Appendix 1: Exploratory and Confirmatory Factor Analysis, Demographics

Exploratory Factor Analysis (EFA): We randomly split the sample (wave 38) in two halves (of 4,250 cases) to allow separate exploratory and validation phases. Distributional assumptions (skew, kurtosis) were satisfied. Partly to enhance replicability, initially we used principal components analysis (PCA) with varimax and orthogonal rotation, by far the most popular method (Kim and Mueller, 1978; Kline, 2002). This method suggested a three factor solution when evaluated using: (a) variance explained (after rotation, the first factor explained 33.88% per cent of total variance, with subsequent factors explaining 15.53% and 12.64%); (b) cut-offs of 1 for Eigenvalues (does not always yield optimal solutions but was consistent with the three factor solution); and (c) the scree plot. Using conventional tests of suitability for EFA, Bartlett's sphericity test was highly significant, the KMO measure of sampling adequacy well above a suggested cut-off of 0.5, at 0.953.Assuming factors are uncorrelated (orthogonal rotation) is less realistic, and PCA is not always considered factor analysis as it does not discriminate between shared and unique variance of manifest variables (Costello and Osborne, 2005; Floyd and Widaman, 1995). However, these limitations tend to be exaggerated in studies with smaller samples. There is enough in common among various approaches that a clear structure is likely to be identifiable with any method used appropriately. As an additional step we used a variety of other methods of extraction and oblique rotation. The same three factor solution emerged. Table 1 shows grouping of variables (see Table 1).

Confirmatory Factor Analysis (CFA): After reducing the item pool (as explained in the paper), we tested the model from our EFA specifying a three factor structure. Having found support for this, we tested and rejected three competing models: a single factor structure, a two factor structure (where we allocated variables based on another EFA, specifying a 2 factor solution), and a version that addressed potential misspecification in error terms and dropped one low-loading variable (patrol). We used a range of common indices (excluding some chi-square based statistics which are vulnerable to sample size and would exaggerate fit). These indicated good fit for our 3-factor model, which used completely new scales and a very large, heterogeneous sample: GFI (0.931) and NFI (0.936) both above a suggested threshold of 0.9 (Byrne, 1994); CFI (0.94) just above a suggested threshold of 0.93 (Byrne, 1994), PCFI (0.750) below a suggested threshold of 0.8; RMSEA (0.079) just below 0.8 (Fan, Thompson and Wang, 1999). As a brief illustration, comparing the single factor / two factor models, fit indices were far worse, e.g.: GFI (0.738 single factor model / 0.800 two factor model), CFI (0.705 / 0.778). For a modified version we examined indices to look at covariances between error terms on the same factor and found some evidence of this in three paired terms: e2 <> e3, e7 <> e8, e1 <> e4. Allowing these to covary and removing the comparatively low-loading patrol variable led to fit improvements: GFI (0.964), NFI (0.967), CFI (0.971), RMSEA (0.061) but at the expense of parsimony (PCFI worsened to 0.625). We judged these were modest improvements so retained the framework in Figure 1. We were cautious of making post hoc changes since this would undermine our claims to be validating a model generated in the EFA, and in effect be introducing a second, exploratory stage which is counter to the logic of CFA. To replicate our findings we carried out two further waves of fieldwork with larger samples (because we did not split these into exploratory and validation phases). Fit indices improved, supporting the structure set out in figure 1: for wave 38 GFI was 0.971, for wave 39 GFI was 0.972. Also, in wave 39 we ran a second CFA model, this time imputing missing values (Acock, 2005; Groves, Dillman, Eltinge and Little, 2002) for “don't knows” to see if these were potential sources of systematic bias. These are not necessarily “missing data” as some answering will have had no police contact. Model structure was extremely similar (comparing path coefficients) as was model fit. (We did not include “don't knows” in correlation analyses, using pairwise deletion, but had over 6,000 responses for each bivariate correlation.)

Gender, Ethnicity, Age: *Gender* was binary, coded by interviewers (though for many, gender is neither binary nor identifiable from appearance). *Ethnicity* was self-chosen from 17 categories widely in use in social surveys in the UK and which we aggregated to 4 sub-groups used in the same schema: “White”, “Mixed”, “Asian”, “Black” (else numbers in some categories were too small to test). Even after aggregating, and though the sample was representative, the “Mixed” (n=158, 1.9%) and “Black” groups (n=486, 5.8%) were much smaller in relation to “Whites” (n=6,175, 73.4%) and “Asians” (n=1,412, 16.8%) so we were cautious interpreting differences across categories (some respondents do not report ethnicity so percentages do not sum to 100). *Age* was self-reported (“What was your age on your last birthday?”). More procedures are possible with a ratio variable, we used a 65+/under 65 group to allow for easier comparison across gender, ethnicity, age. As mentioned “p” values are not helpful with such a large overall sample for interpretation, so we used mean scores in evaluating differences. Alphas for each sub-scale were high across *gender* (Presence 0.736 female, 0.727 male; Fairness 0.897 female, 0.901 male; and Trust both 0.876) *ethnicity* (Presence between 0.714 and 0.763; Fairness between 0.892 and 0.932; and Trust between 0.869 and 0.901) and *age* (Presence 0.723 65+, 0.730 under 65; Fairness 0.877 65+, 0.901 U65; and Trust 0.847 65+, 0.879 U65), evidence the measures were generalisable and stable. For *gender*, we found slightly higher ratings across the 3 factors among women (Trust - 5.5355 vs 5.4443 ; Fairness 5.2981 vs 5.2298 ; Presence 4.9666 vs 4.9182). That there were no stronger differences is in line with existing literature on the relationship between gender and attitudes to policing, where findings are mixed (see Jang, Joo and Zhao for a review). For *ethnicity* we found higher ratings for “Asians” on both our Trust factor (mean of 5.543 vs means of 5.5196 for “Whites”, 5.2534 for “Mixeds”, 5.2739 for “Blacks”) and our Presence factor (mean of 5.0647 vs means of 4.9078 for “Whites”, 4.6859 for “Mixeds”, 5.0225 for “Blacks”). We found higher ratings for “Whites” on our Fairness factor (mean of 5.3268 vs means of 4.7470 for “Mixeds”, 5.1586 for “Asians”, 4.8670 for “Blacks”). Many studies find majority ethnic groups rate the police more favourably but it is outside the scope of the paper to make inferences about these differences or their generalisability to other settings, other than to note that it is common for much higher differences to be reported. For *age*, we found higher ratings among 65+ for Trust and Fairness but lower ratings for Presence (Trust - 5.5043 for under 65s vs 5.6396 over 65s; Fairness - 5.2545 for under 65s vs 5.5349 over 65s ; Presence 5.0232 for under 65s vs 4.9284 over 65s).