December 2021

The Almost Perfect Scale in Medical Students: Model Confirmation, Measurement Invariance, and Differential Item Functioning By Gender

Elizabeth Hollenback Ellinas
University of Wisconsin-Milwaukee

THE ALMOST PERFECT SCALE IN MEDICAL STUDENTS: MODEL CONFIRMATION, MEASUREMENT INVARIANCE AND DIFFERENTIAL ITEM FUNCTIONING BY GENDER

by

Elizabeth H. Ellinas

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Educational Psychology

at

University of Wisconsin-Milwaukee

December 2021

ABSTRACT

THE ALMOST PERFECT SCALE IN MEDICAL STUDENTS:
MODEL CONFIRMATION, MEASUREMENT INVARIANCE AND DIFFERENTIAL ITEM FUNCTIONING BY GENDER

by

Elizabeth H. Ellinas

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor Bo Zhang

This study examined the factor structure of two common perfectionism scales – the Almost Perfect Scale – Revised (APS-R) and the Short Almost Perfect Scale (SAPS) - in medical students. It was found that both two-factor models hold for them, albeit marginally for the APS-R. Measurement invariance by gender showed that while configural invariance and metric invariance hold, scalar invariance does not, indicating that the means for men and women may not be meaningfully compared by using these scales. Additionally, several items exhibited differential item functioning, most of which are in the Discrepancy scale of the APS-R. Overall, the SAPS provides better fit with fewer biased items, and therefore is likely to be a better instrument for comparing perfectionism in men and women medical students, although direct comparison of group means should still be exercised with caution.

To

my dad,

John J. Hollenback Jr., MS, MBA,

a statistician

who never had negative horses.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to begin by thanking Dr. Bo Zhang, my major professor, and Dr. Razia Azen, my advisor, without whose patience and genial support I would never have finished my master's degree. And Dr. Nadya Fouad, whose appearance and encouragement started this whole thing.

I would further like to thank and acknowledge Dr. Tavinder Ark, who helped me learn R, inspired my interest in scale development and factor analysis, and provided access to the data set for this thesis. At the time she arrived in my life, I truly had come to a place where I had to admit that I was not going to finish the master's thesis. She inspired me to continue and was instrumental in my success. It's a debt I can't repay, but only pay forward in some way, and I intend to do that. Thank you Tav.

And to my family:

To my husband, Herodotos, who has my back. Every day. All the time. I send to him my forever love and appreciation.

To my children, I send my love, thanks, and gratitude for their support, patience, cheerleading, and discussion of sunk costs, complete with Oxford commas.

Finally, to my daughter, my mother, and all the women students and learners:

I see you.

I'm going to make sure the data sees you too.

INTRODUCTION

THE ALMOST PERFECT SCALE IN MEDICAL STUDENTS: MODEL CONFIRMATION, MEASUREMENT INVARIANCE, AND DIFFERENTIAL ITEM FUNCTIONING BY GENDER

Medical student wellness has been a topic of great concern for medical schools, with over 70,000 papers on the topic available through the search engine "PubMed" as of July 2021 and half of those in the last five years. Beyond overall wellness, concerns specifically regarding mental health of medical students have led to discovery of high rates of anxiety (Quek et al., 2019) and depression (Blacker et al., 2019) in students, along with stigma against asking for mental health assistance (Blacker et al., 2019; Hankir et al., 2014) in a population that is expected to be strong, selfless, invincible, or in a word, "perfect."

As we seek both causes and solutions for mental health concerns, many have turned to underlying traits of students, employing scales to identify those at risk. Perfectionism was proposed as a factor with both positive and negative elements as early as the 1970's (Hamachek, 1978), with most of the focus on its association with psychopathology. In order to quantify amounts of perfectionism, several scales have been developed, including the Almost Perfect Scale (APS) (Slaney et al., 1995). The most common version of this scale, the APS-Revised or APS-R (Slaney et al., 2001), has items for orderliness (usually ignored (Stoeber & Otto, 2006)) and items for both positive perfectionism (having high personal "Standards" for oneself) and maladaptive perfectionism (feeling a "Discrepancy" between personal standards and performance). Sums of scores on Discrepancy and Standards scales sort people into three groups: non-perfectionists, perfectionists (associated imperfectly with positive psychological

outcomes (Stoeber & Otto, 2006), and maladaptive perfectionists (associated with negative psychological outcomes (Bußenius & Harendza, 2019)).

The Almost Perfect Scale (APS) has become a widely used psychological scale that evaluates perfectionism in many types of people, especially students. Two different versions of this scale are currently in use: the Almost Perfect Scale – Revised (APS-R), and the Short Almost Perfect Scale (SAPS). We validate the use of the APS-R and SAPS in medical students by determining whether measurement invariance (MI) holds for these scales by gender in this population. If measurement invariance is violated, differential item functioning (DIF) is used to determine which questions may have contributed to the difference.

LITERATURE REVIEW

Development of the Almost Perfect Scale

The first publications of the APS appeared in the early 1990s (Slaney et al., 1995). The authors discussed the potential to measure perfectionism through a number of related concepts including orderliness, positive strivings, relationship difficulties and procrastination. The authors believed that of these related concepts, "high standards" was the best measure of perfectionism, with the greatest relationship to adaptive or maladaptive behaviors.

In 2001, Slaney and colleagues substantially revised the original scale into the Almost Perfect Scale Revised or APS-R (Slaney et al., 2001), retaining questions for orderliness and high standards, but dropping other elements of the original scale (e.g. relationship difficulties). They made specific attempts to distinguish positive and negative aspects of perfectionism, mirror dictionary definitions of perfectionism, and have clear and logical implications for psychologists. The authors retained 12 items from their original scale (6 for order and 6 for perfectionism) and added 7 new items for "Standards" (high performance standards) and 20 items for "Discrepancy" (self-critical performance evaluations or the discrepancy between expectations and perceived performance) which represented a new factor not previously captured. Using exploratory factor analysis, they constructed a 23-item scale that included 4 items for "Orderliness," 7 for high "Standards" or positive perfectionism, and 12 for "Discrepancy" or maladaptive perfectionism.

In 2006, Stoeber and Otto (Stoeber & Otto, 2006) reviewed the literature on perfectionism, with an emphasis on whether positive perfectionism was actually a benefit for people's psychological health. The authors suggested that the four items for orderliness could be dropped. Most authors since have not included the results of the orderliness scale. With Stoeber and Otto's paper, the APS-R essentially became a 19-item scale. Rice and Ashby then used

cluster analysis to standardize classification of perfectionists in the APS-R (K. G. Rice & Ashby, 2007) and develop the cutoff criteria for classifications that we currently use, with sums of scale items sorting respondents into one of three possible categories: Non-perfectionists (low on both Standards and Discrepancy), Perfectionists (high Standards and low Discrepancy) and Maladaptive Perfectionists (high Standards and high Discrepancy)

No further changes were suggested to the APS-R until 2014 when Rice and colleagues (K. G. Rice et al., 2014) proposed an 8-item "short form" for the APS-R called the Short Almost Perfect Scale (SAPS). SAPS utilizes questions directly taken from the APS-R and consists of four questions each on adaptive and maladaptive perfectionism. Rice's stated goals for the SAPS were to reduce redundancy and ambiguity to create a shorter but equally psychometrically valid perfectionism scale.

Although the APS-R continues in use, since its introduction in 2014, SAPS has gained in acceptance, has been translated into several languages (Lins de Holanda Coelho et al., 2021), and at least one attempt has been made to classify perfectionists using this scale (Wang et al., 2016). See Table 1 for items in the APS-R and SAPS.

Table 1: Items in the Almost Perfect Scale – Revised (APS-R) and the Short Almost Perfect Scale (SAPS).

Standards	
stand_1	I have high standards for my performance at work or at school.
Stand_5	If you don't expect much out of yourself, you will never succeed.
Stand_8*	I have high expectations for myself.
Stand_12*	I set very high standards for myself.
Stand_14*	I expect the best from myself.
Stand_18	I try to do my best at everything I do.
Stand_22*	I have a strong need to strive for excellence.
Discrepancy	
disc_3	I often feel frustrated because I can't meet my goals.
Disc_6	My best just never seems to be good enough for me.
Disc_9	I rarely live up to my high standards.
Disc_11*	Doing my best never seems to be enough.
Disc_13	I am never satisfied with my accomplishments.
Disc_15	I often worry about not measuring up to my own expectations.
Disc_16*	My performance rarely measures up to my standards.
Disc_17	I am not satisfied even when I know I have done my best.
Disc_19	I am seldom able to meet my own high standards of performance.
Disc_20*	I am hardly ever satisfied with my performance.
Disc_21	I hardly ever feel that what I've done is good enough.
Disc_23*	I often feel disappointment after completing a task because I know I could have done better.

Note: stand = standards, disc = discrepancy
*Items also in the SAPS

Scoring of the APS-R and SAPS

APS-R scoring classifies respondents into three groups: non-perfectionists, adaptive perfectionists, and maladaptive perfectionists (K. G. Rice & Ashby, 2007). These groups are created based on the total score from 19 items (7 for Standards, 12 for Discrepancy). Each item uses a 7-point Likert scale: Strongly Disagree (1) to Strongly Agree (7). Scoring follows the following steps: Sum the items for Standards (maximum possible 49). If your Standards score sum <42 you are a Non-perfectionist. If your Standards score >=42, check your Discrepancy score (maximum possible 84). If your Discrepancy score <42 you are an Adaptive Perfectionist.

If your Discrepancy score is >=42 you are a Maladaptive Perfectionist. If all Standards items are answered, this corresponds to a mean Standards score of >=6, and if all Discrepancy items are answered, this corresponds to a mean Discrepancy score of >= 3.5. While most authors utilize this classification, a few have argued that more categories than the three currently scored are needed. In particular, a fourth category has been proposed – low on Standards but high on Discrepancy – which may result from not meeting high expectations set by others (Wang et al., 2007). Latent profile analysis has indicated that the same three classes of perfectionists typically found using the APS-R are also found using the SAPS (Wang et al., 2016), but a scoring system for this scale has yet to be developed.

Perfectionism in medical students

Per Stoeber and Otto, perfectionism is "a personality style characterized by striving for flawlessness and setting of excessively high standards for performance accompanied by tendencies for overly critical evaluations of one's behavior" (Stoeber & Otto, 2006, p. 295). These traits are particularly likely to be present in medical students, who must strive for flawlessness in order to be accepted into medical school, and then are subject to both high performance standards and abundant opportunities for self-critical comparison to outstanding peers. Studies of perfectionism in medical students have chosen multiple different scales to measure perfectionism (Thomas & Bigatti, 2020), and therefore this section reports on perfectionism more generally. Hu and colleagues (Hu et al., 2019) found that maladaptive perfectionism in medical students (as measured by the APS and found in 25% of students) was significantly associated with shame and embarrassment, which were in turn associated with depression and anxiety. Their finding is consistent with the meta-analytic finding that the correlation between depression and perfectionistic concerns overall was 40% (Limburg et al.,

2017)), and with the suggestion that a focus on perfection contributes to poor coping with medical errors (Robertson & Long, 2018). In the Hu study, more than 60% of medical-student participants compared their academic performance to others at least moderately, and two-thirds reported tying their academic performance to their self-worth. In work comparing Canadian medical students to arts-college students using the Multidimensional Perfectionism Scale (Enns et al., 2001; Frost et al., 1990), the perfectionism profile for medical students was found to include higher personal standards scores and lower rates of maladaptive perfectionism relative to their arts-college peers. The authors conclude that medical students may systematically differ from general arts students and hypothesize that the Canadian medical school application process may select for positive perfectionism, a situation directly tested in a German study that suggested that maladaptive perfectionism was more prevalent in medical-school applicants that were eventually rejected for acceptance (Bußenius & Harendza, 2019). Leung combined multiple scales in a latent profile analysis of medical students (Leung et al., 2019) to describe a profile that might be at risk of poor coping in the medical setting. They included gender in their analysis and found that women scored higher on maladaptive measures and men scored higher on high standards. My literature search did not produce any studies regarding confirmatory factor analysis (CFA) or Measurement Invariance (MI) regarding the APS and medical students, medical residents, fully licensed medical doctors, or the effect of their perfectionism on their patients.

Gender differences and measurement invariance in the APS

Most studies on the APS have been completed in convenience samples of university students in psychology classes, which resulted in female-predominant samples. Although scores on the APS have not frequently been differentiated by gender, Rice and Ashby (K. G. Rice &

Ashby, 2007) reported that women had higher means on both Standards and Discrepancy, while another study of mixed ages found no differences in scores by gender (Ashby et al., 2008).

Rice and collaborators have considered measurement invariance for SAPS in Italian and US, and Korean and US, undergraduates (K. G. Rice et al., 2019; S. P. M. Rice et al., 2020), and in clients and non-clients in a university counseling center (K. G. Rice & Taber, 2019). In those studies, MI supported the two-factor SAPS structure but measurement non-invariance was found at the scalar (intercept) level, indicating that measured differences may result from the ways groups perceive perfectionism or reacted to the scale questions, rather than actual differences in the mean levels of perfectionism.

Research questions

Two research questions are addressed in this study:

1. Does the two-factor model for perfectionism as determined by the Almost Perfect Scale – Revised (APS-R) and/or the Short Almost Perfect Scale (SAPS) hold in medical students?
2. Does measurement invariance indicate that the perception of perfectionism as determined by the APS-R and/or SAPS holds in women and men medical students?

METHODS

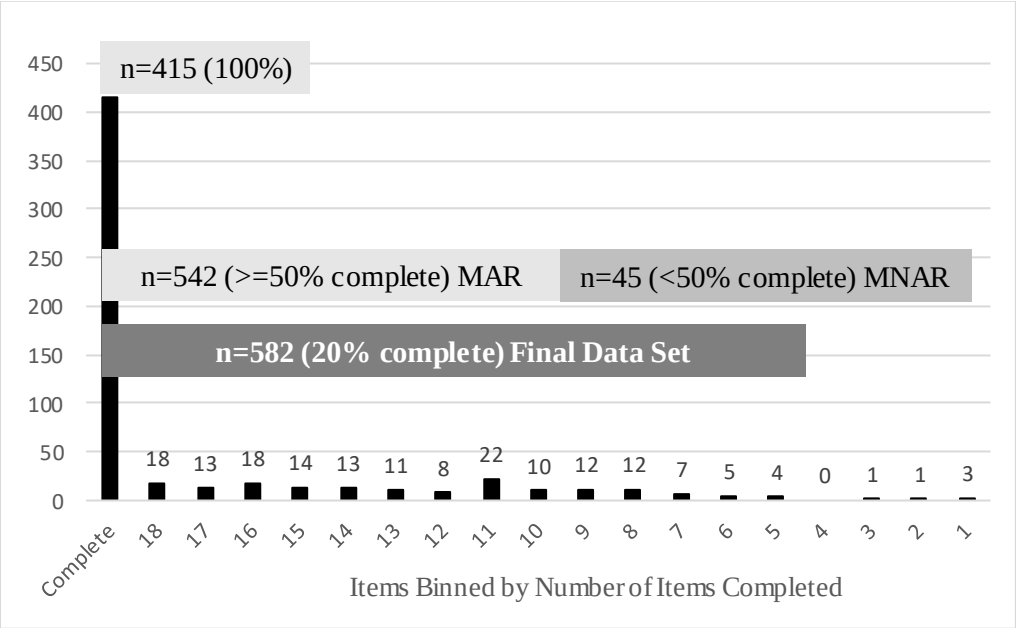
Sample

Of the 667 responses to the APS that included gender, there were 370 that self-identified as women and 295 that identified as men. Two respondents who listed their gender as non-binary were removed from subsequent analysis. Of the 665 binary-gender responses, 415 were entirely complete (all 19 Standards and Discrepancy items answered), 127 were less than 50% missing, 45 were greater than 50% missing but had at least 1 item answered, and 78 were entirely blank.

Some missing data were found to be not missing at random (NMAR). Dividing the responses into three groups: Complete, <50% missing, and >50% missing, analysis of variance and Tukey's post hoc test indicated that the >50% missing group was significantly lower on Standards and higher on Discrepancy than the other two groups ($p < 0.001$ for both omnibus and comparison of means), indicating that those 45 respondents may be NMAR.

Realizing that there is no perfect solution to data NMAR data, we considered whether we should use only complete data, 50% complete data, (which were found to be equivalent to the complete group) or include responses from low-completion respondents to attempt to account for the NMAR data. We elected to include the data down to 20% complete, which added an additional 40 responses to the 542 that were 50% complete or more. This is indicated graphically in Figure 1: a histogram indicating the number of responses by level of completeness.

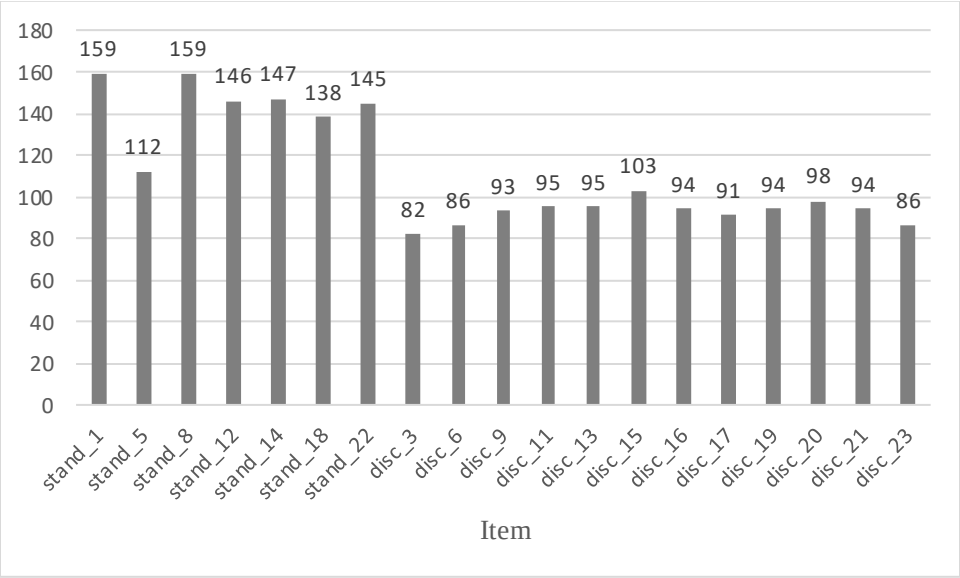
Figure 1: Histogram of binned responses by number of items completed.



We further considered responses by item. This is visualized in Figure 2, which shows the number of responses by item (the 415 complete cases that were present for every item are not shown to enhance readability). To illustrate, Standards 1 has 415 complete case plus 159 additional responses from the incomplete cases, providing a total of 574 responses out of a potential 582 responses, or 98% complete. Discrepancy 3, with the lowest number of responses, has 415 + 82 = 497 respondents, or 85% complete. Overall, we see that the Standards items had a better response rate than the Discrepancy items.

Full information maximum likelihood was utilized for the missing item responses (Peugh & Enders, 2004), realizing that this is an imperfect choice given our suspected MNAR items.

Figure 2: Responses per item for incomplete cases.*



*Each item has 415 additional responses from the 415 entirely complete cases. They are not included here for readability.

Analyses

Analyses were conducted in R (The R Core Team, 2021), using LAVAAN to conduct CFA (Rosseel, 2012) and lordif (Choi et al., 2011) to conduct DIF.

To address the first research question, which is **"**Does the model for perfectionism as determined by the Almost Perfect Scale – Revised (APS-R) and/or the Short Almost Perfect Scale (SAPS) hold in medical students?", we need to know whether the structure of perfectionism in the minds of medical students corresponds with previous findings for the APS scales. Specifically, we perform a CFA (Brown, 2015) on the whole population using the 19 items in the APS scale and the 8 items in the SAPS.

CFA is a multivariate model in which the latent factor or factors is unobserved or "latent," and follows the equation (Dimitrov, 2010; UCLA: Statistical Consulting Group, n.d.):

$$y_i = \tau_i + \lambda_i \eta + \varepsilon_i$$

Where y is the item or question in a scale, τ is the intercept (mean), λ is the loading of that item on the latent factor (or correlation of the item with the factor), η is the latent factor, and ε is the error variance (or the amount of variance in the item that is not attributable to the latent factor).

CFA further relates the observed covariance matrix (Σ) of the measured scale items to the latent factor variance using the following model:

$$\Sigma(\theta) = \Lambda \Psi \Lambda' + \Theta_\varepsilon$$

Where Λ is the matrix of factor loadings (uses λ's from the equation above), Ψ is the variance-covariance matrix for the latent factors (η's), and Θ_ε is the variance-covariance matrix of the residuals (ε's).

For example, a three-item scale can be expressed as follows using the equations from above for a single latent factor:

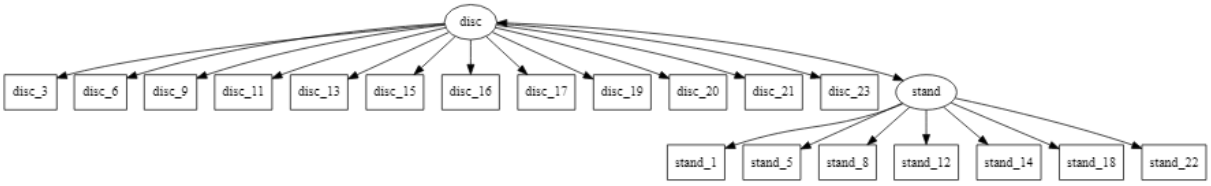
$$\Sigma(\theta) = \begin{pmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{pmatrix}(\psi_{11})(\lambda_1 \quad \lambda_2 \quad \lambda_3) + \begin{pmatrix}\theta_{11} & \theta_{21} & \theta_{31}\\ \theta_{12} & \theta_{22} & \theta_{32}\\ \theta_{13} & \theta_{23} & \theta_{33}\end{pmatrix}$$

For our case of the 19-item APS scale and two latent factors (Standards and Discrepancy) as predictors, Σ is a 19x19 observed population covariance matrix, Λ is a 2x19 factor loading matrix (with factor loadings set to zero if the item does not load on that factor), Ψ is a 2x2 variance-covariance matrix, and Θ_ε is a 19x19 matrix with the variances on the diagonal and the covariances held to zero because we have no reason (e.g. a single rater for some items) to believe the items will covary. The SAPS follows similarly, but with a reduced number of items.

Factor analysis generally requires a sample size of about 10 responses per total items, making adequate sample size about 190 participants for the APS-R, easily satisfied with the current sample. To address problems with non-normality or skew (e.g., a Standards score skewed toward perfectionism), we use robust maximum likelihood. Additionally, the 7-point Likert scale limits responses to the range 1-7, with any data entry errors noted and eliminated. As the APS is completed on a 7-choice Likert scale, responses were treated as continuous. We check specifically for the frequency of empty cells on extreme ends, looking to see if the 7-point scale is actually utilized.

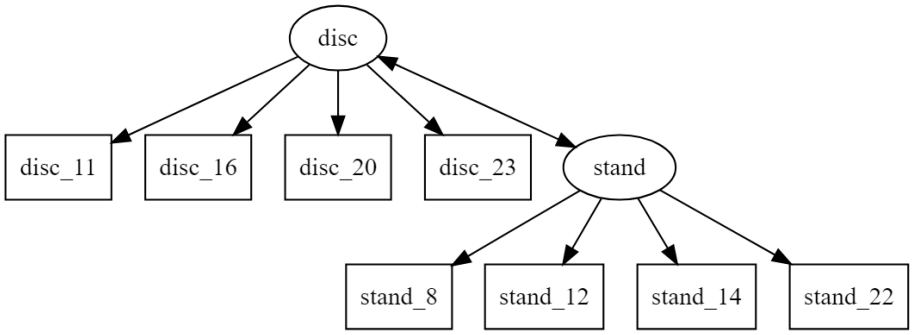
Factor analysis generally presumes that for a collection of observed variables or items, the interrelationships among those items can be explained by their relationship to latent, unobserved factors. In CFA, those relationships have been defined a priori, and we test those relationships. For the APS-R, we use the LAVAAN (Rosseel, 2012, 2021) package in R to test a two-factor model in which (Figure 1) the 7 items for Standards (items 1, 5, 8, 12, 14, 18, and 22) load to the latent factor Standards, and the 12 items for Discrepancy (items 3, 6, 9, 11, 13, 15, 16, 17, 19, 20, 21, and 23) load to the latent factor Discrepancy.

Figure 3: The confirmatory factor model for the APS-R.



Similarly for the SAPS (Figure 2), the 4 items for Standards (items 8, 12, 14, and 22) load to the latent factor Standards, and the 4 items for Discrepancy (items 11, 16, 20, and 23) load to the latent factor Discrepancy.

Figure 4: The confirmatory factor model for the SAPS.



To test model fit for, we use multiple fit indices, including exact fit indices, such as the chi-square goodness of fit, and approximate fit indices, such as the root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), the comparative fit index (CFI), and the Tucker-Lewis Index (TLI). In a well-fitting model we would expect a non-significant chi-square test, CFI and TLI at or above 0.90, and an RMSEA and SRMR at or lower than 0.08 (Chen, 2007; Hu & Bentler, 1999; Vandenberg & Lance, 2000). A well-fitting model

confirms that the model for perfectionism, as described by the SAPS and/or the APS-R fit the data for medical students.

To address the second research question, which is, "Does measurement invariance indicate that the perception of perfectionism as determined by the APS-R and/or SAPS holds in women and men medical students?", we examine MI by using multigroup CFA (Brown, 2015) for women and men. In general, multigroup CFA compares models in which increasingly restrictive parameters (e.g., factor loadings and group intercepts) are constrained to be equal, and compared to models in which those parameters are free to vary. If the constrained and unconstrained models fit equally well, measurement invariance is present. If the models are significantly different, then measurement non-invariance is present. Because the model with equality constraints is nested within the model that is free to vary, chi-square testing (along with other fit indices for CFA above) can be used to compare the models. Specifically, if the chi-square is not significant, the models are deemed equivalent at that level, we conclude that there is measurement invariance by gender, and we can proceed to the next more restricted model. If MI is not found (at any step) the analysis stops, and we can conclude that the model is not comparable at that level.

MI analysis is done using LAVAAN (Rosseel, 2021; UCLA: Statistical Consulting Group, n.d.) in R. We use the step-by-step process described in Brown (Brown, 2015, pp. 241–285) to hold the groups to successively greater levels of testing invariance. There are four levels of MI, with each level holding the model to successively greater constraints. We begin with configural invariance by asking whether the factor structure that we found in the CFA analysis of all medical students is the same regardless of gender, that is, do the factors load in the same pattern. Metric invariance examines whether the weight of the loading is equivalent between

groups. We do this by comparing the configural model, which provides a baseline estimate of the relationship between each item and the factors, to the metric model in which the loadings are constrained to be equal between genders. If we find metric invariance, we know that each perfectionism item has the same relationship to the perfectionism construct for both men and women, or put another way, we can say that a unit change in one group's perfectionism is equivalent to the other group's change in perfectionism. The next restriction is scalar invariance, or equal intercepts. We compare a model with equal factor loadings to a nested model with both equal loadings and equal intercepts. This necessitates the introduction of group means, accomplished by fixing a reference group's mean to zero. Of note, equal intercepts are what's needed to compare group means, i.e., to know that you can meaningfully compare the mean scores of women to the mean scores for men. If scaler invariance is not found, then differential item functioning is suspected and considered as the source of this level of measurement invariance. Finally, the last level of MI (strict invariance) asks whether the error residuals are equal (constrain the residuals and compare the model to the scalar model). If found, and it is rarely found, it indicates that the leftover variance or unique variance due to each item and measurement error is similar in both groups. Most authors do not require strict invariance in MI testing (Brown, 2015).

If measurement invariance is violated, differential item functioning (DIF) is used to identify which items may have contributed to the difference. More specifically, DIF determines whether women and men with the same underlying trait level of perfectionism endorse a different level of perfectionism on the APS scales (Teresi & Fleishman, 2007). DIF occurs when persons with equal amounts of trait, but from different groups, have different expected item responses. A classic example of this is items assessing "crying" in depression scales. Because

men are typically socialized to cry less, this item exhibits DIF – women with the same level of depression score higher than men on this item. DIF begins with an estimate of the "conditioning variable," typically one summed score of all items in a test or scale measuring a single construct. In the APS scores, the correlation between latent variables is nearly zero, and we therefore consider Standards and Discrepancy items as separate constructs. Because a sum is needed, and we are retaining responses with some missing items, missing data must be addressed. Missing scores on any item are replaced with the mean for that item across all respondents. While this reduces the variance of that item, it allows us to retain the same data set as used for CFA and MI. To test for DIF, we use an ordinal logistic regression model which compares a full model which includes a conditioning variable (APS), a grouping variable (gender), and an interaction term between the conditioning and grouping variables, with reduced models that include only the conditional or conditional and grouping variables. The models can be expressed as follows:

Model 1: conditional model: $\text{logit } P(u_i \geq k) = \alpha_k + \beta_1\theta$

Model 2: grouping variable model: $\text{logit } P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2\text{group}$

Model 3: full model: $\text{logit } P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2\text{group} + \beta_3(\theta \text{ x group})$

where $P(u_i \geq k)$ is the cumulative probability that the item response falls into category k or higher, α_k is the intercept, θ is the true trait level, and β is the respective regression coefficient. If no significant differences are found between nested models, there is no DIF for that item. If a comparison between Model 1 and 3 is significant, DIF is present, and comparison can be sought between other models to determine the type of DIF. If the grouping variable is significant in comparing Model 2 to Model 1, this indicates uniform DIF: DIF of the same direction and magnitude between genders across the spectrum of perfectionism. A significant interaction term from comparing Model 3 and Model 2 indicates non-uniform DIF: different amounts or direction

of DIF at higher or lower amounts of theta (perfectionism)) (Choi et al., 2011; Crane et al., 2007).

RESULTS

Descriptive statistics

Descriptive statistics for the final data set are presented in Table 2, along with differences between means on items and factors by gender.

While the entire scale (from 1-7) was generally utilized (3 of the Standards items did not utilize "1" on the scale), multiple items were found to fail multivariate normality and several had significant skew (Table 2), and therefore robust measures (robust maximum likelihood were subsequently used in the analysis.

Note in Table 2 that while women and men have nearly identical average Standards scores, their scores differ on Discrepancy. A t-test of the Discrepancy means by gender indicates that the means for women are "significantly higher" than the means for men ($p < 0.05$ in both cases). A relevant question from this study is whether these means can be meaningfully compared, as the difference may not be due to the trait difference but rather to how the two groups interpret the APS items.

Table 2: Descriptive variables in the APS scales data set.

	All (n=582)							Women (n=325)					Men (n=257)					Difference
	n	mean	sd	min	max	skew	kurt	n	mean	sd	min	max	n	mean	sd	min	max	in means
Latent Factor Means																		
APS-R Stand	581	6.30	0.60	3.4	7	-1.33	2.56	324	6.34	0.57	3.4	7	257	6.25	0.64	3.7	7	0.09
APS-R Disc	574	3.63	1.59	1	7	0.46	-0.95	322	3.81	1.66	1	7	252	3.40	1.47	1	7	0.41
SAPS Stand	579	6.38	0.67	2.8	7	-1.54	3.60	324	6.40	0.67	2.8	7	255	6.36	0.67	3.5	7	0.04
SAPS Disc	558	3.32	1.69	1	7	0.68	-0.74	308	3.47	1.76	1	7	250	3.14	1.58	1	7	0.32
Individual Item Means																		
stand_1	572	6.61	0.64	1	7	-2.58	12.96	320	6.63	0.64	1	7	252	6.58	0.64	3	7	0.06
stand_5	526	5.79	1.26	1	7	-1.59	2.67	291	5.79	1.29	1	7	235	5.80	1.23	1	7	-0.01
stand_8*	574	6.54	0.69	2	7	-1.99	6.27	321	6.59	0.66	3	7	253	6.49	0.72	2	7	0.10
stand_12*	561	6.36	0.90	1	7	-2.05	6.44	316	6.38	0.92	1	7	245	6.33	0.86	1	7	0.06
stand_14*	562	6.33	0.84	2	7	-1.75	4.57	315	6.33	0.88	2	7	247	6.34	0.79	3	7	-0.01
stand_18	553	6.20	0.98	1	7	-1.75	4.15	308	6.35	0.81	2	7	245	6.00	1.13	1	7	0.34
stand_22*	560	6.31	0.85	2	7	-1.80	5.27	320	6.30	0.92	2	7	240	6.32	0.76	4	7	-0.02
disc_3	497	4.65	1.72	1	7	-0.42	-1.01	282	4.86	1.67	1	7	215	4.39	1.74	1	7	0.47
disc_6	499	3.67	1.84	1	7	0.29	-1.19	276	3.94	1.90	1	7	223	3.34	1.71	1	7	0.60
disc_9	508	3.56	1.85	1	7	0.47	-1.11	279	3.70	1.93	1	7	229	3.39	1.74	1	7	0.30
disc_11*	510	3.17	1.87	1	7	0.65	-0.86	281	3.34	1.95	1	7	229	2.96	1.74	1	7	0.38
disc_13	509	3.14	1.86	1	7	0.68	-0.79	281	3.20	1.95	1	7	228	3.07	1.74	1	7	0.12
disc_15	517	4.71	1.89	1	7	-0.47	-1.11	289	4.84	1.90	1	7	228	4.55	1.86	1	7	0.29
disc_16*	509	3.29	1.77	1	7	0.73	-0.75	277	3.42	1.85	1	7	232	3.13	1.66	1	7	0.29
disc_17	506	3.39	1.90	1	7	0.42	-1.15	276	3.50	1.95	1	7	230	3.26	1.83	1	7	0.24
disc_19	509	3.37	1.82	1	7	0.60	-0.96	279	3.50	1.88	1	7	230	3.20	1.74	1	7	0.29
disc_20*	513	3.01	1.77	1	7	0.88	-0.36	277	3.04	1.85	1	7	236	2.97	1.69	1	7	0.07
disc_21	509	3.07	1.81	1	7	0.78	-0.59	275	3.17	1.89	1	7	234	2.96	1.69	1	7	0.22
disc_23*	500	3.63	1.89	1	7	0.37	-1.13	277	3.79	1.99	1	7	223	3.42	1.74	1	7	0.38

Note: *Items in the SAPS, n = number of responses, stand = standards, disc = discrepancy, sd = standard deviation, kurt = kurtosis. Latent factor means are the means for all items in the latent factor, e.g., the mean for all 7 Standards items in the APS-R.

Results for the APS-R

Confirmatory factor analysis for the APS-R

A two-factor model for the APS-R was specified using LAVAAN software in R, utilizing robust multivariate analysis and full information maximum likelihood (FIML) to account for missing data. The complete model specification is depicted in Figure 1. The model presumed that all indicators loaded onto only one of two latent variables, and all item measurement errors were uncorrelated. The latent variables of Standards and Discrepancy were allowed to correlate r=0.079 (p=0.058), indicating that the latent factors have good discrimination between Standards and Discrepancy.

Goodness of fit indices (Table 3) indicate that the APS-R had marginal fit in medical students: while CFI, TLI, and SRMR were within accepted limits, RMSEA was just barely below the 0.08 threshold. P values for Chi-square (<0.05), although out of bounds for good fit, were likely due to the large sample size (Brown, 2015, p. 69).

Table 3: Confirmatory factor analysis goodness-of-fit results for the APS-R for all respondents, and by gender.

Group	n	df	ChiSq	pChiSq	SRMR	RMSEA	CFI	TLI
All	582	151	652.18	0.000	0.05	0.08	0.90	0.91
Women	325	151	494.08	0.000	0.06	0.08	0.88	0.89
Men	257	151	338.54	0.000	0.05	0.07	0.93	0.93

Note: ChiSq = chi-square, df = degrees of freedom, CFI = comparative fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual

Modification indices indicated that allowing some of the factor loadings to correlate would improve the fit of the APS-R model. While allowing these to correlate did improve RMSEA for the APS-R to <0.08, there was no compelling substantive rationale that those factors should correlate while others did not, and these solutions were not further pursued.

Dividing the respondents by gender (Table 3) indicates that the model fit for those who self-identified as men was better than the fit for those who identified as women. For men, the model fit includes lower RMSEA's than the model for women, including a non-significant p value for the chi-square calculation, although this could be partly influenced by sample size. The APS-R model fits reasonably well for men and marginally for women.

Table 4: Factor loadings for the APS-R for all respondents, and by gender.

Loadings for All Respondents				Loadings by Gender	
Item	Standardized	p	R^2	Women	Men
stand_1	0.69	<0.001	0.47	0.62	0.76
stand_5	0.23	<0.001	0.05	0.19	0.29
stand_8	0.82	<0.001	0.68	0.77	0.88
stand_12	0.78	<0.001	0.61	0.77	0.78
stand_14	0.70	<0.001	0.49	0.68	0.76
stand_18	0.48	<0.001	0.23	0.51	0.45
stand_22	0.70	<0.001	0.48	0.68	0.75
disc_3	0.50	<0.001	0.25	0.50	0.51
disc_6	0.81	<0.001	0.65	0.82	0.78
disc_9	0.79	<0.001	0.63	0.81	0.75
disc_11	0.90	<0.001	0.80	0.90	0.89
disc_13	0.87	<0.001	0.76	0.89	0.84
disc_15	0.65	<0.001	0.42	0.67	0.61
disc_16	0.88	<0.001	0.78	0.89	0.88
disc_17	0.81	<0.001	0.65	0.84	0.76
disc_19	0.89	<0.001	0.79	0.89	0.88
disc_20	0.92	<0.001	0.85	0.92	0.93
disc_21	0.90	<0.001	0.82	0.89	0.93
disc_23	0.76	<0.001	0.57	0.76	0.75

Factor loadings for the APS-R are presented in Table 4. Standards 1 and Discrepancy 3 were indicator items. All freely estimated factor loadings (Z-values) were significant at the $p<0.001$ level, indicating that all the items load significantly onto their respective latent variables. That said, loadings for some items, particularly stand_5, "If you don't expect much out

of yourself, you will never succeed," loaded poorly onto Standards: R-squared for that item was 0.05, indicating that only 5% of the variance in responses to that item was explained by the underlying factor, Standards. Factor loadings by gender (Table 4) for the APS-R illustrate a pattern where men load more strongly than women onto the Standards items while women loaded more strongly than men on Discrepancy items.

Measurement invariance with multi-group CFA in the APS-R

Table 5 presents the results of the MI analysis, and includes both the parameters for the individual models (e.g. "Chi-square") and the change in parameters between successive nested models (e.g. "Chi-square difference") using the Santorra-Bentler method to compensate for non-normality.

Table 5: Measurement invariance results for the APS-R by gender

APS-R	Goodness of Fit						Difference in Fit	
	df	RMSEA	SRMR	CFI	TLI	ChiSq	ChiSq diff	p
Configural	302	0.08	0.06	0.92	0.91	838.57		
Metric	319	0.08	0.06	0.92	0.91	860.47	21.90	0.189
Scalar	336	0.08	0.06	0.91	0.91	912.96	52.49	**<0.001**

Note: df = degrees of freedom, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual, CFI = comparative fit index, TLI = Tucker-Lewis Index, ChiSq = chi-square, diff = difference; bolded p<0.05

Configural invariance compares the structural components of the models for men and women (do the items load onto factors in the same pattern). The configural model for the APS-R is found to be structurally equivalent by gender. This model is then compared to the nested model of metric invariance in which the factor loadings for men and women are constrained to be equal. The non-significant chi-square difference test (p=0.189) for the comparison between these configural and metric models indicates that the constraint of equal factor loadings does not

significantly degrade the fit of the solution, and therefore the factor loadings for the APS-R are equivalent by gender. This finding demonstrates that the meaning and structure of the models are equivalent in men and women medical student respondents. It also indicates that the regression slopes are parallel, and therefore a unit change in the underlying dimension (Standards or Discrepancy) results in a statistically equivalent change in the observed measure for both women and men.

This does not, however, indicate that the mean scores for the groups can be directly compared. To do that, we must add scalar invariance, which holds both the loadings and intercepts for men and women to equivalence. The comparison to a scalar invariance model for the APS-R results in a significant chi-square test ($p<0.001$), indicating a significant difference between the models. The rejection of the null hypothesis indicates that while the regression slopes are equal, the intercepts are not, and comparison of latent means for men and women (for example, the "significant t test" by gender on Discrepancy previously described) is not interpretable, because any mean differences may be due to differences in meaning of the construct. Additionally, testing for measurement invariance stops once measurement non-invariance is found, and therefore strict invariance (equal error variance) is not tested.

Differential item functioning in the APS-R

Having failed to achieve scalar invariance, we look to differential item functioning (DIF) to determine which items might have contributed to the variance between men and women on the APS-R. "Uniform DIF" represents a discrepancy between the conditional and grouping models (models 1&2), "non-uniform DIF" a discrepancy between the grouping and interaction models, and "total DIF effect" guards against type 1 error (in large samples), and should be significant if either uniform or non-uniform DIF is present. Table 6 shows the DIF results produced using the

lordif package in R (Choi et al., 2011) for ordinal logistic regression in polytomous items in the APS-R. Testing was completed for the Standards items separately from the Discrepancy items. Using an alpha level of 0.05, five items exhibited DIF, one Standards and four Discrepancy items. For four of five items flagged for DIF in the APS-R, uniform DIF is found, indicating an even distribution of DIF across theta for both genders. This is consistent with the MI finding of scalar invariance. Discrepancy item 13, however, has some elements of non-uniform DIF, consistent with metric invariance or problems with factor loadings, and may help to explain the marginal fit of the overall CFA model in APS-R. McFadden's R^2 is consistently less than 0.013, indicating that although DIF is present, its effect is small (Jae Jeong, 2016).

Table 6: Differential item functioning in the APS-R by gender

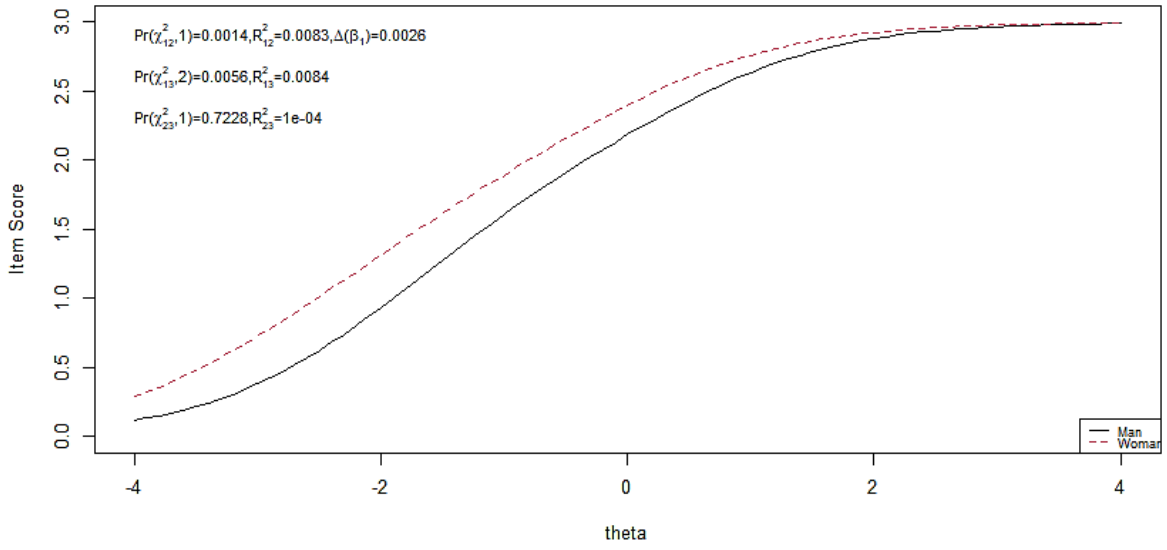
item	#cat	DIF present*	total DIF effect*	uniform DIF*	non-uniform DIF*	R^2
stand_1	2		0.935	0.715	0.983	0.000
stand_5	5		0.384	0.501	0.227	0.000
stand_8	3		0.071	0.121	0.089	0.003
stand_12	4		0.197	0.420	0.107	0.001
stand_14	4		0.109	0.151	0.124	0.002
stand_18	4	YES	**0.006**	**0.001**	0.723	0.008
stand_22	4		0.779	0.481	0.956	0.000
disc_3	7	YES	**0.003**	**0.003**	0.072	0.005
disc_6	7	YES	**0.002**	**0.000**	0.580	0.007
disc_9	7		0.568	0.707	0.320	0.000
disc_11	7		0.962	0.858	0.830	0.000
disc_13	7	YES	**0.010**	**0.041**	**0.026**	0.002
disc_15	7		0.487	0.311	0.519	0.001
disc_16	7		0.814	0.523	0.947	0.000
disc_17	7		0.409	0.784	0.191	0.000
disc_19	7		0.479	0.496	0.315	0.000
disc_20	7	YES	**0.036**	**0.011**	0.669	0.004
disc_21	7		0.708	0.456	0.715	0.000
disc_23	7		0.426	0.317	0.400	0.001

Note: #cat = number of categories , R2 = McFadden's effect size.
*p values for model comparisons, bolded p< 0.05

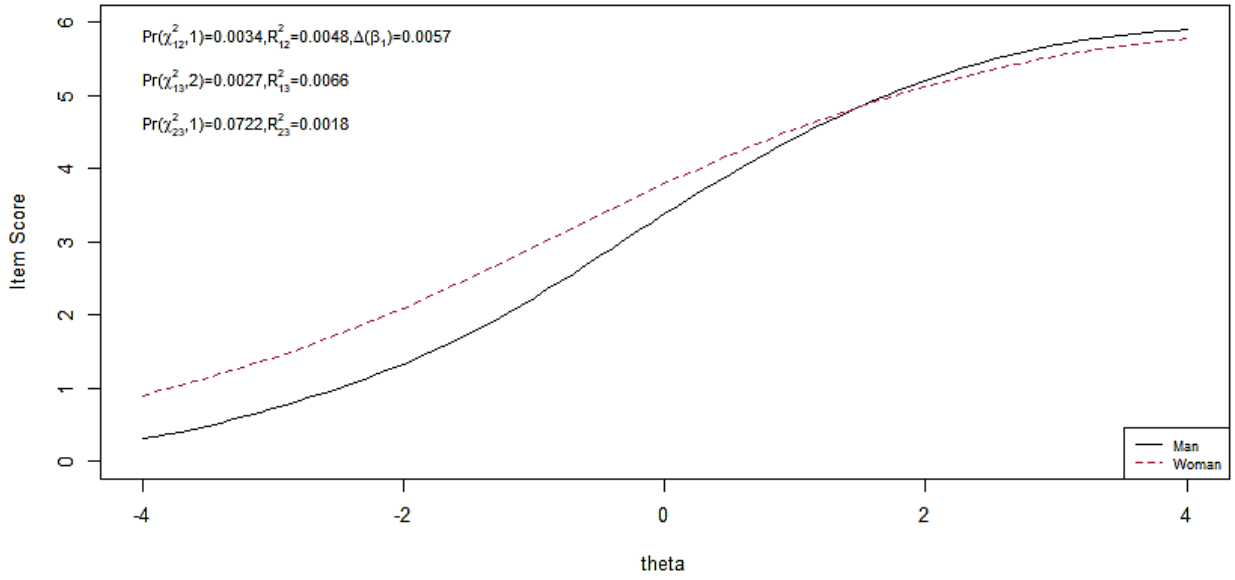
Figure 3 shows the item characteristic curves for the items exhibiting DIF in the APS-R. The x-axis indicates theta score (ability on the item), and the y-axis is the mean item score across individuals of every level of theta. Each line represents one gender, and a gap between the two lines indicates DIF – for women and men with the same level of ability, the item score is different. For example, for Standards 18, especially at low levels of theta, DIF causes women to score higher than men. The crossed curves for item Discrepancy 13 indicate that the DIF is not consistent throughout the scale, i.e., non-uniform DIF.

Figure 5: Item characteristic curves for DIF items in the APS-R

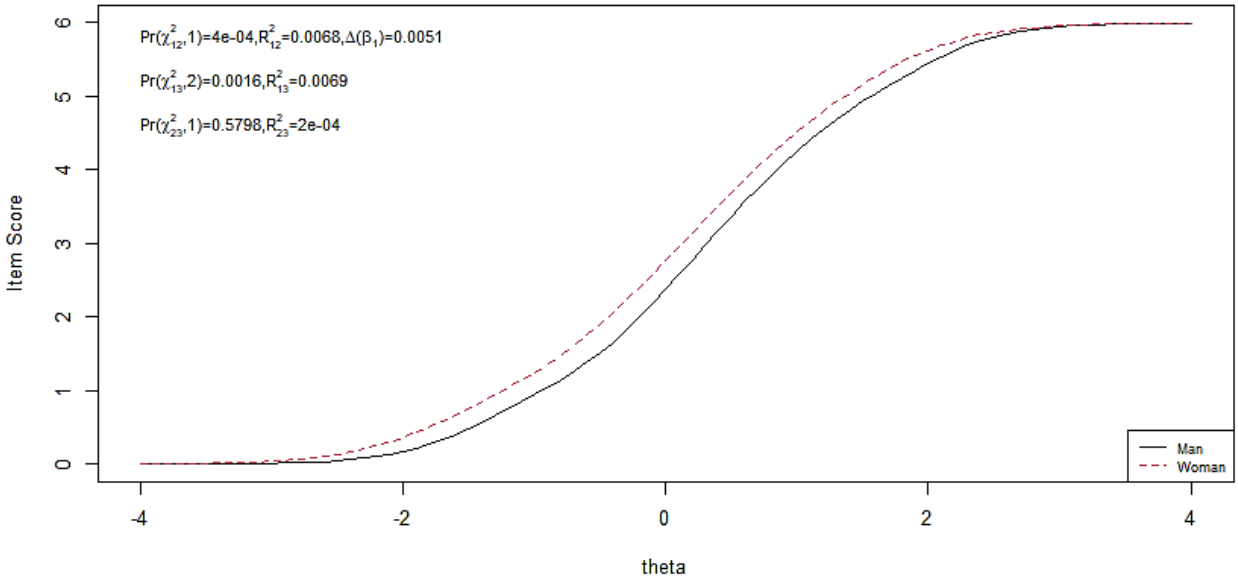
5a: Standards 18: "I try to do my best at everything I do."



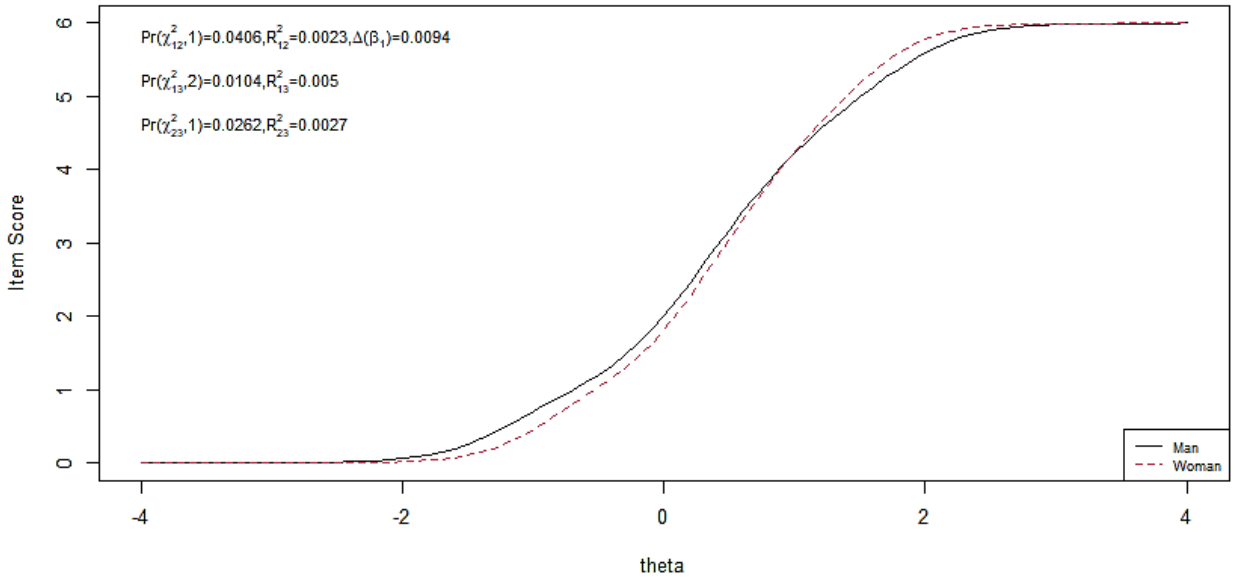
5b: Discrepancy 3: “I often feel frustrated because I can’t meet my goals.”



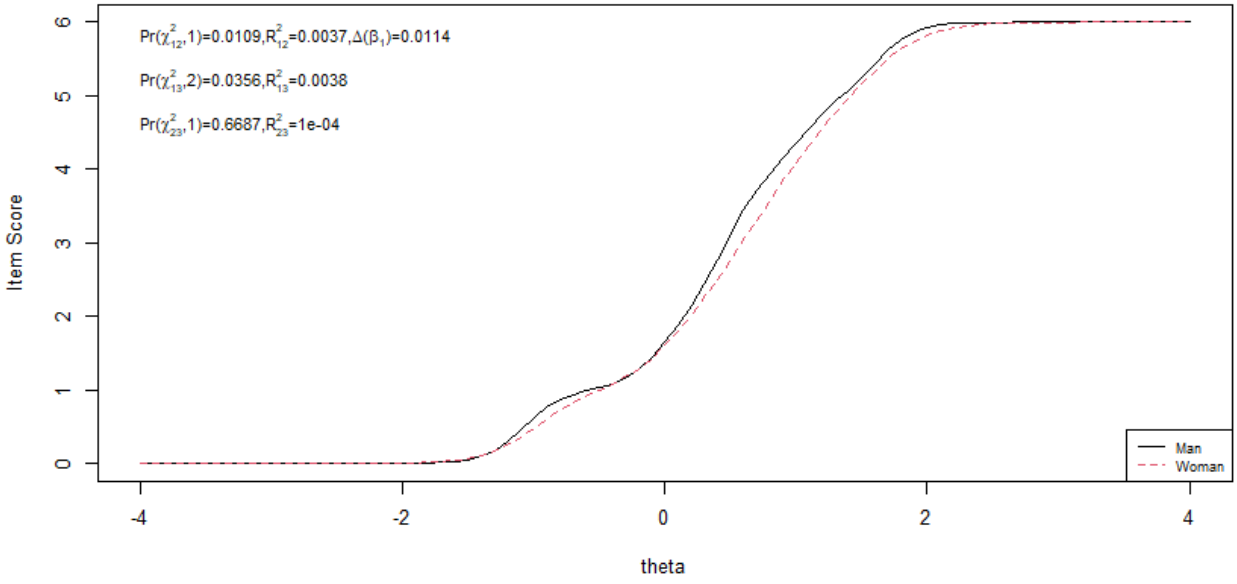
5c: Discrepancy 6: “My best just never seems to be good enough for me.”



5d: Discrepancy 13: “I am never satisfied with my accomplishments.”



5e: Discrepancy 20: “I am hardly ever satisfied with my performance.”



Results for the SAPS

Confirmatory factor analysis for the SAPS

A two-factor model for SAPS were specified using LAVAAN software in R, utilizing robust multivariate analysis and full information maximum likelihood (FIML) to account for missing data. The complete model specification is depicted in Figure 2. The model presumed that all indicators loaded onto only one of two latent variables (Standards or Discrepancy), and all item measurement errors were uncorrelated. The latent variables were allowed to correlate, resulting in a small but significant correlation r=0.088 (p=0.040) and indicating that standards and discrepancy together measure the over-arching construct, perfectionism.

Table 7: Confirmatory factor analysis goodness-of-fit results for the SAPS for all respondents, and by gender.

Group	n	df	ChiSq	pChiSq	SRMR	RMSEA	CFI	TLI
All	582	19	50.24	0.000	0.03	0.05	0.99	0.98
Women	325	19	41.73	0.002	0.03	0.06	0.96	0.95
Men	257	19	20.32	0.376	0.03	0.02	1.00	1.00

Note: df = degrees of freedom, ChiSq = chi-square, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, CFI = comparative fit index, TLI = Tucker-Lewis Index

Goodness of fit indices (Table 7) indicate that the SAPS had good fit in medical students, with SRMR, RMSEA, CFI and TLI all within limits of good fit. Note that this good fit is driven by the men students, whose fit was consistently better than women's. RMSEA in particular meets criteria for excellent fit in men (<0.06) but good fit for women (<0.08). Men additionally had a non-significant chi-square, although their smaller sample size may be contributing to that effect. (Brown, 2015, p. 69).

Factor loadings for both models are presented in Table 8. All freely estimated factor loadings (Standards 8 and Discrepancy 11 are indicator items) were significant at the p<0.001

level, indicating that all the items load significantly onto their respective latent variables. R-squared for all items indicated that at least 40% of the variance in responses to each item was explained by the underlying factors. In particular, over 80% of the variance in Discrepancy 16 "My performance rarely measures up to my standards," and Discrepancy 20 "I am hardly ever satisfied with my performance" was explained by the latent variable, Discrepancy. As in the APS-R, visual inspection suggests that women load less strongly on the Standards items than men. To determine whether this indicates a significant difference in loadings, we turn to MI.

Table 8: Factor loadings for the SAPS for all respondents, and standardized loadings by gender.

Loadings for All Respondents				Loadings by Gender	
Item	Standardized	p	R^2	Women	Men
stand_8	0.80	<0.001	0.64	0.75	0.86
stand_12	0.82	<0.001	0.68	0.82	0.83
stand_14	0.69	<0.001	0.47	0.64	0.75
stand_22	0.70	<0.001	0.49	0.67	0.75
disc_11	0.87	<0.001	0.76	0.87	0.87
disc_16	0.89	<0.001	0.80	0.89	0.90
disc_20	0.91	<0.001	0.83	0.92	0.91
disc_23	0.75	<0.001	0.57	0.75	0.76

Measurement invariance with multi-group CFA in the SAPS

Table 9 presents the results of the MI analysis, and includes both the parameters for the individual models (goodness of fit) and the change in parameters between successive nested models (difference in fit) using the Santorra-Bentler method.

As with the APS-R, configural and metric invariance were present, but the SAPS model failed at the scalar level (p=0.016), indicating non-invariance of intercepts. As with the APS-R, this signals that the regression slopes are equal, but the intercepts are not. As a result,

comparison of latent means for men and women is not interpretable as a difference in that trait by gender. As measurement non-invariance was found at the scalar level, testing was stopped at this level and strict invariance was not tested.

Table 9: Measurement invariance results for the SAPS by gender

SAPS	Goodness of Fit						Difference in Fit	
	df	RMSEA	SRMR	CFI	TLI	ChiSq	ChiSq diff	p
Configural	38	0.05	0.03	0.99	0.99	61.31		
Metric	44	0.05	0.04	0.99	0.99	69.67	8.36	0.213
Scalar	50	0.05	0.04	0.98	0.98	85.33	15.66	0.016

Note: ChiSq = chi-square, df = degrees of freedom, CFI = comparative fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual, diff = difference

Differential item functioning in the SAPS:

Having failed to achieve scalar invariance, we looked to differential item functioning to determine potential contributions to noninvariance between men and women on the APS-R.

Table 10: Differential item functioning in the SAPS by gender

item	#cat	DIF present	total DIF effect*	uniform DIF*	non-uniform DIF*	R^2
stand_8	3		0.117	0.134	0.153	0.002
stand_12	4		0.287	0.478	0.158	0.000
stand_14	4		0.085	0.147	0.093	0.002
stand_22	4		0.794	0.497	0.984	0.000
disc_11	7		0.773	0.788	0.506	0.000
disc_16	7		0.981	0.999	0.844	0.000
disc_20	7	YES	**0.003**	**0.001**	0.815	0.007
disc_23	7		0.547	0.373	0.520	0.000

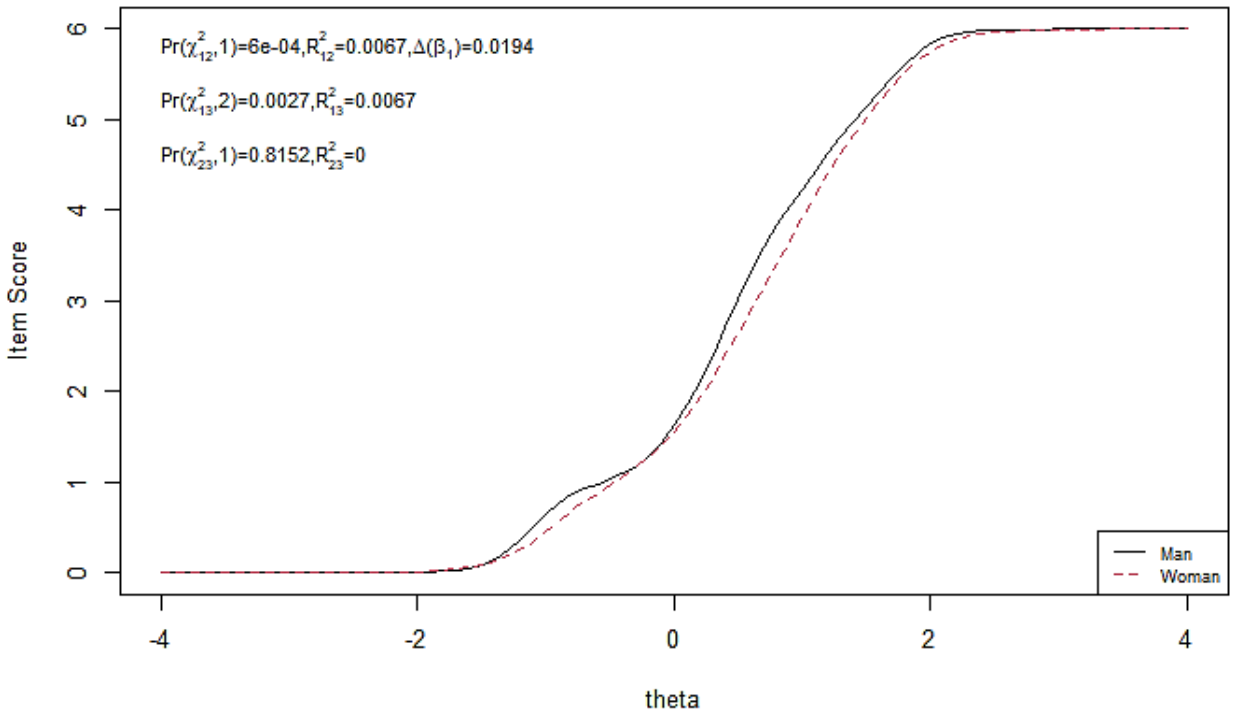
Note: #cat = number of categories for that item, R^2 = McFadden's effect size
*p values for model comparisons, bolded p<0.05

Table 10 shows the DIF results produced using the lordif package in R (Choi et al., 2011) for ordinal logistic regression in polytomous items. Testing was completed for the Standards

items separately from the Discrepancy items. Using an alpha level of 0.05, one item exhibited DIF, Discrepancy 20, "I am hardly ever satisfied with my performance." This is consistent with the DIF results for the APS-R that also found uniform DIF for this item. McFadden's R^2 is consistently less than 0.013, indicating that although DIF is present, its effect is small (Jae Jeong, 2016). Figure 6 shows the DIF curve for Discrepancy 20 in SAPS.

Figure 6: Item characteristic curves for DIF items in the SAPS

Discrepancy 20: "I am hardly ever satisfied with my performance."



DISCUSSION

This study shows that the models for perfection as determined by the Almost Perfect Scale – Revised (APS-R) and the Short Almost Perfect Scale (SAPS) hold in medical students, albeit marginally for the APS-R. The measurement invariance analyses by gender indicates that scalar invariance is violated for both scales. Further DIF analyses reveals that that is mainly due to the Discrepancy items, four in the APS-R, and one in the SAPS. Overall, this research suggests the SAPS should be a better instrument for comparing perfectionism in men and women medical students, although direct comparison of group means should be exercised with caution.

Relative to the general population, our medical students showed similar scores for Discrepancy (averaging ~3.4), but higher scores on Standards, averaging ~6.3 (Table 2) (typical scores being closer to 6.0 (Rice and Ashby). While high Standards is consistent with other findings in medical students, other studies indicate lower rates of maladaptive perfectionism relative to the general population (Enns et al., 2001; Thomas & Bigatti, 2020). In medical students, maladaptive perfectionism is associated with imposter syndrome and depression (Bußenius & Harendza, 2019; Thomas & Bigatti, 2020), and connections are beginning to be made between perfectionism and coping with mistakes (Leung et al., 2019), emphasizing the importance of both measuring these traits in students and ensuring measurement invariance in these scales.

The APS-R has been tested in multiple languages and cultures (Kira et al., 2018; Wang et al., 2007), largely in convenience samples of university students that consisted predominantly of women. Although medical students are typically just one year older than some undergraduate students, the drivers and incentives required for successful medical school application may attract people whose conceptions of perfection differ (Enns et al., 2001). We found that despite

frequent good fit in university populations, the APS-R had marginal fit in our medical student respondents. What constitutes "good fit" is a matter of contention between authors (Hooper et al., 2008; Nye & Drasgow, 2011), most of whom recommend considering a cluster of measures, because current measures both have flaws and measure different aspects of fit. For example, chi-square is very sensitive to sample size, and therefore even though our evaluation of the APS-R "failed" chi-square, we chose to favor of SRMR and TLI, which also test absolute fit of the model without that sensitivity (Brown, 2015, p. 70). Further, we based our assessment of "marginal fit" for the APS-R largely on an RMSEA approaching a cutoff of 0.08, but other authors have proposed cut-offs that were as low as 0.05 (Xia & Yang, 2019), and another fair interpretation of our results would be to say that the APS-R failed on that measure in medical students. Improving goodness of fit was one of the stated drivers for the creation of the SAPS, which removed ambiguous and duplicative items in order to improve fit (K. G. Rice et al., 2014). This tactic should directly improve RMSEA, as its formula contains a parsimony correction (Brown, 2015, p. 71). Our study shows that limiting the items to the eight found in the SAPS did improve fit, with all (excepting chi-square) fit indices within accepted ranges.

When considering these scales by gender, we find that both models consistently fit men better than women (Tables 3 and 7), even in the SAPS. This is important, because previous work has found differences in APS-R scores by gender (K. G. Rice & Ashby, 2007), and knowing whether this is a true difference in *score*, rather than interpretation of items or construct, depends on examining measurement invariance. Unfortunately, neither scale achieved measurement invariance at the intercept level by gender, indicating APS-R and SAPS items may have a different meaning to each group. As this gender noninvariance is inconsistent with the original SAPS work (K. G. Rice et al., 2014), it may be a unique function of medical students. For the

five items exhibiting DIF, no obvious issues with the wording in these questions (like "crying" in depression scales) are apparent to the authors. Two of those four DIF items are similar questions regarding satisfaction: Discrepancy 13 and Discrepancy 20, the latter of which is the item that remains problematic in the SAPS. Flett (Flett et al., 2016) thought that asking about satisfaction – or rather *dis*-satisfaction – could be problematically measuring something other than pure Discrepancy. It's possible that a "satisfaction" component is differently interpreted by gender, although one scale measuring satisfaction (Satisfaction with Life Scale) shows good MI by gender (Emerson et al., 2017).

When Rice and colleagues developed the SAPS (K. G. Rice et al., 2014), they did not test DIF by gender in the APS-R, but did find measurement invariance by gender in their tests of the SAPS. When choosing items for the SAPS, they specifically considered Discrepancy 20 (our DIF item) and Discrepancy 21 for inclusion. They found that the correlation for those items was high (r=0.81), concluded that either item would be appropriate for inclusion, and choose Discrepancy 20 over Discrepancy 21 because it was shorter in length. Our APS-R analysis did not indicate gender related DIF in Discrepancy 21. Further work should consider whether a change in SAPS is appropriate.

Finally, we can ask if these gender differences are relevant. The violation of scalar invariance indicates that the means can't be meaningfully compared by gender because we can't determine whether the difference is due to person-trait difference or measurement difference. While Standards means for women and men (Table 2) were nearly equal (and contained only one DIF item in the APS-R and zero DIF items in SAPS), the Discrepancy means differed by 0.41 and 0.32, and had four and one DIF items respectively. Conceivably, this could be amplifying a difference between means, potentially leading to Type 1 error. Additionally, the mean

Discrepancy score needed to be categorized as a maladaptive perfectionist is 3.5. The group means for our respondents (Table 2) hover about that score, with men consistently below it, and women nearly at or above it. Again, it is conceivable that a small additive effect could cause either gender to be miscategorized.

This study has several limitations. First, data was shown to be missing-not-at-random without a solution. In particular, it appeared that a group of low-Standards, high-Discrepancy students filled out just a few answers. In an attempt to include that group, we included 40 responses (of the 582 total) with <50% complete data, which resulted in individual items being between 85% and 98% complete overall. This solution is imperfect, especially the use of FIML in MNAR data (Peugh & Enders, 2004), and the use of summed scores in DIF could affect results by substituting the mean for missing responses. Future work should encourage participation of as many students as possible.

Second, for this study, all students took the 23-item APS-R (including the Order items). For the SAPS scale analysis, the authors selectively analyzed the SAPS items. As context matters – responses on individual items influence choices on adjacent items (Şahin, 2021) – future iterations should include a group of students that takes only the 8-item SAPS.

Third, this work represents the first-year medical students in a single, private medical institution in the Midwest United States; therefore caution should be used if generalizing these findings to other medical schools or other years in medical school. Further, this study considered only binary gender, and discarded the responses from the low numbers of students (less than five) sharing a gender that was non-binary. As best practices expand regarding survey items for non-binary gender, guidelines remain elusive regarding statistical best practices for working with those voices, as they are often few within the data set. Authors have advocated for sampling

methods that increase sample size (Glick et al., 2018), or have added non-binary voices arbitrarily to the group of women (essentially changing that group to non-cis men), both of which are imperfect solutions. In future work, the author recommends asking a follow up question to non-binary gendered persons regarding what they would like done with their responses.

References

Ashby, J. S., Rice, K. G., & Kutchins, C. B. (2008). Matches and mismatches: Partners, perfectionism, and premarital adjustment. *Journal of Counseling Psychology*, *55*(1), 125–132. https://doi.org/10.1037/0022-0167.55.1.125

Blacker, C. J., Lewis, C. P., Swintak, C. C., Bostwick, J. M., & Rackley, S. J. (2019). Medical Student Suicide Rates: A Systematic Review of the Historical and International Literature. *Academic Medicine*, *94*(2), 7.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (Second). The Guilford Press.

Bußenius, L., & Harendza, S. (2019). The relationship between perfectionism and symptoms of depression in medical school applicants. *BMC Medical Education*, *19*(370), 1–8. https://doi.org/10.1186/s12909-019-1823-4

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), 1–30.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R. D., & Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, *16*(S1), 69–84. https://doi.org/10.1007/s11136-007-9185-5

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–149. https://doi.org/10.1177/0748175610373459

Emerson, S. D., Guhn, M., & Gadermann, A. M. (2017). Measurement invariance of the Satisfaction with Life Scale: Reviewing three decades of research. *Quality of Life Research*, *26*(9), 2251–2264. https://doi.org/10.1007/s11136-017-1552-2

Enns, M. W., Cox, B. J., Sareen, J., & Freeman, P. (2001). Adaptive and maladaptive perfectionism in medical students: A longitudinal investigation. *Medical Education*, *35*(11), 1034–1042.

Flett, G. L., Mara, C. A., Hewitt, P. L., Sirois, F., & Molnar, D. S. (2016). How Should Discrepancy Be Assessed in Perfectionism Research? A Psychometric Analysis and Proposed Refinement of the Almost Perfect Scale–Revised. *Journal of Psychoeducational Assessment*, *34*(7), 718–732. https://doi.org/10.1177/0734282916651382

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research*, *14*(5), 449–468. https://doi.org/10.1007/BF01172967

Glick, J. L., Theall, K., Andrinopoulos, K., & Kendall, C. (2018). For Data's Sake: Dilemmas in the Measurement of Gender Minorities. *Culture, Health & Sexuality*, *20*(12), 1362–1377. https://doi.org/10.1080/13691058.2018.1437220

Hamachek, D. E. (1978). Psychodynamics of normal and neurotic perfectionism. *Psychology: A Journal of Human Behavior*, *15*(1), 27–33.

Hankir, A. K., Northall, A., & Zaman, R. (2014). Stigma and mental health challenges in medical students. *BMJ Case Reports*, *2014*, bcr2014205226. https://doi.org/10.1136/bcr-2014-205226

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, *6*(1), 53–63.

Hu, K. S., Chibnall, J. T., & Slavin, S. J. (2019). Maladaptive Perfectionism, Impostorism, and Cognitive Distortions: Threats to the Mental Health of Pre-clinical Medical Students. *Academic Psychiatry*, *43*(4), 381–385. https://doi.org/10.1007/s40596-019-01031-z

Jeong, H. (2016). Does Differential Item Functioning Occur Across Respondents' Characteristics in Safety Attitudes Questionnaire? *Biometrics & Biostatistics International Journal*, *4*(3), 103-111. https://doi.org/10.15406/bbij.2016.04.00097

Kira, I., Shuwiekh, H., Rice, K., & Ashby, J. (2018). Is the "Almost Perfect Scale" Almost Perfect? The Psychometric Properties of the Arabic Version of APS-R and Its Short Form. *Psychology*, *9*(7), 1875–1897. https://doi.org/10.4236/psych.2018.97109

Leung, J., Cloninger, C. R., Hong, B. A., Cloninger, K. M., & Eley, D. S. (2019). Temperament and character profiles of medical students associated with tolerance of ambiguity and perfectionism. *PeerJ*, *7*, e7109. https://doi.org/10.7717/peerj.7109

Limburg, K., Watson, H. J., Hagger, M. S., & Egan, S. J. (2017). The Relationship Between Perfectionism and Psychopathology: A Meta-Analysis. *Journal of Clinical Psychology*, *73*(10), 1301–1326. https://doi.org/10.1002/jclp.22435

Lins de Holanda Coelho, G., Pereira Monteiro, R., Vilar, R., H. P. Hanel, P., Cunha Moizéis, H. B., & Gouveia, V. V. (2021). Psychometric Evidence of the Short Almost Perfect Scale (SAPS) in Brazil. *The Counseling Psychologist*, *49*(1), 6–32. https://doi.org/10.1177/0011000020949146

Nye, C. D., & Drasgow, F. (2011). Assessing Goodness of Fit: Simple Rules of Thumb Simply Do Not Work. *Organizational Research Methods*, *14*(3), 548–570. https://doi.org/10.1177/1094428110368562

Peugh, J., & Enders, C. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, *74*, 525–556. https://doi.org/10.3102/00346543074004525

Quek, T. T.-C., Tam, W. W.-S., Tran, B. X., Zhang, M., Zhang, Z., Ho, C. S.-H., & Ho, R. C.-M. (2019). The Global Prevalence of Anxiety Among Medical Students: A Meta-Analysis.

International Journal of Environmental Research and Public Health, *16*(15), E2735. https://doi.org/10.3390/ijerph16152735

Rice, K. G., & Ashby, J. S. (2007). An Efficient Method for Classifying Perfectionists. *Journal of Counseling Psychology*, *54*(1), 72–85. https://doi.org/10.1037/0022-0167.54.1.72

Rice, K. G., Park, H., Hong, J., & Lee, D. (2019). Measurement and Implications of Perfectionism in South Korea and the United States. *The Counseling Psychologist*, *47*(3), 384–416. https://doi.org/10.1177/0011000019870308

Rice, K. G., Richardson, C. M. E., & Tueller, S. (2014). The Short Form of the Revised Almost Perfect Scale. *Journal of Personality Assessment*, *96*(3), 368–379. https://doi.org/10.1080/00223891.2013.838172

Rice, K. G., & Taber, Z. B. (2019). Measurement Invariance and Latent Profiles of Perfectionism in Clients and Nonclients. *Journal of Counseling Psychology*, *66*(2), 210–223.

Rice, S. P. M., Loscalzo, Y., Giannini, M., & Rice, K. G. (2020). Perfectionism in Italy and the USA: Measurement invariance and implications for cross-cultural assessment. *Journal of Psychological Assessment*, *36*(1), 207–211. https://doi.org/10.1027/1015-5759/a000476

Robertson, J. J., & Long, B. (2018). Suffering in Silence: Medical Error and its Impact on Health Care Providers. *The Journal of Emergency Medicine*, *54*(4), 402–409. https://doi.org/10.1016/j.jemermed.2017.12.001

Rosseel, Y. (2012). LAVAAN: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y. (2021, June 27). *The lavaan Project*. LAVAAN Latent Variable Analysis Version 0.6-9. https://lavaan.ugent.be/about.html

Şahin, M. D. (2021). Effect of Item Order on Certain Psychometric Properties: A Demonstration on a Cyberloafing Scale. *Frontiers in Psychology*, *12*, 590545. https://doi.org/10.3389/fpsyg.2021.590545

Slaney, R. B., Ashby, J. S., & Trippi, J. (1995). Perfectionism: Its Measurement and Career Relevance. *Journal of Career Assessment*, *3*(3), 279–297.

Slaney, R. B., Rice, K. G., Mobley, M., Trippi, J., & Ashby, J. S. (2001). The Revised Almost Perfect Scale. *Measurement and Evaluation in Counseling and Development*, *34*(3), 130–145.

Stoeber, J., & Otto, K. (2006). Positive Conceptions of Perfectionism: Approaches, Evidence, Challenges. *Personality and Social Psychology Review*, *10*(4), 295–319. https://doi.org/10.1207/s15327957pspr1004_2

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research*, *16*(S1), 33–42. https://doi.org/10.1007/s11136-007-9184-6

The R Core Team. (2021). *R: The R Project for Statistical Computing*. https://www.r-project.org/

Thomas, M., & Bigatti, S. (2020). Perfectionism, impostor phenomenon, and mental health in medicine: A literature review. *International Journal of Medical Education*, *11*, 201–213. https://doi.org/10.5116/ijme.5f54.c8f8

UCLA: Statistical Consulting Group. (n.d.). *Confirmatory Factor Analysis (CFA) in R with lavaan*. UCLA Institue for Digital Research and Education. Retrieved July 31, 2021, from https://stats.idre.ucla.edu/r/seminars/rcfa/

Wang, K. T., Permyakova, T. M., & Sheveleva, M. S. (2016). Assessing perfectionism in Russia: Classifying perfectionists with the Short Almost Perfect Scale. *Personality and Individual Differences*, *92*, 174–179. https://doi.org/10.1016/j.paid.2015.12.044

Wang, K. T., Slaney, R. B., & Rice, K. G. (2007). Perfectionism in Chinese university students from Taiwan: A study of psychological well-being and achievement motivation. *Personality and Individual Differences*, *42*(7), 1279–1290. https://doi.org/10.1016/j.paid.2006.10.006

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2