Fairness of Protestant work ethic scale items among male and female gender

Investigating if the items show bias towards the male gender

Sumbul Jafri

PSY121 - Psychological assessment and diagnostics
Summer 2022

Prof. Dr. Andrea Hildebrandt

ABSTRACT

Protestant Work Ethic (PWE), a personality trait coined by Max Weber in 1905, expounded the relationship between Protestant cultural values of labor and discipline, and the economic productivity of Northern European countries as compared to the Catholic Southern European ones. Herbert Mirels and James Garret published the Protestant Work Ethic Scale in 1971, focusing on the concept of Protestant Work Ethic on an individual level rather than a consequence of culture.The religious connotations were replaced with the description of people who prioritize work in their lives, hold admiration for hard work and disdain for leisure (Furnham, 1990b). The occupational monopoly of men and the social constraints faced by women in various professions have changed over the years. As a result, each item on this scale required evaluation to verify that men and women with similar levels of PWE hold an equal probability of response endorsement (Plouffe et al., 2021). To assess the test items for a bias toward the male gender, the item difficulty, discrimination power, and reliability of the items across genders were investigated to identify any potential deficits in measurement invariance. The objective of this report was to employ Differential Item Functioning (DIF) to determine whether there are true mean differences across test items for the two genders (male and female), and therefore assess the fairness of test items for the female gender. The rationale for the six items recognized as DIF has been addressed.

Keywords: Protestant work ethic, DIF, invariance, gender, bias

Max Weber (1905), a sociologist, postulated a causal link between the economic prosperity of Northern European countries and their Protestant ideals as capitalism flourished in the western civilization. The essential notion behind this theory was that Protestant Ethics supplied a moral justification for wealth generation and promoted the values of labor and discipline whereas the southern European countries that subscribed to the Catholic values of ceremony and confession preached denouncement of wealth. Although puritan theologians considered economic success as a sign of God's grace, they also saw capital as a source of ongoing temptation for self-indulgent behaviors (Mirels & Garrett, 1971). Fullerton et al. (1959), described the austere nature of the Protestant Ethic as,

"…the exhortation toward restless enterprise [which] was based on the view that disciplined work was the best prophylactic against what the Puritan called the "unclean life" against the sloth and sensuality which riches so often engender. Work in one's calling [was a] prescription against sexual temptation as well as against religious doubts [p. 16]."

Weber et al. (1958) defined two concepts: the "Protestant ethic" and the "spirit of capitalism." The first concept relates to austere Protestantism's professional ethic, which includes valuing work and striving to perform it well. And the second one refers to a set of attitudes and behaviors required for capital accumulation and success in the modern capitalistic world (Bendix, 1975; see also Grabowski et al., 2016; Kozyr-Kowalski, 1967). Work is vital to the spirit of capitalism because it aids in the accumulation of wealth, while the protestant ethic appraises the quality of work to the amount of wealth acquired (Grabowski et al., 2016). Therefore, both concepts accentuate each other's goals and are heavily intertwined.

People with a high Protestant work ethic attach a sense of moral obligation to hard work, making it a thing to revere. Work ethic can be defined as a framework of values about work and leisure. It is a behavioral paradigm centered on the importance of work; in other words, to work is a privilege, and leisure is denying oneself of that privilege (Furnham, 1990a; Grabowski et al., 2016; Mirels & Garrett, 1971; Mudrack, 1997). Work ethic is viewed as both a personality variable and as an element of culture. According to the protestant ethic, people who respect labor are opposed to idleness and despise squandering resources or being overindulgent.Their appreciation of hard work includes frugality, efficiency, deferment of gratification and being socially helpful (Cherrington, 1980; Furnham, 1990a; Grabowski et al., 2016; Miller et al., 2002).

Although the Protestant work ethic was initially believed to be a cultural trait, some psychologists attempted to explore it on an individual level in the 1960s. The religious connotations were replaced with the characterization of this construct's psychological meaning in terms of its relationships with other personality variables, and with occupational interests (Mirels & Garrett, 1971). It became a descriptive term for people who prioritize work in their lives, hold admiration for hard work and disdain for leisure (Furnham, 1990b; Zhang et al., 2012).

Herbert Mirels and James Garret's Protestant Work Ethic (PWE) scale was the first instrument based on the notion that ethic was a personality characteristic and

corresponded to the level of acceptance for the Protestant ethic philosophy. It consists of 19 items that were selected from a large number of statements using factor analysis.The authors initially obtained four highly correlated factors, but eventually reduced the measurement of Protestant ethic to a single dimension that had manifested itself strongly.This questionnaire was formerly a widely utilized instrument all across the world. The items of this scale show satisfactory internal consistency with Cronbach's α values that are greater than 0.70 (e.g. Mudrack, 1997; Grabowsk et al., 2016).

In the original scale, PWE was considered a unidimensional construct however there have been multiple claims as to its multi-dimensionality. Furnham (1990b) highlighted various interpretable elements as factors, including respect, admiration, willingness to participate in hard work, contempt for leisure, morality, religion, independence from others, asceticism, and the disadvantages of having too much time and money. Tang (1993) developed four factors using the PWE Scale produced by Mirels and Garrett (1971) and a sample of Taiwanese students: hard work, internal motive, asceticism, and attitudes toward leisure. McHoskey (1994) provided four criteria using the same scale: success, asceticism, hard work, and anti-leisure. Furthermore, Wentworth and Chell (1997) discovered five key components, such as a disdain for leisure, a person's use of time, hard work, rewards of work, and disdain for indolence. Jones Jr (1997) also classified the PWE in terms of five facets: hard work, time management, saving, innovation, and honesty. Arslan (2000) identified five factors as well: work as an end in itself, success brought on by hard work, money and time saving, internal locus of control, and unfavorable attitudes toward leisure (Zhang, et al., 2012).

Grabowski and colleagues (2016) evaluated the psychometric properties of the Polish version of the PWE scale and the four factors examined displayed a poor degree of reliability (due to the vagueness of certain statements), reducing the significance of the multidimensional version of the PWE scale.The one-factor model also gave a poor fit but was legitimized by the standardized root mean square residual (SRMR) values that were lower than or equal to 0.08 (Grabowski et al., 2016; Hu & Bentler, 1999). Nevertheless, they advised caution while considering the scale as uni-dimensional and therefore a sub-part of our test adaptation goal was to explore the scale's factor structure.

The Mirels and Garrett scale is a relatively dated measure of work ethic and the social attitudes towards genders have changed ever since. The degree of endorsement of PWE ideology was nearly identical for both men (£ = 85.7, SD = 15.5) and women (£ = 85.5, SD = 16.2) in the original study. However, given the change in the occupational monopoly of men over the last decades, an investigation into the measurement invariance of the scale across the two genders (male and female) was necessary. The test items could show a bias towards the male gender and to reveal such possible deficits in measurement invariance, item difficulty, discrimination power, and reliability of the items across genders was assessed using Differential Item Functioning (DIF). The main aim of this test analysis was, therefore, to investigate whether there are real mean differences across the two genders (male and female) and thus evaluate the fairness of test items towards the female gender.

METHODS

Objectives:

The purpose of this study was to use Item Response Theory and Differential Item Functioning to explore whether the item properties of the Protestant Work Ethic (PWE) scale (Mirels & Garrett, 1971) are invariant across the two genders (male and female). Therefore, it was necessary to empirically assess and define the scale's dimensionality as a sub-goal. An attempt was also made to provide an interpretation for the items flagged as biased by DIF concerning their relevance in today's time.

Subjects and Procedure:

The data used for this study was acquired from the Open Psychometrics website, wherein the total sample size was 1,350, with three gender categories (male, female and other). As there was a huge disparity between the data for male and female categories, and the 'other' category, for this specific analysis, data for the 'other' category was removed. Therefore, only male and female responses were considered (n = 1302) and dummy coded (Male: n = 675, code = 0, Female: n = 627, code = 1). Since the user's responses to the statements were the only values relevant for this study, they were stored in a new variable, and reverse items (Q9, Q13, and Q15) were recoded.

Protestant Work Ethic Scale (Mirels & Garrett, 1971):

It has 19 statements either endorsing or not endorsing attitudes corresponding to the Protestant work ethics. Three values are collected for each question, i.e., the user's response on a five-point Likert scale (1 = Disagree, 5 = Agree), the position in the survey that the question was administered, and the time spent on each question in milliseconds. It has more than 30 demographic/general variables. An optional survey with questions from the Ten Item Personality Inventory (Gosling et al., 2003) is also presented after the test, with a validity check.

Data analyses

All statistical analyses were computed in the R environment.

Dimensionality Assessment:

This is an important step before fitting an IRT model, the dimensionality of the scale can be determined by using either of the three methods -

1. Categorical Principal Component Analysis (Princals)

It is a data reduction technique appropriate for categorical data and useful when researchers are interested in identifying the underlying components of a scale or a test while maximizing the amount of variance accounted for by its items (or the principal components).

Princals is accessed through the **Gifi** package in R.

2. Exploratory Factor Analysis (EFA) using polychoric correlations

EFA is a classical approach for locating latent variables. It aids in the description and identification of the number of latent constructs or factors that affect responses to observed variables or indicators. It is different from PCA as it creates a statistical model including an error term (unique factors) and PCA uses a simple linear combination to characterize the principal components (Suhr, 2005). EFA attempts to explain covariance in data, whereas PCA attempts to explain variance in data (although it does not completely dismiss the covariances). EFA differs from confirmatory factor analysis in that each indicator loads on each component individually, and the factors are often uncorrelated. In contrast to CFA, where the user choose which indicator is linked with which factor based on some underlying substantive theory (Mair, 2018).

Implemented in the **psych** package (Revelle W, 2022) in R.

3. Item Factor Analysis (IFA)

Item Factor analysis is another way of performing factor analysis without correlation computations. It analyzes item-level response data and is used to summarize the dependence structure of a set of categorical variables by a small number of latent variables (Chen & Zhang, 2020). Exploratory IFA was used in the study to confirm the uni-dimensionality assumption. This kind of IFA is based on the assumption that the user has little or no prior understanding of the data. Through exploratory study processes, it seeks to understand the latent structure underlying the data.

All three methods were performed to adapt the scale to a uni-dimensional factor structure because the literature and the output of the statistical analysis supported this claim (see Results) .

Item Response Theory (IRT):

Item response theory is a psychometric modeling technique used for analyzing categorical data from tests and other instruments. It is used to examine the relationships between latent variables and item responses (DeMars, 2010; Embretson & Reise, 2013; Tang, 1993). It has numerous advantages over Classical Test Theory, the most important of which is that it models outcomes at the item level rather than the test level. In terms of the information it offers on test performance, IRT is more complicated and extensive. As a result, establishing an IRT framework will allow for a more accurate invariance analysis across gender for PWE.

Partial credit model (PCM):

It was used to confirm the factor structure and analyze the psychometric properties of the adapted version of the PWE scale after removing four items.

Masters (1982) developed the partial credit model (PCM; a unidimensional latent trait model for both dichotomous and polytomous response types) as an extension of the Rasch Model for a more comprehensive estimation of a person's ability than a simple pass/fail score (source). It was designed for instances involving partial credit (e.g., 0 = "Disagree," 1 = "Partially disagree," 2 = "Partially agree," 3 = "Agree") in which the scores are reported on an ordinal scale. Items do not need to have the same amount of categories unlike the Rating Scale Model, and each item has its own set of item category parameters. It estimates person locations, item difficulty, and thresholds specific to each item (Glas & Verhelst, 1989).

The package **eRm** in R was used to fit PCM.

Differential Item Functioning (DIF):

Differential Item Functioning analysis with IRT allow for parameter group comparisons by assessing the consistency of item discrimination and item difficulty parameters across groups/time (Zhang, et al., 2012). According to Osterlind and Everson (2009), the differences in the item parameters can be interpreted as displaying bias in that item.The notion of fairness is linked to this bias. DIF analysis aims to identify (i.e., "flag") such biased items (Mair, 2018). When individuals with the same latent ability from different groups have differing likelihood of responding to the same item, they are said to have DIF. Non-DIF items are those in which individuals with the same latent ability have an equal opportunity of responding to an item regardless of their group membership (see also Columbia et al., 2019; Holland & Wainer, 1993). DIF items can then be evaluated and removed by experts during the development of the scale. The person parameters for these items can also be adjusted without omitting them if the purpose is to rate individuals on a well-established scale.The results of DIF can be divided into two forms:

Uniform DIF: the Item characteristic curves (ICC) are shifted across subgroups but stay parallel which is the group main effect. That is, uniform DIF reflects true group variations in the presentation of the underlying trait or attribute (Mair, 2018).

Nonuniform DIF: the ICCs are shifted between subgroups and cross, which demonstrates an interaction effect between the group and the trait. It represents errors or biases in the measurement process (Mair, 2018).

Logistic Regression DIF Detection (Zumbo, 1999) was used because the PWE scale has polytomous response scores. The fundamental idea behind this method is to compare the fit of a model that considers the factor loadings (discrimination) and thresholds (severity) parameters as the same (no DIF) to a model that permits them to differ (DIF) across groups (Columbia et al., 2019).

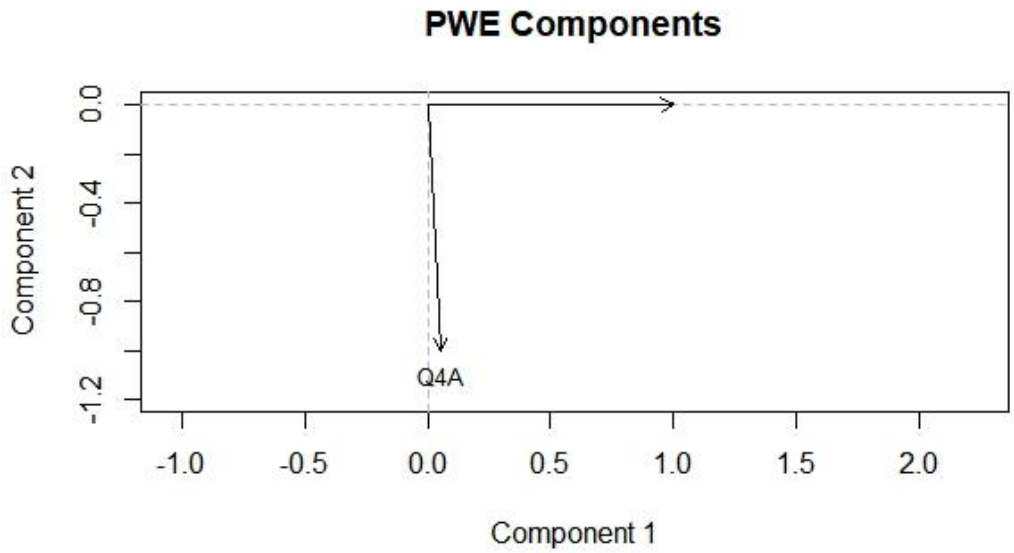
This approach is implemented in R using the **lordif** package, which incorporates empirical criteria defined in the Monte Carlo simulation (Choi et al., 2011). Specifically, this approach uses Stocking-Lord equating for item parameters and estimates person parameters using DIF and non-DIF items to fit a GRM (default) or a GPCM with **mirt** (Mair, 2018).

RESULTS

The results of the categorical principal component analysis (or Princals; see Fig 1.1) illustrate that a considerable portion of the variance in the PWE scale can be explained by a single component with one variable (Q4) projecting in the direction of the second component. However, the difference between the eigenvalues (i.e., the amount of variance retained by each principal component) is more than 3, implying uni-dimensionality of the scale with one outlier (Hattie, 1985).

Fig 1.1

Result of Princals



Note. The eigenvalue for component one (18.0319) was much higher than the eigenvalue for component two (0.9681)

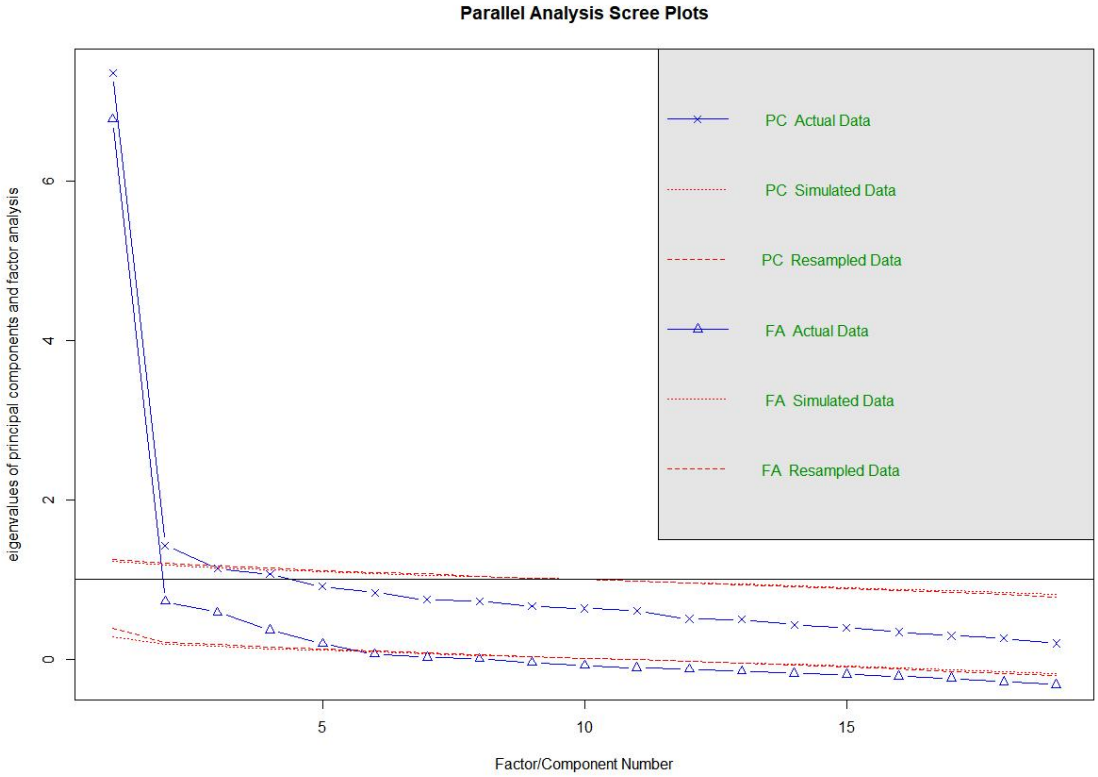
Princals assumes a linear mapping (that the data lives in a smaller-dimensional Euclidean space); many times, this doesn't match reality.
To get a more in-depth look at the factor structure of the scale, EFA based on polychoric correlations was done. The results supported the unidimensionality claim because more than 30% of the variance was explained by **one** factor (Grabowski et al., 2016), wherein 38.8% variance was explained by the first and only 7.8% variance was explained by the second factor.

Parallel analysis was used to rule out the idea that the reported dimensionality may be the result of chance. This analysis allows us to compare the EFA results to the average eigenvalues that would be generated by chance, enhancing the evidence for the previously reported dimensionality. A factor is deemed **significant** if its eigenvalue is greater than the 95% quartile (red line in the scree plot, see Fig 1.2) of those computed from random or resampled data (Mair, 2018). In this case, two components were above the 95% quartile with a strongly dominating first factor. The Velicer

Minimum Average Partial test (MAP; Velicer, 1976) was used to determine the number of interpretable factors. **One** and **three**-factor structures minimized the average partial correlations best (Henson & Roberts, 2006; Zwick & Velicer, 1986).

Fig 1.2

Result of Parallel analysis



Note. The scree plot displays the eigenvalues of the full solution as well as the eigenvalues of a random data matrix of the same size as the original that was generated using a parallel analysis. This plot suggests five factors according to the Kaiser criterion with a dominating first factor.

The suggestions of the various methods employed for the factor number don't always coincide because they express different characteristics of the factorial solution. Therefore, applying Item Factor analysis (IFA) was necessary to interpret the factor loading matrix of both one and three-factor solutions.

Table 1

PWE items and their loadings on the three theorized factors

PWE Items	Factor Loading		
	1	2	3
Factor 1: *Protestant Work Ethic*			
1. Most people spend too much time in unprofitable amusements.	**-0.45007**	0.31755	0.01964
3. Money acquired easily (e.g. through gambling or speculation) is usually spent unwisely.	**-0.56321**	-0.00762	0.06134
4. There are few satisfactions equal to the realization that one has done one's best at a job.	**-0.53399**	0.02444	0.11119
5. The most difficult college courses usually turn out to be the most rewarding.	**-0.46414**	-0.05577	0.07523
6. Most people who don't succeed in life are just plain lazy.	**-0.25550**	0.19430	0.42661
7. The self-made person is likely to be more ethical than someone who is born to wealth.	**-0.52214**	-0.16544	0.00375
8. I often feel I would be more successful if I sacrificed certain pleasures.	**-0.47068**	0.11410	-0.06962
11. People who fail at a job have usually not tried hard enough.	**-0.37852**	0.08062	0.33777
12. Life would have very little meaning if we never had to suffer.	**-0.52952**	0.12769	-0.03275
14. The credit card is a ticket to careless spending.	**-0.48547**	-0.09300	-0.14277
16. The person who can approach an unpleasant task with enthusiasm is the one who gets ahead.	**-0.39168**	-0.04905	0.34445
18. I feel uneasy when there is little work for me to do.	**-0.45109**	0.06228	0.04358
19. A distaste for hard work usually reflects a weakness of character.	**-0.58486**	0.14960	0.11556
Factor 2 : *Leisure*			
2. Our society would have fewer problems if people had less leisure time.	-0.24738	**0.56404**	0.05520

9. People should have more leisure time to spend in relaxation.	-0.00932	**0.90100**	-0.00727
15. Life would be more meaningful if we had more leisure time.	0.04365	**0.85124**	0.04253
Factor 3: *Hard work*			
10. Anyone who is able and willing to work hard has a good chance of succeeding.	0.02236	-0.02361	**0.92014**
13. Hard work offers little guarantee of success.	0.08702	0.11526	**0.65821**
17. If one works hard enough they are likely to make a good life for themselves.	-0.05879	0.00588	**0.87324**

[1]The classical guidelines for RMSEA evaluations, given by Browne and Cudeck (1993), consider values between **0.6-0.8** a fair fit (Mair, 2018), and Bentler and Bonett (1980) recommended that TLI greater than **0.90** indicates an acceptable fit.

Note. Rotated factor loadings: Since there were some reverse items in the scale, only the value of the loading is relevant, not the sign on it.

The results of IFA hint that the violation in uni-dimensionality may be caused by very similar or homogenous leisure (2,9 and 15) and hard work items (10,13 and 17) creating unwanted factors. All the other items are more heterogenous and have strong loadings on the general PWE factor. Therefore to adapt the scale to a one-factor structure, content-wise factor analysis was run to choose one hard work item and one leisure item that fit the one-factor model best. The items were chosen following a series of critical considerations based on their factor loadings, and the values of the Tucker Lewis Index (TLI) and RMSEA index when they were fit into a one-factor model. The best results were achieved with the inclusion of item 9 for mapping the individual's attitude on leisure and item 13 for hard work. This list of items also displayed a better RMSEA fit (**0.077** with 10 % CI [0.072, 0.082])[1] and TLI (**0.857**)[1] than the ones with other items. Thus, a unidimensional PWE scale was achieved after removing four items (2,10,15, and 17).

The Differential Item Functioning analyses flagged only **6** of the 15 PWE items of the modified unidimensional scale as biased across the two gender categories (male and female); these were all items that had higher loadings for the general PWE factor in the original 19 item version of the scale. The 9 remaining items displayed invariance across the two groups. The items are tagged depending on detection criteria ("Chisqr", "R2", or "Beta"). When the detection criteria is "Chisqr" , an item is flagged as DIF if any of the three likelihood ratio $\chi 2$ statistics are significant (Choi et al., 2011).

To obtain gender group-specific item discrimination and category location estimations that account for the items' differential functioning, the item parameters for all 15 items were recalculated (see Table 2). Category location parameters relate to the DIF-adjusted trait degree of agreement where respondents crossed the threshold from one response category to the next (i.e., from category 0 to 1, from 1 to 2, and so on). Category locations ranged from **-2.61** to **-0.79** at the first threshold, **-1.63** to **0.53** at the second threshold, **-0.99** to **0.91** at the third threshold, and from **0.72** to **1.72** at the fourth threshold (Vaughn-Coaxum et al., 2016).

Table 2

DIF item parameters

PWE Items	Male				
	a	cb1	cb2	cb3	cb4
I1.1	1.537	-1.500	-0.627	-0.083	1.020
I5.1	1.647	-0.507	0.311	0.717	1.696
I7.1	0.896	-2.602	-1.638	-0.994	0.721
I9.1	1.639	-1.127	-0.233	0.345	1.428
I14.1	0.925	-1.493	-0.712	-0.105	1.347
I15.1	2.344	-1.051	-0.462	-0.038	0.730

PWE Items	Female				
	a	cb1	cb2	cb3	cb4
I1.2	1.541	-1.238	-0.428	0.166	1.266
I5.2	2.117	-0.094	0.528	0.910	1.727
I7.2	0.921	-2.164	-1.002	-0.572	1.146
I9.2	1.773	-0.797	0.109	0.590	1.669
I14.2	1.081	-1.877	-0.894	-0.425	0.925
I15.2	1.762	-1.265	-0.600	-0.197	0.915

I1 - Most people spend too much time in unprofitable amusements.
I5 - Most people who don’t succeed in life are just plain lazy.
I7 - I often feel I would be more successful if I sacrificed certain pleasures.

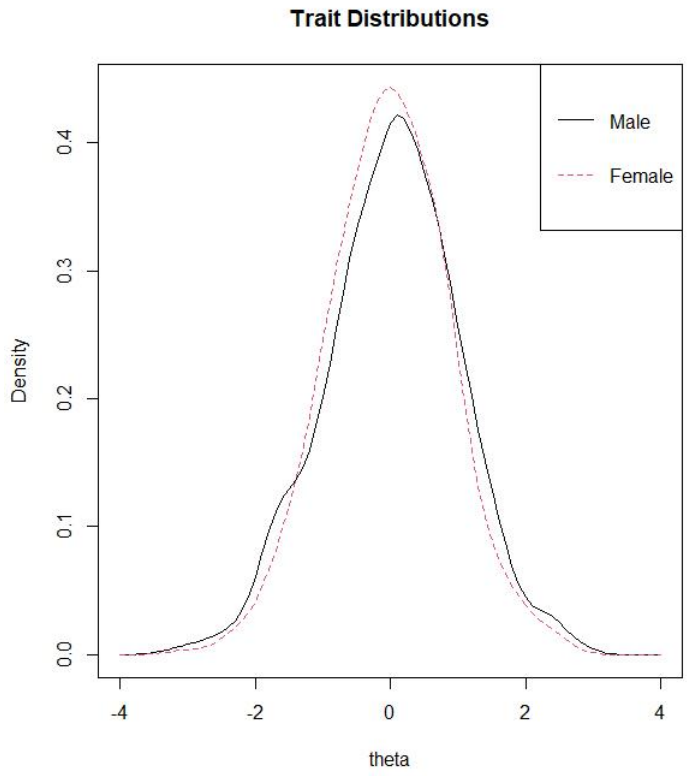
I9 - People who fail at a job have usually not tried hard enough.
I14 - I feel uneasy when there is little work for me to do.
I15 - A distaste for hard work usually reflects a weakness of character.

Note. I.1 and I.2 demonstrate the two gender categories (male and female)

The plot function in **lordif** displays the theta distributions for the female and male groups. The following figure shows that there is a significantly large overlap in the trait distributions between the groups (see Figure 2) but the male participants demonstrated a slightly lower peak than the female participants.

Figure 2

Trait distributions – female vs. male



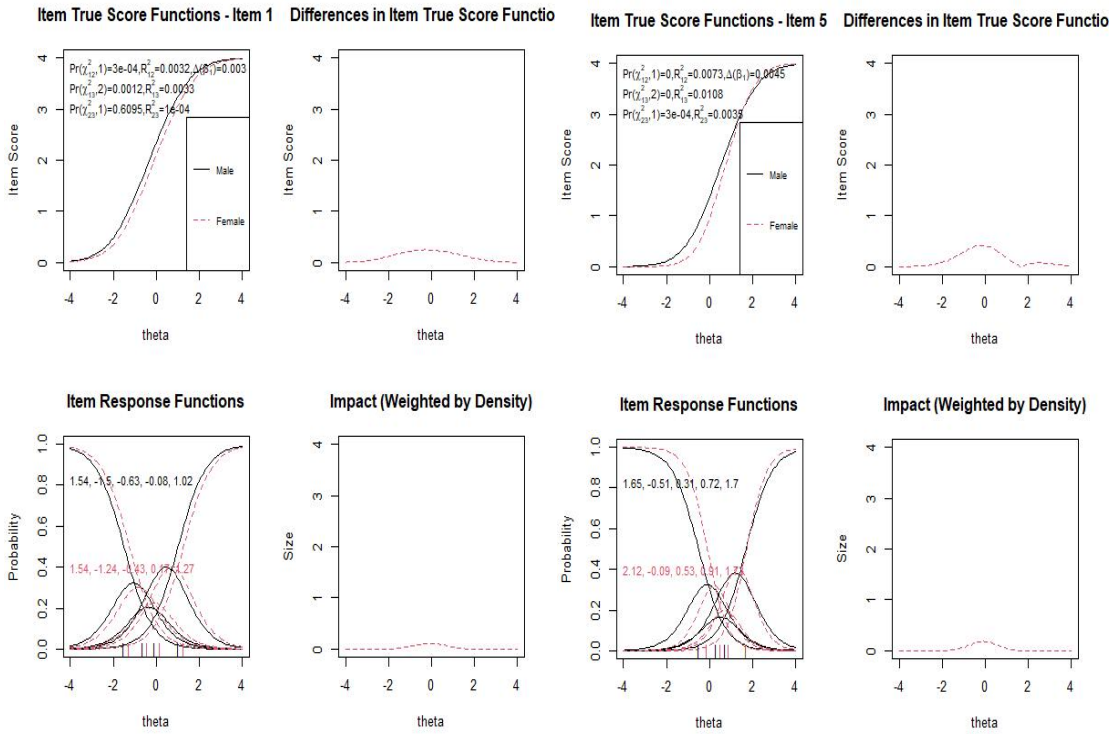
Note: Smoothed histograms of female (dashed line) and male (solid line) PWE levels as measured by the modified PWE scale (theta) are shown in this graph.

The Item characteristic curves (ICCs) and additional diagnostic graphs for each of the DIF items are shown in the figures below (see Figures 3.1 - 3.3). The graphs for each item consist of the female (dashed curve) vs. male (solid curve) ICCs on the top-left. The top–right plot displays the absolute difference in ICCs between the groups, which reflects the difference in PWE trait scores (theta). The item response functions for the two groups are depicted in the lower–left plot, based on demographic–specific item parameter estimations (slope and category threshold values by group). The lower right

graph illustrates the absolute difference between the ICCs (top–right graph) weighted by the score distribution for the targeted group, i.e., female participants (Choi et al., 2011). (For further information, see the text below the figures)

Figure 3.1

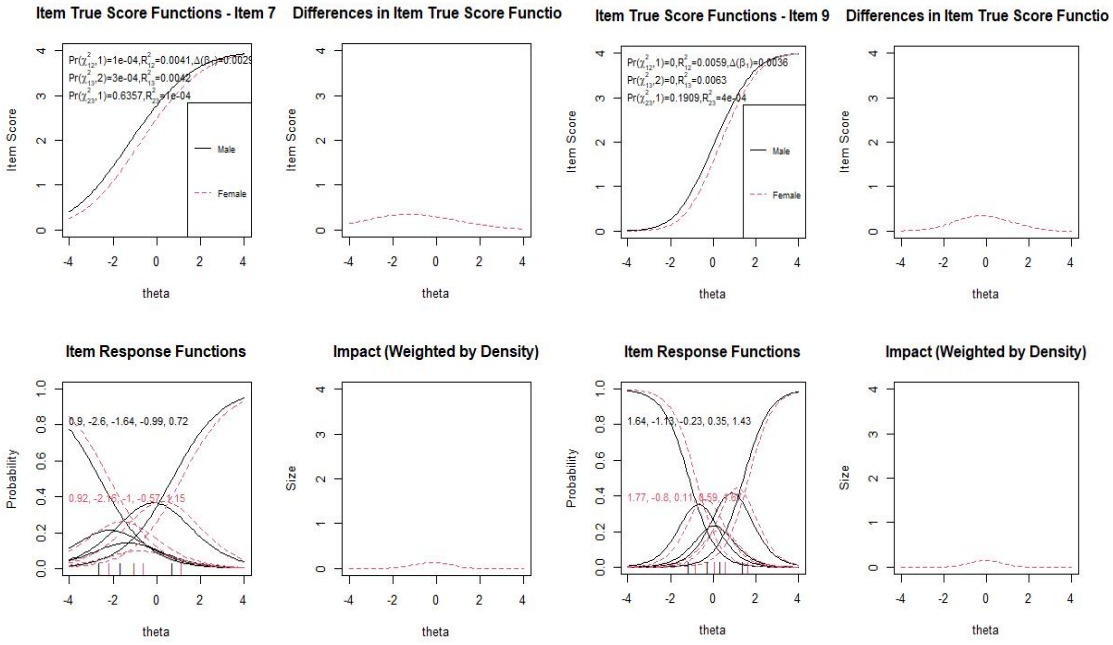
Item Characteristic curves for item 1 (left) and item 5 (right)



Note. The top left plot in each figure depicts item true-score functions based on item parameter estimations specific for each group. The slope for the male group was marginally higher in comparison to the female group (dashed curve), which might indicate non-uniform DIF. However, the LR $\chi 2$ test for uniform DIF, comparing Model 1 and Model 2, was significant (p = 0.0003), whereas the 1-*df* test for comparing Model 2 and Model 3 was not significant (p = 0.6095). So, it's a uniform DIF for item 1 ("Most people spend too much time in unprofitable amusements"). But item 5 ("Most people who don't succeed in life are just plain lazy") has non-uniform DIF with respect to gender because all the p values were significant.

Figure 3.2

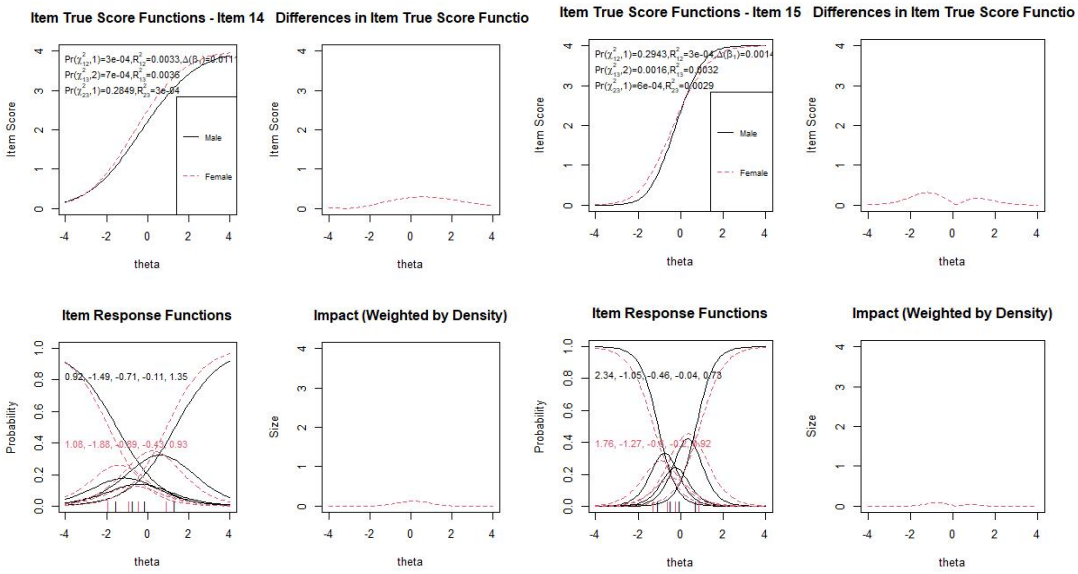
Item Characteristic curves for item 7 (left) and item 9 (right)



Note. Graphical representation of item 7 ("I frequently believe I would be more successful if I sacrificed some pleasures") and item 9 ("People who fail at a job typically do not try hard enough") exhibiting uniform DIF with regards to gender.

Figure 3.3

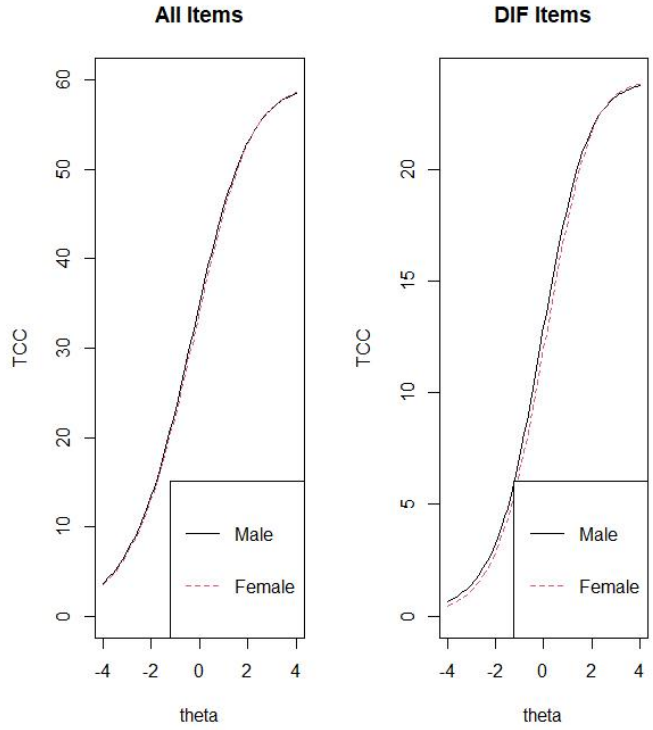
Item Characteristic curves for item 14 (left) and item 15 (right)



Note. Graphical representation of item 14 ("I feel uneasy when there is little work for me to do") exhibiting uniform DIF and item 15 ("A distaste for hard work usually reflects a weakness of character") exhibiting non-uniform DIF with regards to gender.

Figure 4

All vs DIF test characteristic curves (TCCs)

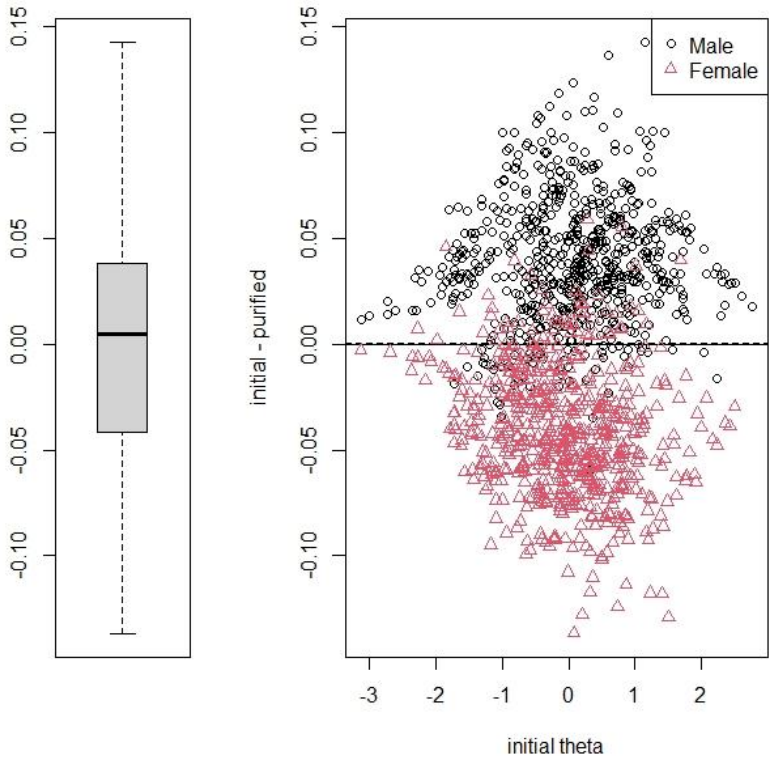


Note. These plots demonstrate how all DIF items affect the test characteristic curves (TCCs).

The left plot (see Figure 4) is based on item parameter estimates for all 15 items, as well as group-specific parameter estimates for the six DIF-identified items. The right plot is based only on the estimations of the group–specific parameters of the DIF items. Although the difference in the TCCs is modest, it suggests that female participants may score lower (show less PWE) if gender group–specific item parameter estimates were employed for scoring. Due to the cancellation of differences in opposing directions, differences in ICCs may become negligible whether averaged over all test items (left plot) or over the subset of items that exhibit DIF (right plot). Although, it is plausible that the influence on trait estimates persists (Choi et al., 2011).

Figure 5

Individual–level DIF impact



Note. To quantify the impact at the individual score level, lordif contrasts the naïve theta estimate to the "purified" theta estimates from the final output that accounts for DIF.

Figure 5 illustrates the disparity in score among those that disregard DIF and those that account for it. The box plot of these differences is given on the left. The interquartile range (the boundary between the bottom and top of the shaded box) spans from -0.04 to +0.04, with a median of roughly +0.01. The graph on the right shows the corresponding difference scores plotted against the initial scores that neglect DIF ("initial theta") separately for female and male participants.The guidelines are set to 0.0 (solid line), which indicates that there is no difference, and the mean of the differences (dotted line; Choi et al., 2011). Accounting for the six DIF items resulted in somewhat more extreme scores around the center, i.e., higher scores in those with moderate levels of PWE. However, this was consistent across male and female participants reaching opposite sides of the mean.

DISCUSSION

Protestant Work Ethic as a construct has been a nebulous construct over the years being associated with religious beliefs, attitude toward economic well-being and leisure, and dedication to hard work. The dimensionality of the scale has been questioned over time and despite showing multiple factors, the reliability of the scale is most prominent in the uni-dimensional form which was achieved in this report by removing 4 items from the original scale of 19 items as they were regarded as redundant forms of the items accounting for the subject's attitude towards hard work and leisure. The modified 15-item PWE scale was then analyzed for any measurement invariance with respect to two genders (male and female). This was deemed necessary as the scale was developed in a time when the social context favored men over women when it came to holding ambition and having occupational interests. According to Angoff (1993), an item exhibits DIF when differing statistical properties between groups after variations in trait levels of the groups are taken into consideration. DIF for a PWE item in this report implies that the item differs in reflecting the degree of PWE trait across the two genders; i.e., latent trait severity by itself does not account for the participants' responses, and the lack of DIF would imply that the item displays measurement invariance (Vaughn-Coaxum et al., 2016; de Ayala, 2008). In this analysis, six items were marked as having variance, and subsequently group-specific item discrimination and category location parameters were computed. A new DIF-adjusted trait (theta) estimate was created for each participant using these group-specific estimates, compensating for cross-gender measurement bias such that all latent trait estimates of trait level had equivalent meaning and could be rated on the same scale. DIF adjusted test information curves indicated a small but almost negligible difference in group effect for the female gender, however, there was a general overlap. When the PWE trait level was moderate, the individual level impact of DIF for male and female participants exhibited extreme scores across the mean in opposite directions. This disparity can be interpreted as having adverse DIF wherein groups differ in their probability of endorsing an item owing to some errors in the measurement process, such as different interpretations of a term or phrase used in the item (Columbia et al., 2019).

CONCLUSION

These findings suggest that PWE is a rather gender invariant scale for the male and female gender, however, the possibility of trait differences in certain items cannot be ruled out. Due to the lack of further evidence or literature, this report is limited in understanding the exact nature of the items that were flagged as having bias, especially when the level of the PWE trait is moderate in both male and female participants. The results are supposedly the consequence of adverse DIF, which could form the basis for further exploration.

REFERENCES

Angoff, W. H. (1993). Perspectives on differential item functioning methodology.

Arslan, M. (2000). A cross-cultural comparison of British and Turkish managers in terms of Protestant work ethic characteristics. Business Ethics: A European Review, 9, 13–19.

Bendix, R. (1975). Max Weber. Portret uczonego [Max Weber. An Intelectual Portrait]. Warszawa: Państwowe Wydawnictwo Naukowe

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), Testing structural equation models (pp. 136-162). Newbury Park, CA: Sage.

Chen, Y., & Zhang, S. (2020). Estimation Methods for Item Factor Analysis: An Overview. arXiv preprint arXiv:2004.07579.

Cherrington, D. (1980). The work ethic. Working values and values that work. New York: Amacom, A division of American Management Associations.

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(8), 1.

Columbia, Mailman School of Public Health. Differential Item functioning. (2019). [Online Source]: https://www.publichealth.columbia.edu/research/population-health-methods/differential-item-functioning

de Ayala, R. J. (2008). The theory and practice of item response theory. New York, NY: Guilford

DeMars, C. (2010). Item response theory. Oxford University Press.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Fullerton, K. Calvinism and capitalism: An explanation of the Weber thesis. In Green, R. W. (1959). Protestantism and capitalism: The Weber thesis and its critics: Heath.

Furnham, A. (1990a). The Protestant work ethic: The psychology of work-related beliefs and behaviors. London: Routledge.

Furnham, A. (1990b). A content, correlational and factor analytic study of seven questionnaire measures of the Protestant work ethic. Human Relations, 43, 383–399.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B.(2003). A very brief measure of the Big-Five
personality domains. Journal of Research in Personality, 37, 504-528.

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. Psychometrika, 54(4), 635-659.

Grabowski, Damian & Chudzicka-Czupała, Agata. (2016). Evaluation Of The Psychometric Properties Of The Polish Version Of The Protestant Work Ethic Scale By Mirels & Garrett. Studia Psychologiczne. 54. 1-17.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and ltenls. Applied psychological measurement, 9(2), 139-164.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. Educational and Psychological measurement, 66(3), 393-416.

Holland, P. W., & Wainer, H. (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.
This is the classic textbook on differential item functioning. It highlights methods for testing test items that function differently for different groups.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55.

Jones Jr, H. B. (1997). The protestant ethic: Weber s model and the empirical literature. Human Relations, 50(7), 757-778.

KozyrKowalski, S. (1967). Max Weber a Karol Marks. Socjologia Maxa Webera jako „pozytywna krytyka materializmu historycznego" [Max Weber and Carl Marx. Max Weber's Sociology as 'Positive Critique of Historical Materialism']. Warszawa: Książka i Wiedza

Mair, P. (2018). Modern psychometrics with R. Cham: Springer International Publishing.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

McHoskey, J. W. (1994). Factor structure of the Protestant work ethic scale. Personality and Individual Differences, 17, 49–52

Mirels, H. L., & Garrett, J. B. (1971). The Protestant ethic as a personality variable. Journal of consulting and clinical psychology, 36(1), 40.

Miller, M. J., Woehr, D. J., Hudspeth, N. (2002). Themeaning and measurement of work ethic: Construction and initial validation of a multidimensional inventory. Journal of Vocational Behavior, 60,
451–489

Mudrack, P.E. (1997). Protestant Ethic Dimensions and Work Orientations. Personality and individual differences, 23, 217–225.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage Publications.

Plouffe, R. A., Kowalski, C. M., Tremblay, P. F., Saklofske, D. H., Rogoza, R., Di Pierro, R., & Chahine, S. (2021). Gender differences or gender bias? Examination of the assessment of sadistic personality using item response theory and differential item functioning. European Journal of Psychological Assessment.

Revelle W (2022). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.2.5, https://CRAN.R-project.org/package=psych.

Suhr, D. D. (2005). Principal component analysis vs. exploratory factor analysis. SUGI 30 proceedings, 203, 230.

Tang, T. L.-P. (1993). A factor analytic study of the Protestant work ethic. Journal of Social Psychology, 133, 109–111.

Vaughn-Coaxum, R. A., Mair, P., & Weisz, J. R. (2016). Racial/ethnic differences in youth depression indicators: An item response theory analysis of symptoms reported by White, Black, Asian, and Latino youths. Clinical Psychological Science, 4(2), 239-253.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41(3), 321-327.

Weber, M., Parsons, T., & Tawney, R. H. (1958). The Protestant ethic and the spirit of capitalism. New York: Scribner.

Wentworth, D. K., Chell, R. M. (1997). American college students and the Protestant work ethic. The Journal of Social Psychology, 137, 284–296

Zhang, S., Liu, W., & Liu, X. (2012). Investigating the relationship between Protestant work ethic and Confucian dynamism: An empirical test in mainland China. Journal of business ethics, 106(2), 243-252.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, *160*.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological bulletin, 99(3), 432.