An Item-level Method for Evaluating Measurement Invariance

Kendall Mather, Sara Weston, & David Condon
Department of Psychology, University of Oregon

Introduction

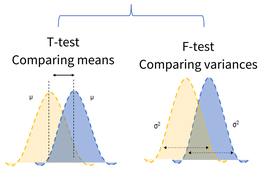
- Most analytical techniques for evaluating measurement invariance focus on detecting differences in responding at the scale-level.
- Recent work has called for a stronger focus on item-level personality assessment[1], but many existing item-level options (e.g., t-tests, F-tests) for detecting group differences are limited because they do not make use of all available information in the data (i.e., the entire distribution of responses).
- The present work compares responding across 12 nations on each of the 135 SAPA Personality Inventory items using the "DTS" test[2]. This test compares the entire distributions of two groups by taking the weighted integral of the difference between the empirical cumulative distribution functions for the two groups.

Comparing groups at the item-level

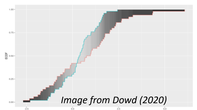
Most tests compare groups on only one parameter → *An alternative test compares* ***entire distributions***

T-test
Comparing means

F-test
Comparing variances



The DTS test takes the variance-weighted integral of the difference between groups



Method

- N = 25,598 Participants from the SAPA project (https://www.sapa-project.org/)
- Mean age 29.36 (*SD* = 11.75); 58% female; 31.85% USA

DTS comparisons will be run between….

12 nations (78 unique pairs):
- United States, Canada, Great Britain, Australia, Sweden, India, Germany, Netherlands, China, Singapore, Philippines, Norway.

135 items:
- SAPA Personality Inventory (SPI) items that are used to measure the big five and a set of 27 narrower traits.

For a total of **10,530 DTS tests** across item x nation pair combinations

Results

…36% of all DTS tests across items and nation pairs were significant

Effect sizes for all significant DTS tests across combinations of item x nation pair



Notes: This figure shows the effect sizes for all significant tests across combinations of SPI item (on the y axis) and nation pair (on the x axis). Items are ordered and color coded by SPI trait. In order on the y axis, these include: irritability, industry, humor, authoritarianism, impulsivity, art appreciation, perfectionism, attention seeking, well being, adaptability, charisma, compassion, honesty, trust, emotional stability, easygoingness, anxiety, sensation seeking, conservatism, intellect, emotional expressiveness, sociability, self-control, conformity, order, introspection, and creativity.

Average DTS stat for each big five and SPI trait

SPI 27 trait	Mean DTS	SD DTS	% p < .05
Conservatism	21.97	17.71	61%
Conformity	14.39	8.80	50%
Authoritarianism	14.25	9.40	50%
Self-Control	13.79	9.66	45%
Intellect	13.73	8.07	51%
Humor	13.71	9.74	47%
Attention Seeking	13.29	9.63	43%
Irritability	12.10	6.75	45%
Anxiety	12.00	6.65	40%
Charisma	11.94	7.07	41%
Industry	11.77	8.06	38%
Sensation Seeking	11.33	7.10	35%
Order	11.27	7.63	34%
Sociability	11.27	6.21	38%
Perfectionism	11.07	7.00	35%
Compassion	10.92	7.69	30%
Emotional Stability	10.61	5.78	30%
Introspection	10.46	6.11	28%
Creativity	10.40	7.11	27%
Honesty	10.20	6.34	27%
EasyGoingness	10.07	5.77	29%
WellBeing	9.96	6.32	23%
Trust	9.73	5.29	22%
Adaptability	9.71	6.11	25%
Art Appreciation	9.46	5.37	22%
Emotional Expressiveness	9.44	4.67	22%
Impulsivity	8.97	5.21	23%

Big five trait	Mean DTS	SD DTS	% p < .05
Conscientiousness	12.37	8.40	41%
Extraversion	12.13	8.06	41%
Neuroticism	11.32	6.56	35%
Openness	10.99	7.02	32%
Agreeableness	10.44	6.61	28%

Notes: This table displays the average DTS effect size across all tests within each big five and little 27 trait. That is, the tests comparing all pairs of nations (78 total) for each item included in the scale (5 for the SPI traits, and 14 for the big five traits). The other two columns show the standard deviation in effect sizes across these tests and the proportion of tests that were significant (p < .05). Traits are arranged in descending order of mean effect size.

Items with the most and fewest differences

Fewest differences

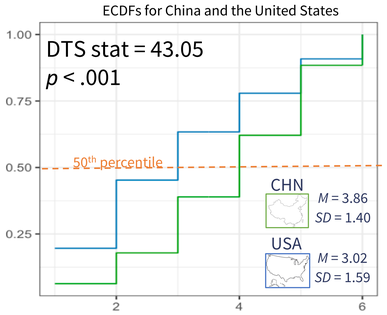
"Believe that others have good intentions" (trust)
"Am not bothered by disorder" (order)
"Adjust easily" (adaptability)
"Am open about my feelings" (emotional expressiveness)
"Let things proceed at their own pace" (easygoingness)
"Tell the truth" (honesty)
"Jump into things without thinking" (impulsivity)

Most differences

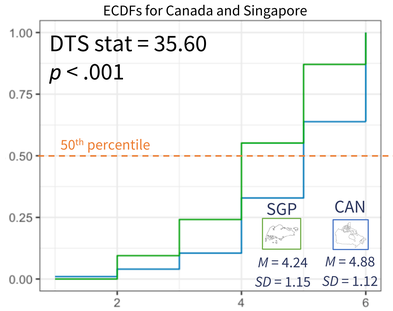
"Like to stand during the national anthem" (conservatism)
"Learn things slowly" (Intellect)
"See myself as an average person" (conformity)
"Believe in one true religion" (conservatism)
"Lack the talent for influencing people" (charisma)
"Fear for the worst" (anxiety)
"Don't consider myself religious" (conservatism)

Individual DTS test results for several openness-related items

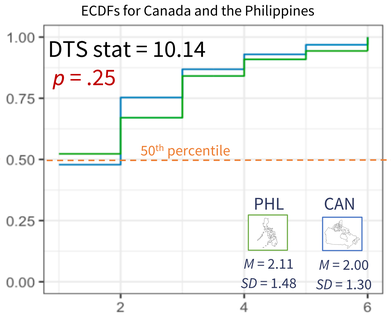
"I would hate to be considered odd or strange"



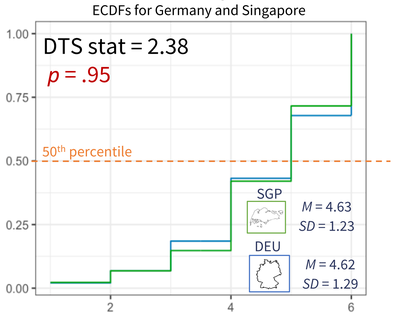
"I am full of ideas"



"I do not like art"



"I believe in the importance of art"



Notes: Each of the figures above shows the ECDFs for two nations on a particular item. Item response options—from 1 (strongly disagree) to 6 (strongly agree)—are shown on the x axis, and the percentile of respondents selecting each option on the y axis. P-values displayed in red indicate a non-significant effect. Note that the current version of the "twosamples" r package does not provide confidence intervals for DTS effect sizes.

Conclusions

- Item-level group differences that might otherwise be overlooked (e.g., due to similar means or variances) can be detected using more powerful tests that compare the entire distributions of the groups.
- Items used to measure the same broad trait are usually heterogeneous in content. Item-level statistical tests could contribute to understanding why some broad dimensions (e.g., openness to experience) are more relevant in some cultural contexts than others.
- When using these tests to identify items that show the most and fewest differences across nation pairs (rather than nation pairs that showed the most/fewest differences across items), it's important to use a highly representative selection of nations.

References

[1] Revelle, W., Dworak, E. M., & Condon, D. M. (2021). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, *169*, 109905

[2] Dowd, C. (2020). A new ECDF two-sample test statistic. *arXiv preprint arXiv:2007.01360*.

[3] Condon, D. M. (2022). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model