



PSYCHOLOGY AND PSYCHIATRY
REPLICATION
SUPPLEMENTARY-RESULT

Validating the behavioral Defining Issues Test across different genders, political, and religious affiliations

Hyemin Han 

Educational Psychology Program, University of Alabama, Tuscaloosa, Alabama, USA

Corresponding author. Email: hyemin.han@ua.edu

(Received 01 February 2023; Revised 18 February 2023; Accepted 21 February 2023)

Abstract

The Defining Issues Test (DIT) has been widely used in psychological experiments to assess one's developmental level of moral reasoning in terms of postconventional reasoning. However, there have been concerns regarding whether the tool is biased across people with different genders and political and religious views. To address the limitations, in the present study, I tested the validity of the brief version of the test, that is, the behavioral DIT, in terms of the measurement invariance and differential item functioning (DIF). I could not find any significant non-invariance at the test level or any item demonstrating practically significant DIF at the item level. The findings indicate that neither the test nor any of its items showed a significant bias toward any particular group. As a result, the collected validity evidence supports the use of test scores across different groups, enabling researchers who intend to examine participants' moral reasoning development across heterogeneous groups to draw conclusions based on the scores.

Key words: Defining Issues Test; differential item functioning; measurement invariance; moral reasoning

Introduction

The Defining Issues Test (DIT) is a widely used tool in the fields of moral psychology and education for evaluating the development of moral reasoning. Its primary purpose is to measure one's ability to apply postconventional moral reasoning when faced with moral dilemmas (Thoma, 2006). The DIT generates a P-score, a postconventional reasoning score, which quantifies one's level of postconventional reasoning development. The score reflects their likelihood of utilizing postconventional reasoning, which involves the ability to re-evaluate existing social norms and laws based on moral principles, rather than personal interests or social norms, across different situations (Rest et al., 1999).

Despite being a widely used tool in the field, concerns have been raised regarding the potential bias of the DIT toward individuals with varying gender, political, and religious affiliations. For instance, Gilligan (1982) argued that the model based on postconventional reasoning development might favor men versus women because women are more likely to be assessed to focus on personal interests, social relations in fact, in solving moral dilemmas from the DIT's perspective. In addition, some argue that the postconventional reasoning presented in the measure is liberal-biased, so conservative populations, including both politically and religiously conservative ones who value traditions and conventions, are also likely to be unfairly penalized due to their political and religious views, not by their actual developmental level (Crowson & DeBacker, 2008).

Several moral psychologists have suggested that people affiliated with different political and religious groups are likely to endorse different moral foundations and, thus, they are likely to render different moral

decisions based on different moral philosophical rationales (Graham et al., 2009). For instance, research on the moral foundations theory has demonstrated that liberals tend to endorse foundations for individualizing, whereas conservatives tend to focus on foundations for social binding (Graham et al., 2009). Hence, without examining whether a test for moral reasoning, the DIT, is capable of assessing one's moral reasoning in an unbiased manner across people with different moral views, it is impossible to assure that the test can generate reliable and valid outcomes across such people (Han et al., 2022b).

Previous studies have addressed these concerns by demonstrating that there have not been significant differences in the mean P-scores across the different groups (Thoma, 1986; Thoma et al., 1999), or P-scores significantly predict socio-moral judgment even after controlling for political and religious affiliations and views (Crowson & DeBacker, 2008). However, such score-based comparisons cannot address concerns regarding whether the test per se or its items are biased. To be able to address the concern, it is necessary to conduct: first, a measurement invariance (MI) test, which examines whether a test measures a construct of interest consistently across different groups (Putnick & Bornstein, 2016); and second, a differential item functioning (DIF) test based on the item response theory, which examines whether a specific item favors a specific group, while the latent scores are the same (Zumbo, 1999).

Once MI is supported and no item demonstrates a significant DIF across different groups, then it is possible to conclude that the items in the test do not measure one's latent ability unequally. Of course, there have been a few previous studies employing such methods to evaluate the cross-group validity of the DIT (Choi et al., 2019; Richards & Davison, 1992; Winder, 2009). However, their sample size was small, or they focused solely on specific groups (e.g., Mormons). Furthermore, the traditional DIT presented a technical challenge for conducting MI or DIF tests due to its scoring method, which uses rank-ordered responses instead of individual item ratings.

In the present study, to address the abovementioned limitations in the previous studies that examined the validity of the DIT across different groups, I tested the MI and DIF of the behavioral DIT (bDIT) with a large dataset collected from more than 1,400 participants. The bDIT, which is a simplified version of the traditional DIT, uses individual item responses to calculate one's P-score instead of rank-ordered responses (Han et al., 2020). Consequently, conducting MI and DIF tests with the bDIT is more straightforward. In contrast to previous studies, the present study analyzed a more extensive dataset collected from participants with diverse political and religious affiliations.

Methods

Participants and data collection

Data were acquired from college students (Age mean: 21.93 years; SD: 5.95 years) attending a public university in the Southern United States of America. All data collection procedures and the informed consent form were reviewed and approved by the University of Alabama Institutional Review Board (protocol number: 18-12-1842). Participants were recruited via the educational and psychological research subject pools. They signed up for the study and received a link to a Qualtrics survey form and received a course credit as compensation.

Table 1 summarizes the demographics of the participants in terms of their gender, political, and religious affiliations, which were the main interests of this study. Due to the convergence issue associated with Confirmatory Factor Analysis (CFA), only groups with $n \geq 100$ were used for the MI and DIF tests (Han et al., 2022a). As a result, for political affiliations, Republicans, Democrats, Independents, and Others were analyzed, and, for religious affiliations, Catholics, Evangelical and Non-Evangelical Protestants, Spiritual but not religious, and Others were analyzed.

Measures

The bDIT and demographics survey form used in the present study (survey.docx) and the codebook (varlist.xlsx) are available in the Open Science Framework repository at <https://osf.io/ybmp6/> for readers' information.

Table 1. Demographics information of participants

	<i>n</i>	%
Gender		
Female	1,145	85.32
Man	197	14.68
Political affiliation		
Republican	644	47.99
Democrat	293	21.83
Independent	201	14.98
Libertarian	49	3.65
Green Party	4	0.30
Others	151	11.25
Religious affiliation		
Catholic	272	20.27
Evangelical Protestant	231	17.21
Non-Evangelical Protestant	174	12.97
Spiritual but not religious	148	11.03
Agnostic	62	4.62
Atheistic	39	2.91
Jewish	22	1.64
Muslim	5	0.37
Buddhist	4	0.30
Hindu	1	0.07
Others	384	28.61

Behavioral Defining Issues Test

The bDIT consists of three dilemmas: Heinz Dilemma, Newspaper, and Escaped Prisoner (see survey.docx in the repository for the sample test form and items). For each dilemma, participants were asked to examine whether a presented behavioral option to address the dilemma is morally appropriate or inappropriate. Then, they were presented with eight items per dilemma asking the moral philosophical rationale supporting their decision. For each item, three rationale options were presented. Each of the three options corresponds to one of three schemas of moral reasoning proposed in the Neo-Kohlbergian model of moral development, personal interests, maintaining norms, and postconventional schemas. The participants were requested to choose the most important rationale.

Once the participants completed the bDIT, I examined how many postconventional options were selected out of 24 items (eight per dilemma \times three dilemmas). Then, one's P-score, which ranges from 0 to 100%, was calculated as follows:

$$P = \frac{\text{\#of selected postconventional options}}{24} \times 100.$$

For instance, if one selected the postconventional options as the most important rationale for 12 items, then the P-score becomes 50. It means that the likelihood of utilization of the postconventional schema while solving moral dilemmas is 50% in this person's case.

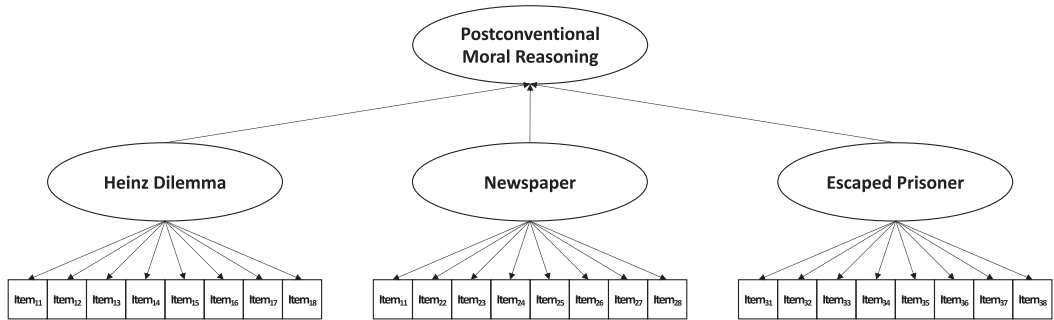


Figure 1. Measurement model of the behavioral Defining Issues Test for confirmatory factor analysis.

Demographics survey form

At the end of the survey, I presented a demographics survey form to collect participants' demographics for MI and DIF tests across different groups. The collected demographics include gender, political, and religious affiliations (see *survey.docx* in the repository for the demographics survey form).

Statistical analysis

First, I evaluated the MI of the bDIT via multigroup confirmatory factor analysis (MG-CFA) implemented in an R package, *lavaan* (Rosseel, 2012). Whether MI is supported was examined by the extent to which model fit indicators, that is, RMSEA, SRMR, and CFI, changed when additional constraints were added to the measurement model (Putnick & Bornstein, 2016). I tested four different levels of invariance: configural, metric, scalar, and residual invariance (refer to the [Supplementary Material](#) for additional methodological details). Because scalar invariance is minimally required for cross-group comparisons, I focused on whether this level of invariance was achieved (Savalei et al., 2015). In this process, I treated a response to each item as a dichotomous variable, that is, a postconventional versus non-postconventional, because that is consistent with how the actual P-score is calculated as described above. Then, I assumed a higher-order model with three latent factors, one latent factor per the presented dilemma (see [Figure 1](#) for the CFA model).

Second, I also performed the DIF test to investigate whether any item of the test demonstrated a statistically significant preference for a particular group compared to others, even when the latent ability, moral reasoning, was the same. To implement the DIF test, I employed the logistic ordinal regression DIF test with an R package, *lordif* (Choi et al., 2011) with an R code for the multiprocessing to distribute the tasks to multiple processors that was previously applied in Han et al. (2022a; 2022c). Once *lordif* was performed, I tested whether there was any significant uniform or nonuniform DIF for each item to examine whether the item significantly unequally favored one group versus other (see the [Supplementary Material](#) for methodological details).

All data and source code files are available in the Open Science Framework repository at <https://osf.io/ybmp6/>.

Results

The results from the MI tests demonstrate that residual invariance, the most restrictive invariance, was supported across genders and political and religious affiliations (see [Table 2](#)). In all cases, the changes in the fit indicators did not exceed the cutoff values. The results suggest that the equal measurement model, factor loading, intercept, and residual assumptions were satisfied, so the postconventional reasoning can be measured by the bDIT consistently across the groups.

Furthermore, the results from the DIF tests also support the point that all items in the bDIT did not significantly favor a specific group. Several χ^2 tests demonstrated significant outcomes, $p < .01$ (see the top

Table 2. Results from measurement invariance tests

	RMSEA	SRMR	CFI	Δ RMSEA	Δ SRMR	Δ CFI
Whole sample	.037	.039	.940			
By gender						
Configural invariance	.033	.042	.946			
Metric invariance	.029	.043	.956	-.004	.001	.010
Scalar invariance	.030	.043	.953	.000	.001	-.003
Residual invariance	.029	.043	.952	.000	.000	-.001
By political affiliation						
Configural invariance	.032	.051	.943			
Metric invariance	.031	.056	.944	-.001	.006	.001
Scalar invariance	.030	.058	.945	-.001	.001	.001
Residual invariance	.031	.060	.936	.001	.002	-.009
By religious affiliation						
Configural invariance	.035	.058	.932			
Metric invariance	.031	.063	.944	-.004	.005	.012
Scalar invariance	.026	.064	.957	-.005	.001	.014
Residual invariance	.026	.065	.953	.000	.001	-.004

panels of [Figures S1–S9](#) in the [Supplementary Material](#)). However, in all cases, both R_2 's and $\Delta\beta_1$'s were below the thresholds, .02 and .10, respectively (see the middle and bottom panels of [Figures S1–S9](#) in the [Supplementary Material](#) for R_2 and $\Delta\beta_1$ values, respectively). The results suggest that there was no item demonstrating a practically meaningful DIF.

Discussion

In the present study, I examined whether the bDIT can measure the development of postconventional moral reasoning across different gender, political, and religious groups consistently without bias. The MI test indicated that the bDIT assessed postconventional moral reasoning consistently across heterogeneous groups at the test level. At the item level, the DIF test reported that no item significantly favored a specific group. These results suggest that the bDIT was not biased across different groups.

Given that the bDIT did not show any significant non-invariance or DIF across different gender, political, and religious groups, it would be possible to conclude that the test can consistently examine moral reasoning. The results may address the concerns related to the potential gender and liberal biasedness in measuring postconventional moral reasoning. In the United States, in terms of political affiliations, Democrats are supposed to be more liberal and more likely to endorse individualizing moral foundations than Republications (Han et al., 2022b). In the case of religious affiliations, Evangelical Protestants are generally considered more conservative and more likely to support binding foundations than other religious groups (Sutton et al., 2020). In the present study, I examined the validity evidence of the bDIT among these groups with diverse political and religious views, which are inseparable from moral standpoints. Hence, moral psychologists and educators may employ the bDIT to test participants' developmental levels of moral reasoning.

However, there are several limitations to the present study that should be acknowledged. First, the study was conducted solely within the United States and, thus, further data should be gathered from a

more diverse range of cultural contexts while also taking into consideration different political and religious factors that exist within different countries. Second, the study relied on self-reported political and religious affiliations, which may not fully represent participants' actual political and religious views. Such variables are categorical and may not capture the complexities of an individual's beliefs accurately. Finally, while the present study tested the cross-group validity of the bDIT and the bDIT is a reliable and valid proxy for the original DIT (e.g., Choi et al., 2019; Han et al., 2020), it is necessary to examine whether the same level of validity can be supported for the original DIT, the DIT-1, and DIT-2. This can be achieved by administering the original DIT with large, diverse samples (e.g., Choi et al., 2020) and gathering additional demographic information.

Open peer review. To view the open peer review materials for this article, please visit <http://doi.org/10.1017/exp.2023.6>.

Supplementary materials. To view supplementary material for this article, please visit <http://doi.org/10.1017/exp.2023.6>.

Data availability statement. All data and source code files are available in the Open Science Framework repository at <https://osf.io/ybmp6/>.

Authorship contributions. H.H. conceived and designed the study, performed statistical analyses, and wrote the article.

Funding statement. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. The author declares none.

Ethical standards. The author asserts that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All data collection procedures and the informed consent form were reviewed and approved by the University of Alabama Institutional Review Board (protocol number: 18-12-1842).

References

- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*, 1–30. <https://doi.org/10.18637/jss.v039.i08>
- Choi, Y.-J., Han, H., Bankhead, M., & Thoma, S. J. (2020). Validity study using factor analyses on the Defining Issues Test-2 in undergraduate populations. *PLoS One*, *15*, e0238110. <https://doi.org/10.1371/journal.pone.0238110>
- Choi, Y.-J., Han, H., Dawson, K. J., Thoma, S. J., & Glenn, A. L. (2019). Measuring moral reasoning using moral dilemmas: Evaluating reliability, validity, and differential item functioning of the behavioural Defining Issues Test (bDIT). *European Journal of Developmental Psychology*, *16*, 622–631. <https://doi.org/10.1080/17405629.2019.1614907>
- Crowson, H. M., & DeBacker, T. K. (2008). Political identification and the Defining Issues Test: Reevaluating an old hypothesis. *The Journal of Social Psychology*, *148*, 43–60. <https://doi.org/10.3200/SOCP.148.1.43-60>
- Gilligan, C. (1982). *In a different voice*. Harvard University Press. <https://doi.org/10.2307/2067520>
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046. <https://doi.org/10.1037/a0015141>
- Han, H., Blackburn, A. M., Jeftić, A., Tran, T. P., Stöckli, S., Reifler, J., & Vestergren, S. (2022a). Validity testing of the conspiratorial thinking and anti-expert sentiment scales during the COVID-19 pandemic across 24 languages from a large-scale global dataset. *Epidemiology and Infection*, *150*, e167. <https://doi.org/10.1017/S0950268822001443>
- Han, H., Dawson, K. J., & Choi, Y.-J. (2022b). Testing the consistency of the moral growth mindset measure across people with different political perspectives. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000710>
- Han, H., Dawson, K. J., Thoma, S. J., & Glenn, A. L. (2020). Developmental level of moral judgment influences behavioral patterns during moral decision-making. *The Journal of Experimental Education*, *88*, 660–675. <https://doi.org/10.1080/00220973.2019.1574701>
- Han, H., Dawson, K. J., Walker, D. I., Nguyen, N., & Choi, Y.-J. (2022c). Exploring the association between character strengths and moral functioning. *Ethics & Behavior*, 1–18. <https://doi.org/10.1080/10508422.2022.2063867>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Rest, J. R., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: A Neo-Kohlbergian approach*. Lawrence Erlbaum Associates.
- Richards, P. S., & Davison, M. L. (1992). Religious bias in moral development research: A psychometric investigation. *Journal for the Scientific Study of Religion*, *31*, 467. <https://doi.org/10.2307/1386857>

- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, **48**, 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Savalei, V., Bonett, D. G., & Bentler, P. M. (2015). CFA with binary variables in small samples: A comparison of two methods. *Frontiers in Psychology*, **5**, 1515. <https://doi.org/10.3389/fpsyg.2014.01515>
- Sutton, G. W., Kelly, H. L., & Huver, M. E. (2020). Political identities, religious identity, and the pattern of moral foundations among conservative Christians. *Journal of Psychology and Theology*, **48**, 169–187. <https://doi.org/10.1177/0091647119878675>
- Thoma, S. J. (1986). Estimating gender differences in the comprehension and preference of moral issues. *Developmental Review*, **6**, 165–180. [https://doi.org/10.1016/0273-2297\(86\)90010-9](https://doi.org/10.1016/0273-2297(86)90010-9).
- Thoma, S. J. (2006). Research on the Defining Issues Test. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 67–91). Psychology Press.
- Thoma, S. J., Narvaez, D., Rest, J., & Derryberry, P. (1999). Does moral judgment development reduce to political attitudes or verbal ability? Evidence using the Defining Issues Test. *Educational Psychology Review*, **11**, 325–341. <https://doi.org/10.1023/A:1022005332110>
- Winder, D. R. (2009). *Macromorality and mormons: A psychometric investigation and qualitative evaluation of the Defining Issues Test-2*. Utah State University.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Department of National Defense, Directorate of Human Resources Research and Evaluation.

Cite this article: Han H (2023). Validating the behavioral Defining Issues Test across different genders, political, and religious affiliations. *Experimental Results*, **4**, e6, 1–11. <https://doi.org/10.1017/exp.2023.6>

Peer Reviews


Reviewing editor: Dr. Teresa Ober

University of Notre Dame, Department of Psychology, E418 Corbett Family Hall, Notre Dame, Indiana, United States, 46556

Minor revisions requested.

doi:10.1017/exp.2023.6.pr1

Review 1: Validating the Behavioral Defining Issues Test across Different Genders, Political and Religious Affiliations

Reviewer: Meghan Bankhead 

Date of review: 15 February 2023

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Conflict of interest statement. Reviewer declares none.

Comment

Comments to the Author: GENERAL COMMENTARY:


This work addresses a long-standing critique of the DIT — it is biased in favor of certain groups. It addresses this critique well, using a large and heterogenous sample. I do wonder if readers would be interested in the extent to which the present results can be extended to the DIT-1 and DIT-2. And if not, might the author suggest that future research should conduct similar studies with these iterations of the instrument (using large & heterogeneous samples)?

VERY MINOR EDITS:


Wording of the first paragraph of introduction is a bit difficult to follow.

Score Card

Presentation

	Is the article written in clear and proper English? (30%)	5/5
	Is the data presented in the most useful manner? (40%)	5/5
	Does the paper cite relevant and related articles appropriately? (30%)	5/5

Context

	Does the title suitably represent the article? (25%)	5/5
	Does the abstract correctly embody the content of the article? (25%)	5/5
	Does the introduction give appropriate context? (25%)	5/5
	Is the objective of the experiment clearly defined? (25%)	5/5

Analysis



Does the discussion adequately interpret the results presented? (40%)

5/5


Is the conclusion consistent with the results and discussion? (40%)

5/5

Are the limitations of the experiment as well as the contributions of the experiment clearly outlined? (20%)

5/5

Review 2: Validating the Behavioral Defining Issues Test across Different Genders, Political and Religious Affiliations

Reviewer: Dr. Kimberly F. Colvin 

SUNY Albany - Downtown Campus, Albany, New York, United States, 12203-1094

Date of review: 16 February 2023

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Conflict of interest statement. Reviewer declares none.

Comment

Comments to the Author: The manuscript Validating the Behavioral Defining Issues Test across Different Genders, Political and Religious Affiliations, documents the measurement invariance and differential item functioning analyses of the Behavioral Defining Issues Test (bDIT).

In the Abstract, the authors state that “the test is valid.” A test by itself is not valid, but the authors could claim that the validity evidence collected supports conclusions based on scores from the bDIT.

The statistical and psychometric procedures were appropriate, well-documented and appear to be conducted correctly.

The text around the DIF analyses implied that the existence of DIF is the same is bias, this is not accurate.

The discussion is confusing. A few examples of confusing phrases: “Democrats are supported to be more liberal” and “more likely to support binding foundations.”

Minor points: Throughout the manuscript there are several places where there are extra words, missing words, or incomplete sentences. Some sections were difficult to interpret. The formatting of Table 1 makes it difficult to read; it’s hard to tell which groups are / are not subgroups. Each page is labeled as “1.”

Score Card

Presentation



Is the article written in clear and proper English? (30%)

2/5

Is the data presented in the most useful manner? (40%)

4/5

Does the paper cite relevant and related articles appropriately? (30%)

5/5

Context



Does the title suitably represent the article? (25%)

5/5

Does the abstract correctly embody the content of the article? (25%)

3/5

Does the introduction give appropriate context? (25%)

4/5

Is the objective of the experiment clearly defined? (25%)

5/5

Analysis



Does the discussion adequately interpret the results presented? (40%)

3/5

Is the conclusion consistent with the results and discussion? (40%)

3/5

Are the limitations of the experiment as well as the contributions of the experiment clearly outlined? (20%)

4/5