaluations.
Hence, fear of crime is a frequent topic in social surveys. Most often, fear of crime is measured with a variant of the item "*Is there any area near where you live – that is, within a mile – where you would be afraid to walk alone at night?*". Using surveys in the United Kingdom as an example, variations of this indicator were used at least once since 2000 in the following studies: the *Crime Survey for England and Wales* (formerly known as *British Crime Survey*), both the *British Household Panel Survey* and the *UK Household Longitudinal Survey* into which the BHPS has merged, the *Offending Crime and Justice Survey*, the *New Deal Community Survey*, the *British Social Attitudes Survey*, the *Continuous Household Survey*, the *(Home Office) Citizenship Survey*, the *Young Life and Times Survey*, the *Health Survey for England*, the *National Child Development Study*, the *General Household Survey*, the *Health Survey for England*, the *Millennium Cohort Study*, the *Scottish Household Survey*, the *Scottish Social Attitudes Survey*, the *Scottish Crime and Justice Survey*, and the *Northern Ireland Crime Survey*. International Surveys using this item in the United Kingdom are the *International Crime Victims Survey* (ICVS)/ *European Survey of Crime and Safety* (EU ICS), several rounds of the *Eurobarometer* and the *European Social Survey* (ESS).
Given the use of this indicator in such reputed surveys for temporal and international comparisons (Hummelsheim et al., 2011; Visser et al., 2013), one might expect the measurement properties of this indicator to be well established. However, we were unable to find quantitative estimates for either validity or reliability for this indicator in the literature. Given the fact that content and face-validity of the crime indicator have been criticized repeatedly (as will be discussed in the section 2), this lack of quantitative evidence of the measurement quality of a policy-relevant item is worrying.
In general, validity is widely regarded as the most important measurement property. Unfortunately, the estimation of statistical models designed for the analysis of validity (e.g. the true-score-MTMM-models of Saris and Andrews (1991); Scherpenzeel and Saris (1997)), requires specific designed surveys (for example, measurements of the same construct using different modes). Despite a 40 year history of the use of the standard indicator, we are not aware of such surveys. However, using reliability as a tool, an upper bound of the validity of this item be estimated (Alwin, 1989). This is due to the fact, that "reliability of a test is a measure of the degree of true-score variation relative to observed-score variation" and that the "validity of a test with respect

to any criterion cannot exceed the index of reliability" (Lord and Novick, 1968, 61,72). Therefore, validity can not exceed $\sqrt{\sigma_t^2/\sigma_x^2} = \sqrt{\sigma_{\text{true}}^2/\sigma_{\text{measured}}^2}$. Further, one has to keep in mind that reliability is a necessary condition for validity, but not a sufficient one. Therefore, an indicator which is not reliable is also not valid, but an indicator which is reliable is not guaranteed to be valid but only that it *can* be valid (Alwin, 2007).
Therefore, this study is the first attempt to quantify reliability and stability of the standard fear of crime indicator. Since the standard indicator is a single item and not a composite scale, reliability cannot be estimated by measures of internal consistency such as Cronbach's α (section 3).
We will estimate reliability with a special version of quasi-Markov simplex models (Jöreskog, 1970) for single indicators as proposed by Wiley and Wiley (1970). Using reasonable assumptions, such models can separate true change and random measurement error given panel data with a sufficient number of waves (Alwin, 2007; van de Pol and de Leeuw, 1986). For the estimation, a four-wave panel based on the British Household Panel Study (BHPS, three waves) and the UK Household Longitudinal Survey (UKHLS, one wave), will be used (section 4). Finally, we will discuss the implications of the results.

2 Measuring fear of crime with the standard indicator

Despite the wide use of fear of crime indicators, the conceptualization and measurement of fear of crime is still debated (Farrall et al., 2009). Most often, a simple measure based on one item is used. Respondents are asked some variation of the questions "*How safe do you feel or would you feel walking alone in your neighborhood at night?*" or "*How safe do you feel walking alone in this area after dark?*" (Franklin et al., 2008).
These indicators are not immune to criticism and doubt regarding their validity. For some authors, they lack a clear conceptualization and presume that fear of crime is a unidimensional phenomenon (Farrall et al., 1997; Hale, 1996; Jackson, 2005; Kreuter, 2002). Further points of criticism are the hypothetical nature and diffuse spatial frame of reference of the situation described in the question, as well as the missing reference to crime, as pointed out by Garofalo (1979). Respondents are confronted with the task of assessing their perceived safety out alone in the dark, which may be a rare activity for many of them. Accordingly, Shapland and Vagg (1988) state that the attempt to measure fear through a potentially non-existent activity is questionable. Further problems arise with some wordings in variants of the standard indicator.

Terms such as "neighborhood" or "worry" are open to varying interpretations by different people and pose the risk of producing misleading results (Farrall et al., 1997; Hale, 1996). Subsequently, Ferraro and LaGrange (1987) address the question of whether the standard indicators measure fear of crime or just the perceived risk of being out alone at night. Persons with a high risk perception might answer that they feel unsafe, while they also avoid situations where they are out alone at night. Thus, they don't develop a high fear of crime because they don't put themselves in situations which they perceive as risky. Given the fact that the classic global indicator makes no reference to a specific crime, or any crime at all, Ferraro and LaGrange (1987) ask: "fear of what?".

Furthermore, the measurement of fear of crime using a single item seems to be hampered by a methodological artifact. The criminological literature on fear of crime suggests gender differences in response behaviour. Male respondents may not admit their fear to others or themselves and might answer in terms of an exaggerated masculine ideal (Goodey, 1997; Smith and Torstensson, 1997; Sutton and Farrall, 2005). It is therefore possible that male respondents over-report their feelings of safety and avoid conspicuous answers (Krosnick, 2002) in accordance with their perceptions of socially desirable responses (Tourangeau and Ting, 2007). Female respondents might answer according to their fear of sexual harassment, which may dominate in their appraisal of fear of crime. Warr (1984) coined the term "master offense", meaning that fear of rape or sexual assault poses a ubiquitous threat, capable of "shadowing" (Warr, 1985) other types of crime. According to Ferraro (1996), fear of sexual harassment manifests itself more strongly in situations outside of the subject's own dwelling, a situation explicitly formulated in the question. Additionally, female respondents are expected to report their fears more willingly than male respondents (Sutton and Farrall, 2005, 213).

Therefore, differences in male and female response patterns are expected. This implies different error correlations, requiring separate estimations of reliability and stability. This will be examined in section 5.

Finally, reliability is no guarantee of validity. A perfect reliable measure could measure something entirely different than intended. As early as the late 80th Shapland and Vagg (1988) suspected that the standard item only nominally measures fear of crime and could cover many other things (fear of the dark, fear of spooky places) as well. Accordingly, Noack (2015) showed, with data from the British Crime Survey, that diffuse fears, such as fear of the dark or fear of being out alone, which are not directly related to crime, have a stronger impact on the standard indicator than on offense-specific fear of crime items. Using qualitative interviews of extremely fearful respondents, Kury et al. (2004) traced fear of crime back to perceived incivilities.

Similar critiques of the standard indicator have been published by a couple of authors, for example Ferraro and LaGrange (1987) and Farrall et al. (1997). Therefore, many authors (for example Farrall et al., 2009; Ferraro, 1995; Jackson, 2005; Keane, 1992; Kreuter, 2002; Noack, 2015; Thompson et al., 1992; Warr, 1993) suggested replacing the standard indicator in the long run by more specific measures.
To sum up, at least face validity and content validity of the standard indicator is contested in the literature. Throughout the criminological literature, more crime-specific fear of crime questions are recommended (Farrall et al., 2009; Ferraro, 1995; Gray et al., 2008; Jackson, 2005; Kreuter, 2002; Noack, 2015; Warr, 1984; Warr and Stafford, 1983). However, in general social surveys, the standard indicator still prevails.

3 Methods for the estimation of reliability

Classical test theory defines reliability as the ratio of true score variance and observed score variance (Lord and Novick, 1968). Since true score variance is usually not available, reliability must be estimated. Two methods for the estimation of reliability are widely known: measures of internal consistency and the test-retest-procedure.
In most social science applications, reliability is estimated with a measure of internal consistency such as Cronbach's α. Despite its wide use, α as an estimator of reliability is not without problems (Bentler 2009, Green and Yang 2009, Revelle and Zinbarg 2009, Sijtsma 2009a, Sijtsma 2009b).
Especially the assumption of linearly related measures which only differ by a constant (essential τ-equivalency, Lord and Novick, 1968) seems to be rarely given in practice. In many cases, the assumption of uncorrelated errors will be violated resulting in overestimating reliability by the use of α (Cortina, 1993; Green and Hershberger, 2000; Green and Yang, 2009). Finally, since the standard indicator is a single item, neither α nor other measures of internal consistency can be applied here at all.
The other well-known method to assess the reliability of items is the test-retest method. Here, the same item is applied at two points in time for the same persons. The correlation between two measurements might be used as an approximation of reliability (Lord and Novick, 1968). This approach has to assume the absence of measurement problems such as memory effects or respondents' fatigue. Such problems seem to inflate the correlation between measures and therefore the estimated reliability. However, the major drawback of this approach is the questionable assumption of time-invariant true values (Lord and Novick, 1968). So using the test-retest method will

probably suffer from memory effects if the time between two measurements is short, or might suffer from changes in the true scores if the time between two measurements is long. Therefore, the test-retest method is rarely used for the estimation of reliability in surveys.

To overcome the problem of time-invariant scores, two aspects of a measurement should be clearly distinguished: The squared correlation between observed and true scores (reliability) and the temporal stability of a true score, measured by the correlation between true scores at two adjacent points in time. In survey settings, the most common problem concerns the measurement of a single variable at several timepoints where the assumption of time-invariant true scores seems unlikely. For the separate estimation of stability and reliability in these kind of problems, different methods have been proposed in the statistical literature. The most widely discussed (Alwin, 2007; Coenders et al., 1999; Saris and Gallhofer, 2007) approach is the family of quasi-Markov simplex models. Given at least three waves of data, these models allow for dynamic true values by decomposing change in true change and fluctuations due to measurement error.

The family of quasi-Markov simplex models can be most easily explained by a path model (see Fig 1). Quasi-Markov simplex models consist of two parts: a measurement model, which relates the manifest measures to the latent variables, and a structural model, which defines the relation of the latent variables. Interdependence between the underlying latent variables is modelled by a Markov process. In this class of models, the distribution of a latent variable at time t depends only on its distribution at time $t-1$. Earlier timepoints have only indirect effects on t (Alwin, 2007).

In matrix notation, the reduced form of the quasi-Markov simplex model is given by

$$\begin{aligned} x &= \Lambda_x \xi + \varepsilon && (1) \\ &= \Lambda_x (I - B)^{-1} \zeta + \varepsilon && (2) \end{aligned}$$

and

$$\Sigma_{xx} = \Lambda_x (I - B)^{-1} \Psi (I - B')^{-1} \Lambda_x' + \Theta^2 \qquad (3)$$

where x defines the $(P \times 1)$ vector of observed scores, $(I - B)$ represents the difference between the $(P \times P)$ identity matrix I and the $(P \times P)$ matrix B of regression coefficients linking the adjacent timepoints t and $t-1$. In this notation, ζ represents the $(P \times 1)$ vector of true score disturbances, ξ is the $(P \times 1)$ vector of true values and ε is the $(P \times 1)$ vector of measurement errors. Ψ is defined as $(P \times P)$ variance-covariance diagonal matrix of true score disturbances that contains the variances in the main diagonal. Θ^2

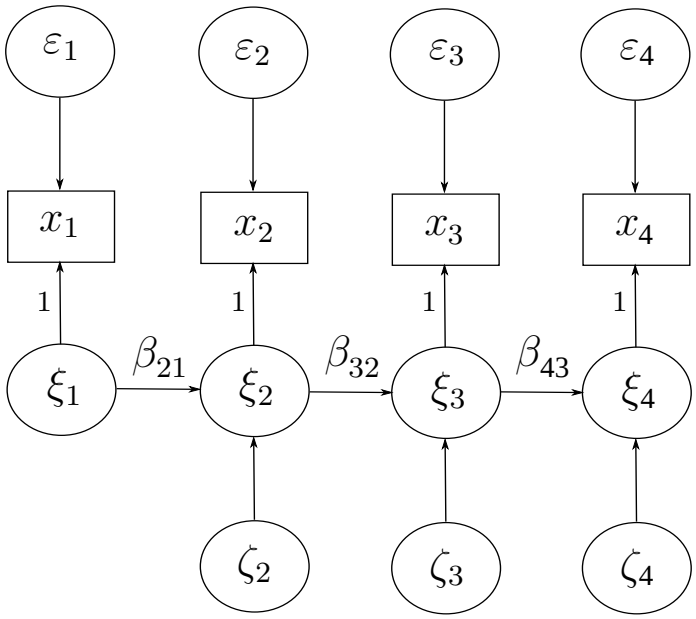

Figure 1: Path diagram of a four-wave quasi-Markov simplex model

represents the variance-covariance diagonal matrix of measurement errors. Λ_x is fixed to an identity matrix (Alwin, 2007).
The models of Heise (1969) and Wiley and Wiley (1970) are the best-known simplex models. The major difference between the models is the use of a correlation matrix in the Heise model, whereas the model of Wiley and Wiley uses a covariance matrix. As a consequence, the Heise-model assumes constant reliabilities over different points in time, whereas the reliabilities can vary in the Wiley and Wiley model. In turn, the residual variances are assumed to be constant over time in the Wiley model (Alwin and Krosnick, 1991). In the following, the approach proposed by Wiley and Wiley is used. Even if the differences tend to be only minor in general (Alwin, 2007), not assuming constant reliabilities seems to be more plausible.
Since the degrees of freedom for both models and P waves are given by

$$df = \frac{1}{2}(P(P+1)) - 2P, \tag{4}$$

these models possess two degrees of freedom in a panel with four waves. The models are therefore over-identified (Bollen, 1989). Hence, their goodness of fit can be tested with fit-indices for structural equation models such as

RMSEA (Alwin, 2007; Raykov and Marcoulides, 2006).[1]
Since the assumption of a multivariate normal distribution makes no sense for categorical data (Loehlin, 2004), we follow the recommendation of Jagodzinski and Kühnel (1987) regarding the computation of structural equation models for categorical items and do not use the Maximum Likelihood method. For estimation, polychoric covariances with time-invariant thresholds and a variant of the Weighted Least Squares-estimator (WLS, Browne, 1984) were used. This variant is known as Diagonally Weighted Least Squares (DWLS) and does not require sample sizes as large as does the WLS approach (Kaplan, 2000).
Overall, eight parameters need to be estimated in the four-wave Wiley and Wiley-model: The variance $\text{Var}(\xi_1)$ of the true scores at the first wave, for which $\xi_1 = \zeta_1$ applies, three lag-1 regression coefficients β_{21}, β_{32} and β_{43} which connect the true scores over time, three true score residual variances $\text{Var}(\zeta_2)$, $\text{Var}(\zeta_3)$ and $\text{Var}(\zeta_4)$ at the last three waves, and the error variance of the measuring instrument $\text{Var}(\varepsilon)$, which is constrained to be equal over time (Alwin, 2007).
Given this model reliability ρ_t^2 can then be estimated with

$$\begin{aligned}
\rho_1^2 &= \frac{\text{Var}(\zeta_1)}{\text{Var}(\zeta_1) + \text{Var}(\varepsilon)} \\
\rho_2^2 &= \frac{\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)}{\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2) + \text{Var}(\varepsilon)} \\
\rho_3^2 &= \frac{\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)}{\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3) + \text{Var}(\varepsilon)} \\
\rho_4^2 &= \frac{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4)}{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4) + \text{Var}(\varepsilon)}.
\end{aligned} \tag{5}$$

Accordingly, stability $\gamma_{t+1,t}$ is estimated with

$$\begin{aligned}
\gamma_{21} &= \beta_{21}\frac{\sqrt{\text{Var}(\zeta_1)}}{\sqrt{\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)}} \\
\gamma_{32} &= \beta_{32}\frac{\sqrt{\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)}}{\sqrt{\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)}} \\
\gamma_{43} &= \beta_{43}\frac{\sqrt{\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)}}{\sqrt{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4)}}.
\end{aligned} \tag{6}$$

[1] For a panel with three waves, both models are saturated with $df = 0$ and $\chi^2 = 0$. Therefore, a panel with at least 4 waves is needed to test the goodness of fit for the estimated model.

Since the paper of Wiley and Wiley (1970) only covers the case of a three wave panel, we added the derivation of β_{43} and γ_{43} in appendix C.

4 Data

As explained above, for an identified Wiley model a panel study with at least four waves is needed. Since inference to general populations was intended, surveys of subgroups such as students, children or offenders are not suited. The same logic excluded panels confined to deprived geographical regions. These requirements limited strongly the number of available studies. Despite an intensive search in academic data repositories and the literature, we were not able to locate a single study until the combined data of BHPS (University of Essex, 2009) and UKHLS (University of Essex, 2013) became available for scientific use in 2014 (described below). To the best of our knowledge, BHPS/UKHLS is currently the only general population data set with this indicator and more than three waves.

The data set used here has two two sources: The British Household Panel Survey (BHPS) and the United Kingdom Household Longitudinal Survey (UKHLS), which is known to its respondents as *Understanding Society*. The UKHLS "both replaces and incorporates" the BHPS (Boreham et al., 2012, 1). As of wave 2, the UKHLS "consists of all members from the BHPS sample who were still active at Wave 18 in the BHPS and who had not refused consent to be issued as part of the Understanding Society sample" (McFall, 2013, 31).

The BHPS started as an annual survey in 1991, interviewing each adult member of a household (children are treated as adults and interviewed once they turn 16 years old), resulting in a sample of more than 5.000 households and approximately 10.000 individuals in Great Britain south of the Caledonian canal and excluding Northern Ireland. Initial selection of households for the first BHPS-panel wave used a two-stage stratified systematic design. The mode of data collection changed from paper-and-pencil (PAPI) to computer-assisted personal interview (CAPI) as of wave 9 (Taylor et al., 2009).

From the BHPS, all waves are used, where the classical fear of crime indicator "*how safe do you feel walking alone in this area after dark?*" was asked. These are waves 7, 12 and 17. The fieldwork for these waves was done between August 1997 and May 1998 for wave 7, from September 2002 to April 2003 for wave 12 and from September 2007 to April 2008 for wave 17 (Taylor et al., 2009). In the UKHLS, the classical fear of crime indicator was used in wave 3. The UKHLS-fieldwork for wave 3 took place between January 2011 and December 2012, but almost 95% of BHPS-respondents were interviewed in

2011. The mode of data collection for the individual adult questionnaires was CAPI, just as in the BHPS (Scott and Jessop, 2013).
Overall 2851 respondents answered the fear of crime indicator in all four waves. Weighting the data would lead to the exclusion of 767 observations, since respondents who did not respond at each wave up to and including the latest wave (Taylor et al., 2009) received a longitudinal weight of zero. Therefore, it is possible that a respondent has answered the fear of crime indicator each time it was part of the questionnaire, but drops out of the analysis due to a zero weight because he or she did not participate in some other wave were the indicator was not asked. For the analysis intended here, this seems to be a suboptimal use of the available data, since about 27% of the respondents who answered the fear of crime-indicator in all four waves would be excluded. Because of that, the analysis is based on the unweighted data set, including the 767 respondents with zero weights.[2]

5 Exploration of gender differences in response behaviour to the fear of crime indicator

As mentioned in section 2, based on the criminological literature, it is anticipated that the response patterns of men and women differ. To explore this, parallel coordinate plots of the response profiles are used here (Inselberg, 2009). To avoid problems by overplotting due to the categorical nature of the fear of crime indicators, jittering and α-blending (Cleveland, 1993; Theus, 2008) was applied. For plotting, R (R Core Team, 2014) with Lattice (Sarkar, 2014) were used.
Since different response profiles for male and female respondents were expected, separate profiles were plotted. Fig. 2 shows the resulting parallel coordinate plots for male and female respondents of the BHPS/UKHLS data. The tendency of male respondents to assign themselves to categories labelled as "safe" is obvious. Few female respondents choose the category "very safe". About 81% of male respondents as opposed to 47% of female respondents feel "(very) safe" at all four points in time. Female respondents show great variability in their responses, whereas male respondents are almost exclusively confined to transitioning between the categories "safe" and "very safe".
Due to this preliminary analysis, the Wiley and Wiley-model is estimated separately for male and female respondents.

[2] Comparisons of covariance matrices based on weighted and unweighted data sets using Box M-tests yield insignificant differences. Furthermore, results for model parameters as well as reliability and stability coefficients are nearly the same for weighted and unweighted

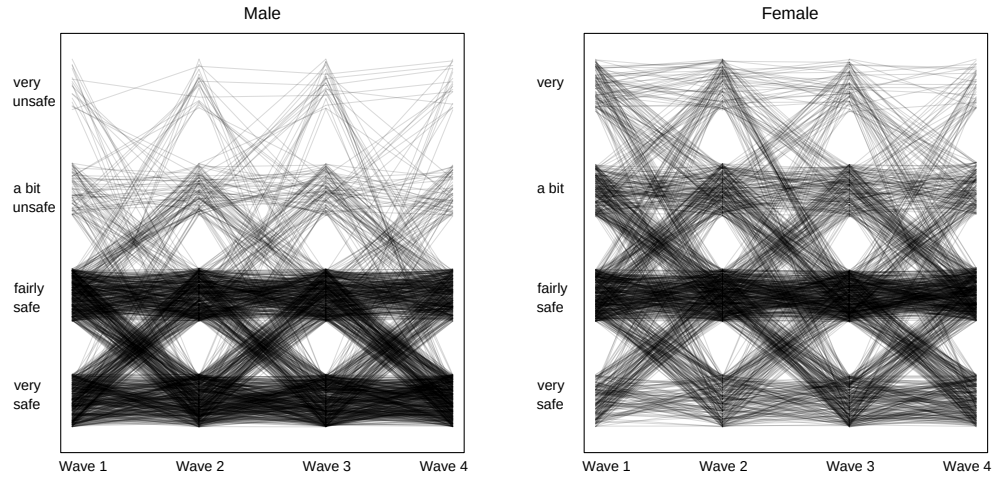

Figure 2: Response profiles on the standard indicator for 1505 male and 1346 female respondents in the BHP/UKHLS panel

6 Estimation results

Lisrel 9.2 (Scientific Software International, 2015) was used for the computation. Table 1 shows the estimates for the eight parameters for all respondents as well as separated for male and female respondents. Using these estimates and the equations (5) and (6) give the final estimates for reliability and stability shown in Table 2.

All goodness of fit measures shown in Table 1 (χ^2, RMSEA, TLI, CFI, SRMR; see Loehlin, 2004) indicate a more than reasonable fit of the model in all three populations considered. Each estimated coefficient is significantly different from zero with the exception of $\text{Var}(\zeta_4)$ for female respondents. Given 24 estimated parameters this may be simply a random outlier. Considering the acceptable fit indices, the model seems to fit the data quite well.

The main results of this study are the estimated reliabilities and stabilities given in Table 2. For the total sample, the stability estimates are above 0.8 for all waves, and between .771 and .899 for the subgroups.

At first sight, the fact that reliabilities ρ_1^2, ρ_2^2 and ρ_3^2 are higher for the total sample than for both subgroups may be surprising. Given the definition of reliability as the ratio of true score variance to observed variance, the observed pattern is most likely the result of reduced true score variance in the subgroups. Considering the observed response profiles in Figure 2, this reduction in true score variance in both subgroups might be plausible.

data.

Table 1: Wiley and Wiley parameter estimates and goodness of fit measures for the standard fear of crime using merged BHPS/UKHLS data.

	All		Male		Female	
	$\hat{\theta}$	$\hat{\sigma}_\theta$	$\hat{\theta}$	$\hat{\sigma}_\theta$	$\hat{\theta}$	$\hat{\sigma}_\theta$
β_{21}	.805	(.059)	.857	(.114)	.726	(.083)
β_{32}	.879	(.031)	.943	(.061)	.877	(.053)
β_{43}	.887	(.031)	.878	(.055)	.832	(.046)
Var(ζ_1)	.366	(.026)	.206	(.028)	.344	(.038)
Var(ζ_2)	.109	(.023)	.094	(.029)	.124	(.031)
Var(ζ_3)	.073	(.013)	.061	(.020)	.095	(.020)
Var(ζ_4)	.065	(.029)	.104	(.039)	.054	(.041)
Var(ε)	.167	(.017)	.149	(.021)	.177	(.025)
$\chi^2_{df=2}$	2.342	p=.310	3.200	p=.202	0.685	p=.710
RMSEA	0.029		0.051		0.000	
TLI	0.996		0.986		1.001	
CFI	0.999		0.995		1.000	
SRMR	0.007		0.013		0.006	
n	2851		1505		1346	

Table 2: Wiley and Wiley reliability and stability estimates for the standard fear of crime indicator for the merged BHPS/UKHLS data.

	All	Male	Female
ρ^2_1	.687	.580	.660
ρ^2_2	.675	.622	.633
ρ^2_3	.671	.652	.651
ρ^2_4	.666	.682	.615
γ_{21}	.828	.785	.771
γ_{32}	.886	.884	.844
γ_{43}	.897	.821	.899

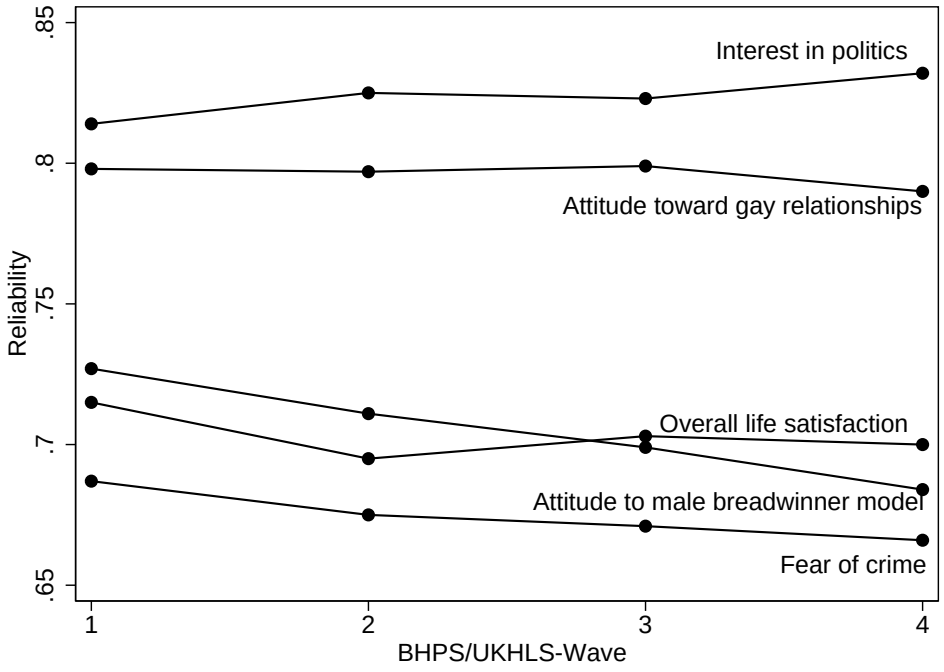

Figure 3: Comparison of reliability estimates: Standard indicator and four attitude items in the BHPS/UKLS. Estimates from Table B.1 in the appendix.

As there is no generally accepted or well-proven absolute threshold for the reliability of a single item in social research, it might be useful to compare the standard indicator with other attitude items in the BHPS/UKLS. We deliberately selected attitudes which we expected to be stable at least for some years (attitude to male breadwinner model), two items with expected average reliability (gay relationships and interest in politics, Alwin, 2007; Prior, 2010) and an item which is known for its low reliability (overall life satisfaction, Lucas and Donnellan, 2012; Schimmack et al., 2010). The questions and estimates are given in the appendix (Table B.1). To make comparisons easier, the results are presented as a lineplot (see Figure 3). At each point in time, the standard indicator has a lower reliability than all other items selected here. Furthermore, at all points in time stability of the indicator is lower all other estimated stabilities.

7 Discussion

Reliabilities between 0.67 and 0.69 for the total sample and 0.58 to 0.68 for the subgroups are better results than expected, given the critical discussion of the standard indicator in the literature. However, there are no fixed cut-off values for reliabilities which might permit the classification of items as acceptable or not. In most applications for reliability estimates, reliability is applied to a test consisting of multiple items. Therefore, most discussions of reliability focus on tests. Since the standard fear of crime indicator is not a scale but a single item, standard guidelines such as 0.7 as minimum value should be considered with caution.

In general, the popular 0.7-cutoff seems to be a misinterpretation of the recommendations in the psychometric literature (Lance et al., 2006). For example, Nunnally (1967, 226) states that a satisfactory level of reliability "(...) depends on how a measure is being used. In the early stages of research on predictor tests or hypothesized measures of a construct, one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .60 or .50 will suffice." But if "(...) important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard." Surveys are not used to make important decisions on individuals, but after 50 years of research on fear of crime, it could hardly be considered to be in its early stages.

So an estimated reliability of 0.67 is certainly not suitable for psychometric purposes, but it might be sufficient for some research purposes, for example comparing aggregated measurements for small geographical areas based on

large numbers of observations.
Therefore, comparison data for single item reliability estimates of survey measures in general populations might be helpful for evaluation of the standard indicator. Recently, two comprehensive studies on survey reliabilities based on simplex models have been published.
Alwin (2007) used the National Election Study (NES), the American Changing Lives Panel (ACL) and the Study of American Families (SAF) to estimate reliabilities of common survey items. The table in the Appendix of his book contains 488 estimates, of which 448 relate to respondents' self-reports. Most of these (347) concern nonfacts. Alwin (2007) gives a mean reliability of 0.634 for this subset. Using a digitized version of the Alwin table, we computed a mean reliability for all items of 0.670, with a median of 0.667. The reliability estimates (0.666–0.687) of the standard fear of crime indicator in the BHPS/UKLS would be close to the 50% percentile of the Alwin estimates. The worst reliability estimate in the BHPS/UKLS (0.552; male respondents wave 1-2) corresponds to the 26% percentile in the Alwin table.
More recently, Hout and Hastings (2012) published reliability and stability estimates of 281 items of the General Social Survey (GSS) three-wave panel (2006, 2008, 2010). For 97 items concerning beliefs and values, they reported 0.690 as mean reliability (median: 0.706). For 63 attitude items, a mean reliability of 0.664 (median 0.658) was observed. This is very close to the estimates reported here.
To summarize the discussion, reliability of the standard fear of crime indicator does not approach psychometric standards for clinical use, but this is true for many other survey items. It should be kept in mind that 50% of all items in both studies cited above have lower reliabilities. Therefore, reliability is either not a central problem of the fear of crime indicator or a central problem of at least half the other items in social science research. We tend to the second interpretation. Improving the measurement of core concepts such as fear of crime seems to us to be of utmost importance. Therefore we hope that the standard indicator of fear of crime will be replaced by offense-specific measures based on rational choice framework (such as Winkel (1981) recommended over 30 Years ago) in the near future. Testing the measurement properties of such a theoretically based reconstruction of fear of crime is subject of our ongoing research.

Appendix

A Questions used in the BHPS/UKHLS

Attitude toward gay relationships: Do you personally agree or disagree with the following statements. [Homosexual relationships are always wrong.]

1 Strongly agree
2 Agree
3 Neither agree nor disagree
4 Disagree
5 Strongly disagree

Interest in politics: How interested would you say you are in politics? Would you say you are

1 Very interested
2 Fairly interested
3 Not very interested
4 Not at all interested

Attitude to male breadwinner model: Here are some questions about family life. Do you personally agree or disagree ... [A husband's job is to earn money; a wife's job is to look after the home and family]

1 Strongly agree
2 Agree
3 Neither agree nor disagree
4 Disagree
5 Strongly disagree

Overall life satisfaction: Here are some questions about how you feel about your life. Please tick the number which you feel best describes how dissatisfied or satisfied you are with the following aspects of your current situation.

1 Not satisfied at all — 7 Completely satisfied

Using the same scale how dissatisfied or satisfied are you with your life overall?

1 Not satisfied at all — 7 Completely satisfied

Fear of crime: How safe do you feel walking alone in this area after dark?

1 Very safe
2 Fairly safe
3 A bit unsafe
4 Very unsafe

B

Table B.1: Reliability and Stability estimates for the standard indicator and 4 other attitude items in the BHPS/UKLS.

	Fear of crime	Attitude toward gay relationships	Attitude to male breadwinner model	Overall life satisfaction	Interest in politics
ρ_1^2	0.687	0.798	0.727	0.715	0.814
ρ_2^2	0.675	0.797	0.711	0.695	0.825
ρ_3^2	0.671	0.799	0.699	0.703	0.823
ρ_4^2	0.666	0.790	0.684	0.700	0.832
γ_{21}	0.828	0.972	0.965	0.885	0.984
γ_{32}	0.886	0.976	0.956	0.967	0.980
γ_{43}	0.897	0.986	1.005	0.936	0.972
$\chi^2_{df=2}$	2.342, p=0.310	4.197, p=0.123	1.116, p=0.572	4.373, p=0.112	1.867, p=0.393
RMSEA	0.029	0.037	0.010	0.024	0.021
TLI	0.996	0.997	1.000	0.998	0.999
CFI	0.999	0.999	1.000	0.999	1.000
SRMR	0.007	0.004	0.002	0.005	0.003
n	2851	9329	9829	10931	12214
BHPS-waves	glqc†	lnpr	kmoq	opqr	opqr

†: Wave c is the third UKHLS wave

C Notes on the estimation of reliability and stability in a 4-wave panel

For the derivation of the estimating equations for waves 1 to 3, see Wiley and Wiley (1970). Extending the formulae for wave 4 is straight forward, given that

$$\text{Var}(\xi_4) = \beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4)$$

and

$$\text{Var}(x_4) = \text{Var}(\xi_4) + \text{Var}(\varepsilon).$$

Reliabilities are defined by Wiley and Wiley (1970, 112-114) as

$$\rho_t^2 = \frac{\text{Var}(\xi_t)}{\text{Var}(x_t)} = \frac{\text{Var}(\xi_t)}{\text{Var}(\xi_t) + \text{Var}(\varepsilon)}$$

following the standard psychometric definition by Lord and Novick (1968) which leads to

$$\rho_4^2 = \frac{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4)}{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4) + \text{Var}(\varepsilon)}.$$

Stability for two adjacent points in time $\gamma_{t+1,t}$ is defined by Wiley and Wiley (1970, 114-115) as correlation between the true scores ξ_t and ξ_{t+1}. Transformation of $\beta_{t+1,t}$ by multiplication with the factor

$$\sqrt{\frac{\text{Var}(\xi_t)}{\text{Var}(\xi_{t+1})}}$$

gives the correlation between ξ_t and ξ_{t+1}

$$\gamma_{43} = \beta_{43}\frac{\sqrt{\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)}}{\sqrt{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2\text{Var}(\zeta_1) + \text{Var}(\zeta_2)) + \text{Var}(\zeta_3)) + \text{Var}(\zeta_4)}}.$$

This could also be obtained by requesting the "completely standardized solution" for $\beta_{t+1,t}$ in Lisrel which gives the standardized latent regression coefficients.

References

Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Qual. Quant.*, 23:277–331.

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

Alwin, D. F. and Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociol. Methods Res.*, 20:139–181.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, 74:137–143.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

Boreham, R., Boldysevaite, D., and Killpack, C. (2012). *UKHLS: Wave 1 Technical Report*. Institute for Social and Economic Research, University of Essex, Colchester.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *Br. J. Math. Stat. Psychol.*, 37:62–83.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, Summit.

Coenders, G., Saris, W. E., Batista-Foguet, J. M., and Andreenkova, A. (1999). Stability of three-wave simplex estimates of reliability. *Struct. Equ. Modeling*, 6:135–157.

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *J. Appl. Psychol.*, 78:98–104.

Farrall, S., Bannister, J., Ditton, J., and Gilchrist, E. (1997). Questioning the measurement of the "fear of crime": Findings from a major methodological study. *Brit. J. Criminol.*, 37:658–679.

Farrall, S., Jackson, J., and Gray, E. (2009). *Social Order and the Fear of Crime in Contemporary Times*. Oxford University Press, Oxford.

Ferraro, K. F. (1995). *Fear of Crime: Interpreting Victimization Risk*. State University of New York Press, Albany.

Ferraro, K. F. (1996). Women's fear of victimization: Shadow of sexual assault? *Soc. Forces*, 75:667–690.

Ferraro, K. F. and LaGrange, R. (1987). The measurement of fear of crime. *Sociol. Inq.*, 57:70–101.

Franklin, T. W., Franklin, C. A., and Fearn, N. E. (2008). A multilevel analysis of the vulnerability, disorder, and social integration models of fear of crime. *Soc. Justice. Res.*, 21:204–227.

Garofalo, J. (1979). Victimization and the fear of crime. *J. Res. Crime. Delinq.*, 16:80–97.

Goodey, J. (1997). Boys don't cry. masculinities, fear of crime and fearlessness. *Brit. J. Criminol.*, 37:401–418.

Gray, E., Jackson, J., and Farrall, S. (2008). Reassessing the fear of crime. *Eur. J. Criminol.*, 5:363–380.

Green, S. B. and Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Struct. Equ. Modeling.*, 7:251–270.

Green, S. B. and Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74:121–135.

Hale, C. (1996). Fear of crime: A review of the literature. *Int. Rev. Victimology*, 4:79–150.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *Am. Sociol. Rev.*, 34:93–101.

Hout, M. and Hastings, O. P. (2012). Reliability and stability estimates for the GSS core items from the three-wave panels, 2006-2010. GSS Methodological Report 119, University of California, Berkely.

Hummelsheim, D., Hirtenlehner, H., Jackson, J., and Oberwittler, D. (2011). Social insecurities and fear of crime: a cross-national study on the impact of welfare state policies on crime-related anxieties. *Eur. Sociol. Rev.*, 27:327–345.

Inselberg, A. (2009). *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*. Springer, New York.

Jackson, J. (2005). Validating new measures of the fear of crime. *Int. J. Soc. Res. Meth.*, 8:297–315.

Jackson, J. (2011). Revisiting risk sensitivity in the fear of crime. *J. Res. Crime. Delinq.*, 48:513–537.

Jagodzinski, W. and Kühnel, S. M. (1987). Estimation of reliability and stability in single-indicator multiple-wave models. *Sociol. Methods Res.*, 15:219–258.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *Br. J. Math. Stat. Psychol.*, 23:121–145.

Kaplan, D. (2000). *Structural Equation Modeling: Foundation and Extensions*. Sage, Thousand Oaks.

Keane, C. (1992). Fear of crime in canada: An examination of concrete and formless fear of victimization. *Can. J. Criminol.*, 34:215–224.

Kreuter, F. (2002). *Kriminalitätsfurcht: Messung und methodische Probleme*. Leske+Budrich, Opladen.

Krosnick, J. A. (2002). The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, pages 87–100. Wiley, New York.

Kury, H., Lichtblau, A., Neumaier, A., and Obergfell-Fuchs, J. (2004). Zur Validität der Erfassung von Kriminalitätsfurcht. *Soziale Probleme*, 15:141–165.

Lance, C. E., Butts, M. M., and Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. what did they really say? *Organ. Res. Methods*, 9:202–220.

Loehlin, J. C. (2004). *Latent Variable Models - An Introduction to Factor, Path, and Structural Equation Analysis*. Erlbaum, Mahwah, 4. edition.

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

Lucas, R. E. and Donnellan, M. B. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Soc. Indic. Res.*, 105:323–331.

Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.*, 50:370–396.

McFall, S. (2013). *Understanding Society: The UK Household Longitudinal Study. Waves 1-3, User Manual*. Institute for Social and Economic Research, University of Essex, Colchester.

Noack, M. (2015). *Methodische Probleme bei der Messung von Kriminalitätsfurcht und Viktimisierungserfahrungen*. SpringerVS, Wiesbaden.

Nunnally, J. C. (1967). *Psychometric Theory*. McGraw-Hill, New York.

Prior, M. (2010). You've either got it or you don't? The stability of political interest over the life cycle. *J. Polit.*, 72:747–766.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raykov, T. and Marcoulides, G. A. (2006). *A First Course in Structural Equation Modeling*. Erlbaum, Mahwah, 2. edition.

Revelle, W. and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the GLB: Comment on Sijtsma. *Psychometrika*, 74:145–154.

Saris, W. E. and Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S., editors, *Measurement Error in Surveys*, pages 575–597. Wiley, New York.

Saris, W. E. and Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley, Hoboken.

Sarkar, D. (2014). *Lattice Graphics*. CRAN, R-Package, version 0.20-29 edition.

Scherpenzeel, A. C. and Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of mtmm studies. *Sociol. Methods Res.*, 25:341–383.

Schimmack, U., Krause, P., Wagner, G. G., and Schupp, J. (2010). Stability

and change of well being: An experimentally enhanced latent state-trait-error analysis. *Soc. Indic. Res.*, 95:19–31.

Scientific Software International (2015). *Lisrel 9.2*. Skokie.

Scott, A. and Jessop, C. (2013). *UK Household Longitudinal Study (UKHLS) Wave 3: Technical Report*. The Institute for Social and Economic Research, University of Essex, Essex.

Shapland, J. and Vagg, J. (1988). *Policing by the Public*. Routledge, London.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74:107–120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, 74:169–173.

Smith, W. R. and Torstensson, M. (1997). Fear of crime. Gender differences in risk perception and neutralizing fear of crime. Toward resolving the paradoxes. *Brit. J. Criminol.*, 37:608–634.

Sutton, R. M. and Farrall, S. (2005). Gender, socially desirable responding and the fear of crime. Are women really more anxious about crime? *Brit. J. Criminol.*, 45:212–224.

Taylor, M. F., Brice, J., Buck, N., and Prentice-Lane, E. (2009). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. University of Essex, Colchester.

Theus, M. (2008). High-dimensional data visualization. In Chen, C.-h., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 151–178. Springer, Berlin / Heidelberg.

Thompson, C. Y., Bankston, W. B., and St. Pierre, R. L. (1992). Parity and disparity among three measures of fear of crime: A research note. *Deviant Behav.*, 13:373–389.

Tourangeau, R. and Ting, Y. (2007). Sensitive questions in surveys. *Psychol. Bul.*, 133:859–883.

University of Essex (2009). Institute for Social and Economic Research, British Household Panel Survey: Waves 1-17, 1991-2008 [computer file]. 6th Edition. Colchester, Essex: UK Data Archive [distributor], May 2009. SN: 5151.

University of Essex (2013). Institute for Social and Economic Research and NatCen Social Research, Understanding Society: Waves 1-3, 2009-2012 [computer file]. 5th Edition. Colchester, Essex: UK Data Archive [distributor], November 2013. SN: 6614.

van de Pol, F. and de Leeuw, J. (1986). A latent markov model to correct for measurement error. *Sociol. Methods Res.*, 15:118–141.

Visser, M., Scholte, M., and Scheepers, P. (2013). Fear of crime and feelings of unsafety in european countries: Macro and micro explanations in cross-national perspective. *Sociol. Quart.*, 54:278–301.

Warr, M. (1984). Fear of victimization: Why are women and the elderly more afraid? *Soc. Sci. Q.*, 65:681–702.

Warr, M. (1985). Fear of rape among urban women. *Soc. Probl.*, 32:238–250.

Warr, M. (1993). Fear of victimization. *The Public Perspective*, 5:25–28.

Warr, M. and Stafford, M. (1983). Fear of victimization: A look at the proximate causes? *Soc. Forces*, 61:1033–1043.

Wiley, D. E. and Wiley, J. A. (1970). The estimation of measurement error in panel data. *Am. Sociol. Rev.*, 35:112–117.

Winkel, F. W. (1981). Angst voor criminaliteit: Verklarende modellen. *Tijdschrift voor Criminologie*, 23:289–308.