
Dmitriy Poznyak, George F. Bishop
University of Cincinnati, USA

Bart Meuleman
University of Leuven, Belgium
ANES Political Trust Scale

People have different ideas about the government in Washington. These ideas don’t refer to democrats or republicans in particular, but just to government in general. We want to see how you feel about these ideas:

1. DO WHAT IS RIGHT: “How much of the time do you think you can trust the government in Washington to do what is right—[1] just about always, [2] most of the time or [3] only some of the time?” ([4] None of the time—is a volunteered response but is also recorded).

2. INTERESTS: “Would you say the government is [1] pretty much run by a few big interests looking out for themselves or that [2] it is run for the benefit of all the people?”


Trust Govt.: “Just about Always” & “Most of the Time”

Change in Levels of Trust
- **75%** in **1964**
- **15%** in **2011**

Trust exhibits large short-term ebbs and flows.

If respondents interpret trust question (s) differently at different points in time, can we compare trust over time?

**Objective:** Test whether political trust items can be validly compared across all waves of ANES.

**Response:** sum of the “Just about always” and “Most of the time” answers to the question: “How much of the time do you think you can trust the government in Washington to do what is right?”
**Issue of Measurement Comparability**

- Brady (1985: 269) “…lack of interpersonal comparability of survey responses” was a greatly neglected and “…serious difficulty… largely ignored by social scientists.”

- “Although the language of the survey questions can be standardized, there is no guarantee that the meaning assigned to the questions is constant over respondents (Groves 1989: 449).”

- Bishop & Mockabee (2011): “A corollary to the comparability principle is that a survey question should mean the same thing to respondents at time two that it did at time one.”

- Before constructs, such as political trust, can be validly compared over time, scale needs to possess characteristic of measurement equivalence (ME)
Concept of Measurement Equivalence (ME)

- Often have several groups “nested” in our analyses:
  - Cross-cultural groups: cultures, regions, countries (WVS, ESS)
  - Cross-time-groups: time-points (ANES, GSS)

- Typically, our goal is to compare attitudes across groups

- But the fundamental question is whether “the meaning assigned to the questions is constant over respondents” (Groves 1989)...
  - Same stimulus for respondents within a sample
  - Constant stimulus for respondents across samples

- In its most general form, ME addresses the question of whether respondents from different samples or groups are responding to a given set of survey items in the same or essentially similar way.
Implications for Trust Research

• If people **conceptualize trust differently** at different time-points, then:
  – People with the **same** score on the latent **trust** variable will have **different** scores on the observed indicators
  – If true, levels of trust may not be **comparable** over time

• If meaning-and-interpretation is **not invariant** across groups, then findings of differences b/w individuals across groups cannot be unambiguously interpreted:
  – Will observed differences be **true** differences or artifacts of **change** in meaning-and-interpretation?

• Equivalence should **not** be taken for granted, but is an **hypothesis that needs to be tested empirically**

• **MGCFA** is the most prevalent statistical tool
Types of ME equivalence

1. **Configural** (conceptual) equivalence is the most basic level
   - Whether there is same configuration of factor loadings across all groups
   - Easy to achieve but doesn’t guarantee score comparability across group

2. **Metric equivalence** implies that factor loadings are invariant over time points
   - One-unit increase on the measurement scale has same meaning across all samples
   - Necessary condition for comparing effect parameters and covariances across groups or over time.

3. **Scalar equivalence** implies, in addition to invariant loadings, invariant intercepts
   - Respondents with same score on latent variable at different points in time have identical expected scores on observed variables
Types of Measurement Equivalence

- **Configural invariance**
  - Comparison scale
  - Measurement scale
  - Time 2
  - Time 1

- **Metric invariance**
  - Comparison scale
  - Measurement scale
  - Time 2
  - Time 1

- **Scalar invariance**
  - Comparison scale
  - Measurement scale
  - Time 1 – Time 2
ME Testing for Ordinal Items

- One of the assumptions of ME testing strategy is that the indicators follow a multivariate normal distribution.
- Trust items, however, are measured using ordered-categorical scales.
- Non-continuous indicators violate assumption of multivariate normality.
- Disregarding categorical nature of Trust item is not an option:
  - Items only have 2-to-3 response options
  - Exhibit high degree of skewness (distrust)
- ME testing for ordered-categorical data involves some notable differences.
Measurement model

\( \xi_1 \) (POLITICAL TRUST)

\( \lambda_1 \), \( \Phi_1 \)

\( \tau_1 \)

\( \delta_1 \)

\( \chi_1^* \) (DO WHAT IS RIGHT)

\( v_{11}, v_{12} \)

\( \chi_2 \) (INTEREST)

\( \lambda_2 \), \( \lambda_3 \)

\( \tau_2, \tau_3 \)

\( \delta_2, \delta_3 \)

\( \chi_3^* \) (WASTE)

\( \lambda_4 \)

\( \tau_4 \)

\( \delta_4 \)

\( \chi_4^* \) (CROOKED)

\( v_{21}, v_{31}, v_{32}, v_{41}, v_{42} \)

\( \chi_1 \): How much of the time do you think you can trust the government in Washington to do what is right—just about always, most of the time or only some of the time? (DO WHAT IS RIGHT)

\( \chi_2 \): Would you say the government is pretty much run by a few big interests looking out for themselves or that it is run for the benefit of all the people? (INTEREST)

\( \chi_3 \): Do you think that people in the government waste a lot of money we pay in taxes, waste some of it, or don’t waste very much of it? (WASTE)

\( \chi_4 \): Do you think that quite a few of the people running the government are crooked, not very many are, or do you think hardly any of them are crooked (CROOKED)

\( \xi \) (Ksi): Latent factor (Political Trust)

\( \kappa \) (Kappa): Latent mean

\( \Phi \) (Phi): Factor variance

\( \lambda \) (Lambda): Factor loading

\( \tau \) (Tau): Intercept

\( \delta \) (Delta): Variance

\( \chi \) (Chi): Observed variable

\( \chi^* \) (Chi): Latent response variable

\( v \) (Nu): Threshold; \( v = (C - 1) \), where \( C \) is the number of categories of the nominal item
Equivalence Testing for Categorical Indicators

**Software:** Mplus

**Model:** MGCFA for ordinal-categorical variables

**Estimator:** Weighted Least Squares (WLMSV) using theta parameterization

**Missing values:** Full Information Maximum Likelihood

**Results:**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>$\Delta$CFI</th>
<th>$\Delta$TLI</th>
<th>$\Delta$RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Configural equivalence</td>
<td>188.0</td>
<td>38</td>
<td>0.000</td>
<td>0.994</td>
<td>0.983</td>
<td>0.048</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Full threshold equivalence</td>
<td>1493.1</td>
<td>128</td>
<td>0.000</td>
<td>0.948</td>
<td>0.954</td>
<td>0.079</td>
<td>-0.046</td>
<td>-0.029</td>
<td>0.031</td>
</tr>
<tr>
<td>3. Partial threshold equivalence</td>
<td>526.4</td>
<td>105</td>
<td>0.000</td>
<td>0.983</td>
<td>0.984</td>
<td>0.048</td>
<td>-0.011</td>
<td>0.001</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Change in Response Pattern to ANES Trust Items: 1964-2008
Comparison b/w ANES Sum-Score and Latent Factor Score

\( \Delta \mu = -5.7; \ 95\% \text{-C.I.:} \ -22.4, 11.6 \)
## Comparison b/w scales

<table>
<thead>
<tr>
<th>Dependent variable: Political trust</th>
<th>Sum-score scale</th>
<th>Latent-factor score scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b</td>
<td>S.E.</td>
</tr>
<tr>
<td>Most important problem perception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economy (0/1)</td>
<td>-0.418</td>
<td>(1.971)</td>
</tr>
<tr>
<td>International (0/1)</td>
<td>3.815+</td>
<td>(1.983)</td>
</tr>
<tr>
<td>Government (0/1)</td>
<td>-7.757***</td>
<td>(2.108)</td>
</tr>
<tr>
<td>Labor issues (0/1)</td>
<td>-1.348</td>
<td>(3.540)</td>
</tr>
<tr>
<td>Environment (0/1)</td>
<td>-1.653</td>
<td>(2.209)</td>
</tr>
<tr>
<td>Public order (0/1)</td>
<td>-1.756</td>
<td>(1.996)</td>
</tr>
<tr>
<td>Racial problems (0/1)</td>
<td>4.602*</td>
<td>(2.178)</td>
</tr>
<tr>
<td>Social welfare (0/1)</td>
<td>-0.197</td>
<td>(1.979)</td>
</tr>
<tr>
<td>Contextual controls</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Policy mood</td>
<td>0.801***</td>
<td>(0.043)</td>
</tr>
<tr>
<td>Index of Consumer Sentiments</td>
<td>0.180***</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Presidential approval ranking</td>
<td>0.087***</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Congress approval ranking</td>
<td>0.034+</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Constant</td>
<td>-34.19***</td>
<td>(3.466)</td>
</tr>
<tr>
<td>N</td>
<td>21001</td>
<td></td>
</tr>
<tr>
<td>N of groups</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.070</td>
<td></td>
</tr>
<tr>
<td>Wald Chi2</td>
<td>1430.42***</td>
<td></td>
</tr>
<tr>
<td>Log-likelihood (null)</td>
<td>-149555.3</td>
<td></td>
</tr>
<tr>
<td>Log-likelihood (model)</td>
<td>-95064.04</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>190212.1</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>190546.1</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

• Because trust construct exhibits partial threshold equivalence over time, we can now model it as DV, using TSCS design.

• Partial threshold equivalence of trust over time ensures that respondents with same score on latent trust variable will have same score on observed variable.

• But latent variable means should be compared instead of sum indices, when only partial invariance is established.

• ANES sum-score undermines reliability and validity of trust index. It assumes that:
  – Observed items relatively free of random measurement error.
  – Trust Items have a stable meaning over time, and
  – Does not account for differential functioning of Waste and Crooked items.
Conclusions

• Modeling trust as latent factor scores (LFS) based on partially equivalent model is superior method because it:
  – Mitigates measurement error by differentiating error component from what is shared with latent factor
  – Reduces random error and improves parameter estimates & model fit statistics
  – Maximizes overall construct validity of the scale as items with highest loadings on latent trust factor have largest effect on factor scores.

• Using LFS approach potentially challenges numerous substantive conclusions in research literature based on standard sum-score scales, which we now know are significantly contaminated with measurement error.