TPM®

ASSESSMENT OF TRAITEDNESS: AN ITEM RESPONSE THEORY APPROACH

EMINE ÖNEN
GAZI UNIVERSITY, ANKARA, TURKEY

In this study, the effectiveness of the person-fit statistics (l_z^p) based on the Graded Response Model, the person discriminal dispersion parameter, and the person reliability parameter based on Dual Thurstonian Models for graded responses, and the responder-specific discrimination parameter based on the Graded Response Differential Discrimination Model in the assessment of traitedness were comparatively investigated. Data were obtained by administering the STAI-State and STAI-Trait inventories (Spielberger et al., 1983) to undergraduates (N = 1102) attending the Education faculty of a public university in Ankara, Turkey. The results indicated that person-fit statistics (l_z^p), person discriminal dispersion parameter, and person reliability parameter might reflect traitedness. In addition, the person reliability parameter was observed to reflect traitedness the most.

Keywords: Traitedness; Person reliability; Person-fit; Individual discrimination; Anxiety.

Correspondence concerning this article should be addressed to Emine Önen, Department of Educational Sciences, Division of Measurement and Evaluation in Education, Gazi University, 06500 Teknikokullar, Ankara, Turkey. Email: emineonen@gazi.edu.tr

There are individual differences in the degree to which a trait is internally and strongly organized by the individual, and the traitedness notion indicates this. The studies conducted by Allport (1937) and Bem and Allen (1974) affected the construction of this concept, and researchers have formed this concept as the relevance of the trait with the individual. The level of internalization of the trait by an individual is related to whether the individual demonstrates consistent behaviors across situations (Bem & Allen, 1974). Tellegen (1988) stated that individuals differ in terms of the degree to which their traits are related to themselves, and the notion of traitedness refers to that. Individual differences in "trait-relevance" could substantially affect the validity of the measures of the affective traits and the usability of these measures for several purposes. In the guidelines of the International Test Commission (2012), investigation of the test score validity is recommended by examining response patterns that could be unrelated to the person's trait level. These explanations show the importance of the investigation of traitedness in educational and psychological assessment (Cucina & Vasilopoulos, 2005).

Therefore, traitedness has been assessed by using several approaches. First in the study of Bem and Allen (1974) by asking individuals how much the measured traits were related to them and how variable they were in terms of this trait across conditions. Baumeister and Tice (1988) suggested using "inter-item variance" to classify individuals as traited or untraited. Chaplin and Locklear (as cited in Warner, 2005) designed in 1989 the "construct similarity" index to assess traitedness. A suggested and widely used way to measure traitedness is to examine the consistency of the item responses. Because traited individuals have a strong perception of their levels in terms of the trait being measured, they will produce a consistent response pattern by responding to items in a way that reflects their traits (Tellegen, 1988). Related to these proposed approaches, various criticisms and limitations are noted in the literature (Cucina & Vasilopoulos, 2005).

However, Item Response Theory (IRT)-based measures seem promising in assessing traitedness in the context of response consistency.

IRT models have been widely used in modeling the responses to tests and scales used in educational and psychological research. Levine and Drasgow (1983) named standard IRT models in which only a single ability parameter was estimated, "constant models." In constant-θ IRT models, the assumption that the individual's latent trait is constant across items is accepted. In applying constant-θ IRT models to real data, the consistency of item response pattern for a given IRT model should be assessed, indicating person-fit. Reise and Waller (1993) stated that likelihood-based person-fit statistics could be used to assess traitedness. Drasgow et al. (1985) proposed the IRT-based l_z person-fit statistics (l_z^p) for polytomous items to examine response consistency.

In this study, the effectiveness of the l_z^p as a person-fit statistic in measuring traitedness was assessed on the base of the Graded Response Model (GRM; Samejima, 1969), and was considered one of the constant-θ IRT models. The main reason for preferring GRM is that in the literature (Ferrando, 2019; Lubbe & Schuster, 2017) it is accepted as a normative counterpart of the other models tested in this study. In GRM, cumulative boundary functions (P^*_{ig}) could be represented by two mathematical functions: logistic distribution function and normal distribution function. As in the literature (Ostini & Nering, 2006; Samejima, 1969), in this study the model was named logistic GRM when the logistic function was used, and it was named normal ogive GRM when the normal function was used. In this study, using GRM (logistic) parameter estimates, the l_z^p values were computed to investigate response consistency. Statistic l_z^p represents the likelihood of an item-score pattern of a respondent at a given latent trait level under a specific unidimensional IRT model and is calculated by using the following equation:

$$l_z^p(x) = \frac{l^p(x) - \mathrm{E}[l^p(x)]}{(\mathrm{VAR}[l^p(x)])^{\frac{1}{2}}} \quad (1)$$

In this equation, while $\mathrm{E}[l^p(x)]$ represents the expected value, $\mathrm{VAR}[l^p(x)]$ represents the variance (Conijn et al., 2013; Emons, 2009). The person response curve for a respondent with low inter-item trait variability will be steeper. In this case, individuals are expected to endorse most items with item location lower than their trait level; they are expected not to endorse the majority of the items with item location higher than their trait level. Such a response pattern will produce positive-high l_z^p values indicating person fit (Ferrando, 2004). So, the l_z^p statistic could be considered to indirectly reflect "trait variability" (LaHuis et al., 2017). Person-fit statistics have been computed based on constant-θ models in which an individual's latent trait is assumed constant across items. When affective attributes are measured, this assumption may not be reasonable, and violating this assumption could affect the validity of latent trait estimations. Based on this assumption, in constant-θ IRT models, a single-person parameter is estimated, which indicates an individual's location on the theta continuum in terms of the trait being measured. However, the literature (Ferrando, 2004, 2019) has shown that individuals respond with different sensitivity to items on different locations of the theta continuum. Nevertheless, in constant-θ IRT models, there is no parameter to capture this difference in sensitivity for individuals, that is, a person parameter corresponding to the item discrimination parameter (Navarro-Gonzalez & Ferrando, 2019).

In the "dual" modeling (Fiske, 1968) approach, a "dual" person discrimination parameter, which corresponds to the item discrimination parameter and reflects the individual difference in the reliability of responses, is estimated. Ferrando (2019) stated that "dual" models could be divided into two basic families: (1) Thurstonian models (TMs) and (2) Multiplicative models (MMs). In Thurstonian models there is a central trait level, and person discrimination is modeled as random fluctuations around this level. In MMs, it is modeled as a person slope. These fluctuations modeled in TMs indicate the temporal changes in the individuals' theta levels, which can be sourced from the uncertainty in their perceptions of the extent

to which they have the trait being measured (Ferrando, 2014a). This uncertainty could be an important "individual-difference variable" which reflects traitedness in terms of "trait-relevance" (LaHuis et al., 2017). Therefore, Dual TMs for graded responses (DTGRM; Ferrando, 2019) is considered an appropriate model to investigate traitedness.

Another model used in this study to investigate traitedness is the Graded Response Differential Discrimination Model (GRDDM), proposed by Lubbe and Schuster (2017) and considered within the MMs family. In GRDDM, the person slope parameter provides information about whether the respondents discriminate between items while responding. This situation identifies the sensitivity of the individuals in responding to items according to different item locations on the latent trait continuum. A high sensitivity points to respondents with a well-defined trait level, while a low sensitivity points to respondents who do not respond to items by their trait level. The high value of the estimated person discrimination in this model indicates that item responses and traits are closer to each other than the situation with the low value. Therefore, the parameter indicating "individual discrimination" in this model was considered appropriate to be used to assess traitedness (Ferrando, 2014a; Lubbe & Schuster, 2017). Accordingly, this study aimed to comparatively investigate to what extent person-fit statistics based on a constant-θ model (GRM), person reliability, and person discriminal dispersion parameters based on Ferrando's (2019) DTGRM and the "responder-specific discrimination" parameter based on GRDDM (Lubbe & Schuster, 2017) may reflect traitedness.

Dual Thurstonian Graded Response Model (DTGRM)

This model was developed by Ferrando (2019) for graded responses based on the Dual Thurstonian approach and IRT and was denoted as DTGRM. In this model, temporary changes (person fluctuation) are allowed in the individual's latent trait level between items. Suppose n items in graded form (with m response categories) developed to measure trait level (θ) are administered to person i. This respondent's trait when he/she responds to item j is called "momentary trait" and is denoted as T_{ij}. In DTGRM, the momentary trait, T_{ij}, is assumed to change around the person's central location (θ_i) by exhibiting a normal distribution with mean θ_i and σ_i^2 variances over the items. Therefore, in this model, two person parameters are estimated as θ_i "person location" parameter and σ_i^2 parameter measuring "person fluctuation" (Ferrando, 2019). As claimed by Ferrando (2019), in DTGRM for fixed θ_i and σ_i^2, the probability of endorsing category k at item j is calculated as:

$$P(X_{ij} = k|\theta_i, \sigma_i^2) = \Phi\left(\frac{1}{\sqrt{\sigma_i^2 + \sigma_{\varepsilon j}^2}}(\theta_i - (\beta_j + \frac{\tau_{jk-1}}{\alpha_j}))\right) - \Phi\left(\frac{1}{\sqrt{\sigma_i^2 + \sigma_{\varepsilon j}^2}}(\theta_i - (\beta_j + \frac{\tau_{jk}}{\alpha_j}))\right) =$$
$$\Phi\left(\Upsilon_{ij}\,(\theta_i - \delta_{jk-1})\right) - \Phi\left(\Upsilon_{ij}\,(\theta_i - \delta_{jk})\right) \quad (2)$$

Ferrando (2019) called σ_i^2 parameter "person-discriminal dispersion" (PDD) and $\sigma_{\varepsilon j}^2$ parameter "item-discriminal dispersion" (IDD). An individual with a low PDD estimate would consistently respond highly according to the different item locations, which indicates that this person has a well-defined trait level. The inverse of the σ_i parameter ($\Upsilon_i = 1/\sigma_i$) is considered the "person reliability parameter" (PRP). In this model, person reliability is considered an "individual-difference variable" reflecting how the trait is represented inside the individual internally, strongly, and clearly. The literature (Ferrando, 2019; LaHuis et al., 2018), therefore, has shown that PRPs may reflect traitedness. In the case of PDDs that are the same for all respondents, while IDDs are allowed to differ, the model is reduced to Samejima's (1969) restricted normal ogive GRM. When IDDs are not allowed to vary across items while PDDs are allowed to vary across

respondents, the model would be DTGRM (Ferrando, 2009, 2019). Therefore, as the standard counterpart of DTGRM, Samejima's (1969) restricted normal ogive GRM is accepted.

Graded Responses Differential Discrimination Model (GRDDM)

GRDDM has been developed by Lubbe and Schuster (2017) by modifying the Differential Discrimination Model (DDM; Ferrando, 2014a) for graded responses. Researchers have stated that they have translated the DDM into a structural equation modeling (SEM) framework. In this framework, they have re-expressed the observed categorical responses (y^*_{ij}) as continuous latent responses (y^*_{ij}). For model identification purposes, a constraint was put on thresholds and the variance of y^*_{ij} was fixed as 1.0, and scale midpoint (c) was fixed as 0. In the SEM framework, the model is expressed as:

$$y^*_{ij} = \gamma_j \tau_i + \delta_j \alpha_i + e_{ij} \quad (3)$$

In this equation, γ_j denotes item loading, α_i denotes responder-specific discrimination, and e_{ij} denotes zero-mean residuals. τ_i is named "scaled trait" and equals to multiply the latent trait level (θ_i) by the responder-specific discrimination ($\tau_i = \alpha_i \theta_i$). δ_j, which shows the difference between item mean and scale midpoint, is equal to $\delta_j = -\beta_j \gamma_j$. In addition to modeling individual discrimination in terms of sensitivity to item-person distance, the responder-specific discrimination parameter could also reflect the "extreme response tendency." Because in GRDDM, individual discrimination is modeled as a person slope, it also indicates the individual differences in "response scale usage" (Ferrando, 2016, 2019; Lubbe & Schuster, 2017). In this study, α_i values have been considered in the context of "individual discrimination" and "response scale usage."

Because the person slope parameter in this model modifies the expected responses, the high value of α_i ($\alpha_i > 1$) shows "increased spaces between responses," and the low values of α_i (values close to 0.0) show that respondents do not discriminate between items when responding (Lubbe & Schuster, 2017). This fact suggests that they were highly insensitive in responding to items located at different places on the latent trait continuum. This implies that the respondents do not respond in accordance with their trait level, and their response patterns are almost random. So, this indicates a low level of traitedness, and α_i value may somehow reflect traitedness (Ferrando, 2019; Lubbe & Schuster, 2017). The responder-specific discrimination (a_{ij}) based on the α_i values can be obtained using the following equation:

$$a_{ij} = \frac{\alpha_i \gamma_j}{u_j} \quad (4)$$

Researchers have also altered GRDDM by using the IRT notation according to Samejima's GRM:

$$P(y_{ij} \geq k | \theta_i, \alpha_i) = \Phi(a_{ij}(\theta_i - \beta_j) - k^*_{jk}),\ k^*_{jk} = u_j^{-1} k^*_{jk} \quad (5)$$

When compared to the equation of the normal ogive version of GRM:

$$P(y_{ij} \geq k | \theta_i) = \Phi(a_j(\theta_i - \beta_{jk}) \quad (6)$$

it is understood that when $\alpha_i$values are equal to 1 for all respondents (therefore $\sigma^2_\alpha = 0$), there will be item-specific discrimination but not person-specific discrimination anymore. Thus, the normal ogive version of GRM and GRDDM would be the same.

Comparing Constant-θ IRT Models and Dual Models in Assessing Traitedness

PDD and PRP show similarities in terms of reflecting response consistency with the likelihood-based person fit-statistics. However, PDD and PRP directly reflect "trait-variability," and likelihood-based person fit-statistic indirectly reflects "trait variability." Trait variability is considered to be sourced from the uncertainty in the individual's perceptions of the extent to which the trait is measured. The individuals with less trait variability are those with well-defined trait levels, namely "traited" individuals (Ferrando, 2019; LaHuis et al., 2018). Although DTGRM-based PDD (σ_i) and GRDDM-based responder-specific discrimination (α_i) parameters are generally considered in the context of individual discrimination, these parameters differ a little in terms of the modeling and what they indicate. In GRDDM, α_i is modeled as a slope parameter (person slope). Thus, the expected responses for individuals having different discriminations also differ between trait levels. The differences in α_i estimates between respondents show the individual differences in the reliability of the responses of these persons to the items. In DTGRM, the σ_i is modeled as a person-specific scale parameter; therefore, the expected responses do not change between different trait levels. Instead, this parameter reflects the consistency of expected responses and measures person reliability.

When the studies using the IRT approach to assess traitedness were reviewed, very few studies were found (LaHuis et al., 2017, 2018; Reise & Waller, 1993). Reise and Waller (1993) showed that individual differences in traitedness based on personality measures could be examined with an IRT-based "scalability index" (Z_L index). Another study (LaHuis et al., 2018) assessed how PRPs estimated based on a variable-θ IRT model (Vθ-GRM) reflect traitedness. In addition to the fact that PRPs may reflect traitedness, the validity was higher for highly traited individuals. LaHuis et al. (2017) found medium to high positive relations between PRP estimates and person-fit statistics (l_z^p). They stated that these relations indicating that PRPs could be used to measure traitedness were observed. Although Ferrando (2004) was not directly interested in traitedness, he reported strong relations between PRPs and person-fit statistics. Relations were also observed between the measures of consciousness and impulsiveness and PRP estimates, which may indicate traitedness (Ferrando, 2009). In addition, he maintained that for the children with high PRP estimates, both trait estimates were more accurate, and stronger relations were observed between test anxiety and test performance (Ferrando, 2016). However, no studies compare the performances of the parameters that suggest individual discrimination estimated based on DTGRM and GRDDM in reflecting traitedness. Investigating to what extent person discrimination could measure traitedness when it is modeled as a scale parameter and as a person slope parameter may be beneficial. Accordingly, in this study, the performances of assessing the traitedness of person-fit statistics, PDD, PRP parameters, and responder-specific discrimination parameters were comparatively investigated.

METHODS

Participants

This study was not aimed to generalize the findings to the population, so a sample was not drawn using a sampling method. Instead, participants were recruited from a state university (Gazi) located in Ankara (the capital of Turkey) where the majority of students belong to different socioeconomic statuses and subcultures, because this might reflect the variety in terms of anxiety (an ethical approval form for the study was obtained from the Gazi University Ethical Committee, Number: E-77082166-604.01.02-223446 26.11.2021). This study was conducted with 1102 students (714 female, 388 male) attending several undergraduate

programs at the Education faculty of this university. The reason why students from the faculty of Education were preferred was that the Educational Psychology course was taught in all undergraduate programs of the faculty of Education and therefore students are familiar with psychological characteristics and especially "anxiety." The participants were first-grade students (n = 211), second-grade students (n = 186), third-grade students (n = 363), and fourth-grade students (n = 342). The study data comprised 1102 observations, and with no missing value (missing data were removed from the analysis). Participants were informed about the research (informed consent was obtained from all participants included in the study), and the study was carried out exclusively with volunteers.

Data and Procedure

The State-Trait Anxiety Inventory (STAI; Spielberger et al., 1983) has two subscales that measure both state anxiety and trait anxiety: STAI-State (STAI-S) and STAI-Trait (STAI-T). Each subscale comprises 20 items rated on a 4-point Likert scale (from 1 = *not at all* to 4 = *very much so*). In the original study, internal consistency coefficients for these measures ranged from α = .86 to α = .95 (Spielberger et al., 1983). Oei et al. (1990) demonstrated that STAI subscales could measure state and trait anxiety separately by obtaining a clear 2-factor solution. Öner and Le Compte (1985) adapted the inventory to Turkish culture. Researchers claimed that internal consistency coefficients varied between α = .82 and α = .85 for the STAI-S measures and between α = .83 and α = .87 for STAI-T. In the study he conducted, Öner (1997) found that the correlation coefficients between STAI measures and other anxiety measures ranged from .52 and .80 for women, from .58 to .79 for men, and from .77 to .84 for the patient sample. In the current study, STAI-S and STAI-T were administered to participants consecutively, and administrations took approximately 40 minutes. These inventories were chosen because they are widely used in hospitals, guidance research centers, and schools in Turkey in diagnosing anxiety disorders, referring students to guidance research centers or hospitals for anxiety disorders, and in research on this subject.

Data Analysis

Checking assumptions of IRT from STAI-S and STAI-T subscales. For the unidimensionality assumption, confirmatory factor analysis was performed in the Mplus version by using robust maximum likelihood estimation (Muthén & Muthén, 1998/2015). The calculated fit indices (CFI = .91 and .96 for STAI-S and STAI-T, respectively; TLI = .90 and .97 for STAI-S and STAI-T, respectively; RMSEA = .08 and .04 for STAI-S and STAI-T, respectively) indicate acceptable fit for both subscales. For the local independence assumption, Yen's (1993) Q3 statistics were computed for each item pair in the STAI-S and STAI-T subscales (separately) in "TAM" and "sirt" packages in R. Yen's Q3 measures (Q3min = –0.20 and Q3max = 0.11 for STAI-S, Q3min = –0.19 and Q3max = 0.14 for STAI-T); they point to no violation for this assumption.

Estimation

The estimation methods used in fitting the models considered in this study were determined based on the literature review (Ferrando, 2019; LaHuis et al., 2017; Lubbe & Schuster, 2017, 2020; Navarro-Gonzalez & Ferrando, 2019). These estimation methods were used to ensure the comparability of the results with

the literature and because no previous study showed that different parameter estimations yield better results in fitting these models.

Initially, normal ogive GRM with constrained item discrimination parameter and normal ogive GRM (unconstrained) were fitted to both STAI-S and STAI-T subscales, and were compared by using the likelihood ratio (LR) test. LR test results suggest that unconstrained normal ogive GRM fitted both subscales better (LR test results for STAI-S is 754.672 with $df = 19$, $p < .000$; LR test results for STAI-T is 280.018 with $df = 19$, $p < .000$). Afterward, to compute person-fit statistics (l_z^p), GRM was fitted by using maximum likelihood estimation with logit link (logistic Graded Response Model), without any constraint on item discrimination. Person-fit statistics were computed by using GRM (logistic) parameter estimates in the "PerFit" package in statistical program R (cut of scores for l_z^p were calculated based on simulation studies in the "PerFit" package). Normal ogive GRM and logistic GRM were estimated using the Mplus Version 7 (Muthén & Muthén, 1998/2015).

In addition, according to GRDDM, as in the study of Lubbe and Schuster (2017), GRM was written as a "one-factorial factor model for ordered categorical responses," and this model was fitted by using robust weighted least square estimation (and with probit link):

$$y_{ij}^{*} = \gamma_j \theta_i + e_{ij} \tag{7}$$

Instead of item discrimination parameter a_j, the standardized loadings, γ_j, were used in this notation (Lubbe & Schuster, 2017). This model was estimated in Mplus using the syntax provided by Lubbe and Schuster (2017). GRDDM was estimated using the robust weighted least square estimation method in the Mplus Version 7 (Muthén & Muthén, 1998/2015). GRM (written as a "one-factorial factor model for ordered categorical responses) and GRDDM were fitted by using the same parameter estimation method. The literature states that in the cases of violation of distributional assumptions, this estimation method was robust in terms of the correction for χ^2 tests for model fit and standard errors. Findings from the simulation studies indicated that for different distributions of the latent variables, this estimation method resulted in consistent parameter estimation and bias that could be ignored (Lubbe & Schuster, 2017). Therefore, this method was selected based on the literature.

DTGRM is fitted using the "InDisc" package available in R (Navarro-González & Ferrando, 2019). In it, DTGRM is fitted by using a "conventional two-stage conditioned procedure" suggested by Ferrando (2019). It consists of two stages: (1) the item calibration stage and (2) the scoring stage. In both, the reason for preferring this method was to keep the methods used in estimating the parameters as simple as possible and to encourage the application of the model to real data (Ferrando, 2019). At the item calibration stage, items are calibrated by fitting a unidimensional FA model to an interitem polychoric correlation matrix using the minimum-residual unweighted least squares method. DTGRM is a model that belongs to the Thurstonian models (TM) family. When the studies on TMs were examined (Ferrando, 2007, 2009, 2014b, 2019), it was seen that the unweighted least squares (ULS) estimation method is used in the item calibration process. The literature states that, despite its limitations, the ULS method is used in fitting FA models, especially in cases where the model is logically correct and large samples are used, because it is both unbiased and provides consistent parameter estimates (Ferrando, 2014b).

At the item calibration stage, item parameters are obtained, and model-data fit is assessed. Then, by using the equation below, the average PDD is estimated:

$$\frac{1-\hat{\alpha}_{(\max)}^2}{\hat{\alpha}_{\max}^2} = \hat{E}(\sigma_i^2) \tag{8}$$

Ferrando (2013) observed that the maximum likelihood (ML) estimation could result in high PDD estimates in estimating person parameters in the original linear model, especially for short tests and where item positions are not evenly distributed. Accordingly, the expected a posteriori (EAP) estimation method was suggested instead of the ML method because it led to more reasonable and finite estimates by using the mean of the PDD estimates resulting from the item calibration process (Ferrando, 2019; Navarro-González & Ferrando, 2019). Therefore, at the scoring stage, Bayes expected a posteriori estimation to be used for θ and σ_i^2 parameters, and the prior distributions for these parameters are specified. For the person location parameter, the standard normal distribution is specified as prior, θ~N(0,1). Because PDDs are variances, the scaled inverse χ^2 distribution is specified as the proper prior for this parameter. The average PDD estimate obtained at the calibration stage is used here as a mean for the prior distribution of PDDs. At the scoring stage, for each respondent, EAP estimations of θ_i, σ_i^2, and their posterior standard deviations (PSDs) are obtained. These posterior standard deviations are considered the standard errors for these parameters. In addition, marginal reliabilities for θ_i and σ_i^2 — $\rho(\hat{\theta})$ and $\rho(\hat{\sigma}^2)$ — are provided as InDisc output.

To assess the appropriateness of DTGRM, initially, the goodness of fit statistics for the unidimensional FA model fitted at the calibration stage should be considered. In addition, the likelihood ratio test result (Λ_i) and the overall index, Q($Q = \sum s_i$), are considered to assess if DTGRM fits better than the corresponding model with constant PDD (Navarro-González & Ferrando, 2019). Ferrando (2019) observed that in the comparison of DTGRM and GRM based on Q statistics, the likelihoods should be assessed at their ML estimates but when fitting DTGRM they are assessed at their EAP estimates. It should be noted here that the differences in parameter estimation methods may affect the accuracy of model comparisons. Based on this limitation of the Q statistic, care should be taken when inferring the appropriateness of the two models.

Assessment of Traitedness

In the studies on person reliability, person-fit, and traitedness, the relationships between person reliability parameter estimates, latent-trait estimates, and person-fit statistics are generally examined, and traitedness is also examined in the context of these relationships (Ferrando, 2004, 2009, 2014a, 2019; LaHuis et al., 2017). As an extension of these studies (also for the sake of comparability with the literature), in the current study, the performances of person-fit statistics and PDD, PRP, α_i (responder-specific discrimination) parameters in reflecting traitedness were assessed by examining the relations between these parameters' estimates and person-fit statistics based on responses to STAI-S and STAI-T items. The relations between these parameters' estimates and person-fit statistics were examined with the Pearson product-moment correlation technique.

Results

In this study, initially, GRM was fitted using the maximum likelihood estimation method, and probit and logit links and results are presented in Table 1. As it is seen in this table, the goodness of fit statistics obtained when GRM was fitted to two subscales using probit and logit links were close to each other. Therefore, in both cases, it could be concluded that GRM was fitted to both subscales similarly. Then GRM (written as "one-factorial factor model for ordered categorical responses") was fitted by

using robust weighted least square estimation (and with probit link). DTGRM and GRDDM were fitted to each of the subscales, and goodness of fit statistics for these models are displayed in Table 2. However, when comparing model fits, it should be kept in mind that they were fitted using different estimation methods.

TABLE 1
Goodness of fit statistics for the normal ogive version of GRM and logistic GRM

Scale	Model	Log-likelihood	AIC	BIC	Sa-BIC	Free parameter
STAI-S	GRM (logistic)	–22540.563	45241.126	45641.517	45387.417	80
	GRM (normal ogive)	–22593.129	45346.258	45746.649	45492.550	80
STAI-T	GRM (logistic)	–23107.694	46375.389	46775.779	46521.680	80
	GRM (normal ogive)	–23197.163	46554.326	46954.717	46700.618	80

Note. GRM = Graded Response Model. AIC = Akaike information criterion; BIC = Bayesian informatin criterion; Sa-BIC = sample size adjusted BIC; STAI = State-Trait Anxiety Inventory; STAI-S = STAI-State; STAI-T = STAI-Trait.

TABLE 2
Goodness of fit statistics for GRM, GRDDM, and DTGRM for STAI measures

Scale	Model	RMSEA	CFI	TLI	$\chi^2(df)$
STAI-S	GRM (RWLSE)	.095	.964	.960	1848.569 (170)
	GRDDM	.042	.993	.992	498.691 (169)
	DTGRM	.104	–	.920	1368.144 (170)
STAI-T	GRM (RWLSE)	.054	.983	.981	721.734 (170)
	GRDDM	.052	.984	.982	678.720 (169)
	DTGRM	.069	–	.933	481.2487 (170)
LRT results					
	average Λ_i	Q(*df*)			
STAI-S	0.756	3384.388 (1102)			
STAI-T	0.765	3400.644 (1102)			

Note. GRM = Graded Response Model; GRDDM = Graded Response Differential Discrimination Model; DTGRM = Dual Thurstonian Graded Response Model. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker-Lrwis index; STAI = State-Trait Anxiety Inventory; STAI-S = STAI-State; STAI-T = STAI-Trait; RWLSE = robust weighted least square estimation method; LRT = likelihood ratio test.

As it is clear from Table 2, although all these statistics indicate an acceptable model fit, the results favor GRDDM for both subscales. These findings suggested that, in modeling responses to STAI-S and STAI-T items, it may be more beneficial to consider the additional person parameter indicating

person discrimination. In addition, fitting GRDDM resulted in a variance of $\sigma^2 = .850$ with a standard error of .076 for the STAI-S subscale, and a variance of $\sigma^2 = .712$ with a standard error of .103 for the STAI-T subscale. Given that in this study, values of σ_i are considered as the individual differences in response scale usage (instead of as an indicator of ERS), these variances suggest that, when responding to STAI-T items, compared to responding to STAI-S items, participants use a more extensive range of the response scale.

According to LR test results, approximate chi-square statistic (Q statistic) indicated that DTGRM was more appropriate than GRM with constant PDD in modeling responses given to items in both subscales. However, statistic Λ_i showed that the variability in the PDD estimates of respondents was not large. Also, this variation was similar for the two subscales. When Q statistic and statistic Λ_i were assessed together, it may be concluded that DTGRM was more appropriate for both subscales. Also, although both GRDDM and DTGRM contain person discrimination parameters, GRDDM shows a better fit than DTGRM. This situation seems to result from the fact that, while equal item discrimination is assumed in DTGRM, it is allowed to change discrimination across items in GRDDM. In parallel with this finding, GRM (normal ogive version), in which item discrimination was allowed to vary between items, showed a better fit to item response data obtained from both subscales than GRM (normal ogive version) with constrained item discrimination parameter (LR test results for STAI-S is 754.672 with $df = 19$, $p < .000$; LR test results for STAI-T is 280.018 with $df = 19$, $p < .000$).

Based on DTGRM, parameters' estimates and person-fit statistics were obtained with the EAP estimation method, and "marginal reliability" estimates for person parameters (based on PSD) were also obtained. The marginal reliability estimates for person parameters obtained based on responses given to STAI-S and STAI-T items are very close to each other. As Ferrando (2019) stated, although reliability values for σ^2 (PDD) estimates of respondents (ρ-PSD = .851 and ρ-PSD = .824 for STAI-S and STAI-T, respectively) are lower than reliability values for person locations (ρ-PSD = .951 and ρ-PSD = .947 for STAI-S and STAI-T, respectively), these values are also acceptable.

Afterward, the person parameters estimated under each model were examined. As suggested in the literature (LaHuis et al., 2018), the models were compared at the person level. The correlations between the person parameters, estimated under each model, and person-fit statistics were calculated separately for each subscale and presented in Table 3. It shows strong relationships between the latent trait estimates (θ) obtained under different models and that these relationships are similar for the two subscales. Weak negative correlations were found between PDD estimates based on DTGRM and alpha (α_i) estimates based on GRDDM. This is an expected finding because, while σ_i values are regarded as a "person-specific scale parameter" in DTGRM, α_i values are modeled as a "person slope parameter" in GRDDM. Accordingly, while low σ_i^2 indicates response consistency and high person reliability, low α_i indicates that respondents do not differentiate between items and select response categories near the scale midpoint. While negative relationships that could be considered strong were observed between l_z^p and PDD estimates, moderate positive relationships were observed with PRPs.

To assess the effectiveness of person parameters estimated in this study in measuring traitedness, the relationships between parameters' estimates and person-fit statistics based on STAI-S measures and their corresponding estimates based on STAI-T measures were examined, and the results are presented in Table 4.

TABLE 3
Correlations among person parameters estimated under Item Response Theory models tested in the study

Model	GRM-θ (Logistic)	GRM-θ (NO)	GRM-θ (RWLSE)	DTGRM-θ	GRDDM τ_i	GRDDM α_i	PDD	PRP	l_z^p
GRM-θ (Logistic)	1.000	.999**	.999**	.999**	.970**	.347**	–.034	.000	–.017
GRM-θ (NO)	.999**	1.000	1.000**	.996**	.974**	.339**	–.032	–.011	–.021
GRM-θ (RWLSE)	.999**	1.000**	1.000	.997**	.973**	.345**	–.032	–.012	–.018
DTGRM-θ	1.000**	.998**	.998**	1.000	.964**	.366**	–0.042	–.002	–.002
GRDDM-τ_i	.977**	.978**	.978**	.978**	1.000	.123**	–.011	–0.038	–.099**
GRDDM-α_i	.520**	.524**	.522**	.515**	.334**	1.000	–.079**	.052	.322**
PDD	–.129**	–.129**	–.130**	–.122**	–.136**	–.036	1.000	–.501**	–.795**
PRP	–.249**	–.254**	–.247**	–.248**	–.225**	–.194**	–.665**	1.000	.366**
l_z^p	.068*	.071*	.071*	.062*	.066*	.057	–.884**	.594**	1.000

Note. Above the diagonal, correlations for STAI-S measures are presented. Below the diagonal, correlations for STAI-T measures are presented. STAI = State-Trait Anxiety Inventory; STAI-S = STAI-State; STAI-T = STAI-Trait. GRM = Graded Response Model; GRM-θ (Logistic) = latent trait estimation under logistic GRM; GRM-θ (NO) = latent trait estimation under the normal ogive version of GRM; GRM-θ (RWLSE) = latent trait estimation under GRM fitted by using robust weighted least square estimation method. DTGRM = Dual Thurstonian Graded Response Model. GRDDM = Graded Response Differential Discrimination Model; GRDDM-τi = responder-specific discrimination parameter based on GRDDM; GRDDM-α_i = latent trait estimation based on GRDDM. PDD = person discriminal dispersion parameter; PRP = person reliability parameter; l_z^p = person-fit statistic.
*p < .05; **p < .01.

TABLE 4
Correlations between parameters' estimates and person-fit statistics obtained from STAI-S and STAI-T subscales

Model	GRM-θ (logistic, STAI-T)	GRM-θ (STAI-T; RWLSE)	DTGRM-θ (STAI-T)	PDD (STAI-T)	PRP (STAI-T)	GRDDM-τ_i (STAI-T)	GRDDM-α_i (STAI-T)	l_z^p (STAI-T)
GRM-θ (logistic, STAI-S)	.869**	.871**	.870**	–.077*	–.217**	.855**	.446	–.003
GRM-θ (STAI-S; RWLSE)	.875**	.877**	.875**	–.073*	–.225**	.860**	.454**	–.001
DTGRM-θ (STAI-S)	.869**	.871**	.869**	–.082**	–.216**	.854**	.447**	.011
PDD (STAI-S)	–.066*	–.062*	–.062*	.341**	–.202**	–.067*	–.004	–.340**
PRP (STAI-S)	–.042	–.047	–.037	–.162**	.482**	–.021	–.120**	.142
GRDDM-τ_i (STAI-S)	.853**	.857**	.854**	–.055	–.224**	.847**	.404**	–.029
GRDDM-α_i (STAI-S)	.301**	.303**	.297**	–.093**	–.067*	.256**	.332**	.114**
l_z^p (STAI-S)	.019	.018	.015	–276**	.137**	.010	.042	.350**
GRM-θ (logistic, STAI-S)	.869**	.871**	.870**	–.077*	–.217**	.855**	.446	–.003

Note. STAI = State-Trait Anxiety Inventory; STAI-S = STAI-State; STAI-T = STAI-Trait. GRM = Graded Response Model; GRM-θ (Logistic) = latent trait estimation under logistic GRM; GRM-θ (RWLSE) = latent trait estimation under GRM fitted by using robust weighted least square estimation method. DTGRM = Dual Thurstonian Graded Response Model. PDD = person discriminal dispersion parameter. PRP = person reliability parameter. GRDDM = Graded Response Differential Discrimination Model; GRDDM-τ_I = responder-specific discrimination parameter based on GRDDM; GRDDM-α_i = latent trait estimation based on GRDDM. l_z^p = person-fit statistic.
*p < .05; **p < .01.

Strong relations (varied between .869 and .877) were reported between latent trait estimates obtained under all models based on STAI-S and STAI-T items. It means that the latent trait estimates based on all models could effectively reflect traitedness. However, moderate and weak relations were observed for PDDs, responder-specific discrimination parameters (α_i), PRPs, and person-fit statistics. Because the strongest relation was observed for PRP, it could be inferred that PRP is the person parameter that reflects traitedness the most. This parameter is measured based on PDD estimates. Therefore, both parameters are expected to be similarly effective in assessing traitedness and could lead to the same inference in terms of traitedness. However, the correlation calculated for PRP estimates is higher than that calculated for PDD estimates. This could result from the fact that the reliability coefficients for participants' latent trait estimates are higher than the reliability coefficients for their σ^2 (PDD) estimates. The fact that correlations calculated for PDD, responder-specific discrimination parameters, and person fit statistics are very close indicates that they are similarly effective in measuring traitedness. Because both subscales measure anxiety and are administered successively, the respondents are not expected to differ much in terms of the response scale usage between the two subscales. In addition, considering that there is a strong positive relation between latent trait estimates based on two subscales under this model, it may be concluded that it could not reflect traitedness and the inferences about traitedness levels of the individuals based on the "responder-specific discrimination parameter" would not be appropriate.

DISCUSSION

Individual differences in terms of traitedness could affect the validity of affective measures and the usability of these measures for several purposes. Accordingly, researchers in the fields of education and psychology have been interested in measuring traitedness and the IRT approach has been used in the studies (LaHuis et al., 2017, 2018; Reise & Waller, 1993; Warner, 2005) aiming to assess traitedness, especially in the context of response consistency. Some studies have examined to what extent the IRT-based person-fit statistics (Reise & Waller, 1993; Warner, 2005) and person reliability parameter (LaHuis et al., 2017, 2018) may reflect traitedness. However, these studies are few, and further studies on this issue are needed. Unlike the previous studies, this one examined comparatively to what extent person-fit statistics and person parameters estimated based on DTGRM and GRDDM may reflect traitedness.

Consistent with the results of other works (Ferrando, 2019; Lubbe & Schuster, 2017), the current study found that DTGRM and GRDDM, which include the person parameter indicating "person discrimination," were more appropriate than the constant-θ IRT model. The results show that there are individual differences in the level of convergence between the participants' responses and their trait levels and in terms of the reliability of their responses. The results also indicate that more information can be obtained regarding both individuals' response behaviors and anxiety levels by using these models.

In assessing traitedness, the strongest relationships for both subscales were found between PDDs and person-fit statistics (l_z^p). In the literature, to the best of the author's knowledge, no research has examined the relations between PDDs and person-fit statistics. However, considering that PRPs are measured based on PDD estimates, this finding is consistent with the literature in this context. Nevertheless, in the current study, lower relations between person reliability and person-fit statistics were found compared to the studies conducted by Ferrando (2004, 2014a) and LaHuis et al. (2017). This may be because the relationships between these parameters were examined according to an IRT model based on Thurstone scaling for binary data in Ferrando's (2004) study and based on Vθ-GRM in other studies (Ferrando, 2014a; LaHuis et al., 2017). Differences in modeling individual discrimination in Vθ-GRM and DTGRM may have led to these differences.

Correlations between person parameters and person fit statistics estimated based on responses to STAI-S items and person parameters and person-fit statistics estimated based on responses to STAI-T items indicated that PRP reflects traitedness the most. These findings are consistent with the literature (LaHuis et al., 2017; Lubbe & Schuster, 2017). La Huis et al. (2017) stated that if PRPs reflected traitedness, PRPs should correlate for similar traits. Researchers found that the cross-scale correlations for PRP and person-fit statistics were low or medium and that PRPs might reflect traitedness. However, the relations observed for PRPs were stronger than those observed for person-fit statistics. Similarly, the current study found higher correlations between PRP estimates based on STAI-S and STAI-T subscales than the correlation found for person-fit statistics. Although not interested in traitedness, Lubbe & Schuster (2017) fitted GRDDM to item response data obtained from five NEO-PI-R scales and estimated α values. Researchers examined relationships among extracted factor scores of these α values and found low to medium relations between them, as also noted in the current study.

Another finding of the study points to the importance of the nature of the trait in assessing traitedness. The correlation calculated for the relationship between PRPs and person-fit statistics based on STAI-S items was lower than that calculated for those based on STAI-T items. The trait measured with STAI-S items is momentary and depends more on conditions than trait anxiety. However, trait anxiety is relatively more stable. The responding behavior is mainly determined by the anxiety of the participants. However, it is more likely that there may be other factors affecting item responses and that the probability of these factors influencing the responses between items may differ more than in the case of responding to STAI-T items. Supporting this, the results of confirmatory factor analysis performed to test the unidimensionality assumption, the goodness of fit statistics for STAI-T measures pointed to a better fit. Therefore, confounding factors may have affected the response consistency when responding to items measuring state anxiety, which is a less stable trait. In this context, the findings of this study imply that the extent to which the person parameters and person-fit statistics reflect traitedness may differ according to the trait being measured. Supporting this idea, Ferrando (2014a) stated that while IRT models in which PRPs are estimated are more suitable for some traits, constant-θ models may be more suitable for some others.

In summary, PDDs and person-fit statistics can reflect traitedness, but the person reliability parameter reflected traitedness the most. It is an expected finding because person reliability is measured based on PDD, and PDD directly indicates "trait variability." Accordingly, individuals with low trait variability would be more sensitive to different item locations and consistently respond to items. This means that these individuals have a well-defined trait level. So, based on the findings of this study and the literature (LaHuis et al., 2017), it could be suggested that the researchers and practitioners in the field of educational measurement and psychology use PRPs to define low-traited individuals in terms of traits they are interested in. However, as noted in the literature (Ferrando, 2019; LaHuis et al., 2018), at least 20 items are required to estimate PDDs and PRPs. Accordingly, in assessing traitedness based on responses given to items included in scales with more than 20 items, the estimates of PDD and PRP may be profitably used by fitting DTGRM. In assessing traitedness, another issue about using PRPs is the need to provide evidence that DTGRM is more appropriate than the corresponding standard model in modeling item responses.

Implications and Suggestions

Individual differences in terms of traitedness could substantially affect the validity of measures of people's affective attributes and the appropriateness and accuracy of the inferences and decisions made based

on these measures. For low-traited individuals, these measures may not be usable and beneficial in predicting their behaviors. In this context, sufficient evidence may not be obtained for these individuals in terms of the validity of these measures. The fact that the respondent does not have a strong internal representation of the related trait could lead to an inaccurate estimation of his/her latent trait level and to an individual not being scaled on the trait continuum. Therefore, assessing traitedness and diagnosing low-traited individuals can enhance the validity and reliability of the affective measures (Ferrando, 2009; LaHuis et al., 2018). To obtain more accurate and valid estimates of the relevant trait for low-traited individuals, these individuals may be assessed using different scales. Necessary adjustments can be made by examining whether the individual differences in traitedness are related to several aspects of the measurement instruments. In this context, the clarity of the instruction can be reviewed and rewritten. Some concepts or statements might cause respondents to give "inappropriate responses to their trait levels" in the scale items. Based on this examination, changes can be made in the statements of the items in accordance with the levels of respondents in terms of personality development.

Assessing to what extent the person parameters and person-fit statistics can reflect traitedness based only on correlations can be considered a limitation of the current study. For future research, the profile membership of responders can be identified in terms of the trait being measured. Traitedness can be assessed by examining possible profile differences in terms of parameters' estimates and person-fit statistics. In this way, information about the source of individual differences in terms of traitedness can be obtained. When examining traitedness based on response consistency, it should be taken into account that although participants' anxiety (measured trait) mainly determines the responses, there are other factors such as response bias that may affect the consistency of the responses. While individual differences in the estimates of the responder-specific discrimination parameter indicate differences in the reliability of responses, they may also reflect "extreme response style" or "careless responding." Similarly, person-fit statistics could point to the other sources of model misfit (e.g., careless responding) besides person reliability (Emons, 2009). Thus, the involvement of response bias may have affected response consistency and thus the level of reflection of traitedness by PDD and person-fit statistics. Accordingly, future research may choose to examine response bias by using some procedures, for example, Mixture Graded Response Model (Eid & Rauber, 2000) or Multi-Process IRT models (Böckenholt & Meiser, 2017), before assessing traitedness based on the IRT approach.

References

Allport, G. W. (1937). *Personality: A psychological interpretation*. Holt.

Baumeister, R. F., & Tice, D. M. (1988). Metatraits. *Journal of Personality*, *56*(3), 571-598. https://doi.org/10.1111/j.1467-6494.1988.tb00903.x

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*(6), 506-520. https://doi.org/10.1037/h0037130

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159-181. https://doi.org/10.1111/bmsp.12086

Conijn J. M., Emons W. H. M., Van Assen M. A. L. M., Pedersen S. S., Sijtsma K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger State-Trait Anxiety Inventory. *Multivariate Behavioral Research*, *48*, 692-718. https://doi.org/10.1080/00273171.2013.815580

Cucina, J. M., & Vasilopoulos, N. L. (2005). Nonlinear personality-performance relationships and the spurious moderating effects of traitedness. *Journal of Personality*, *73*(1), 227-259. https://doi.org/10.1111/j.1467-6494.2004.00309.x

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20-30. https://doi.org/10.1027/1015-5759.16.1.20
Emons, W. H. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*(8), 599-619. https://doi.org/10.1177/0146621609334378
Ferrando, P. J. (2004). Person reliability in personality measurement: An Item Response Theory analysis. *Applied Psychological Measurement*, *28*(2), 126-140. https://doi.org/10.1177/0146621603260917
Ferrando, P. J. (2007). A Pearson-type-VII item response model for assessing person fluctuation. *Psychometrika*, *72*(1), 25-41. https://doi.org/10.1007/s11336-004-1170-0
Ferrando, P. J. (2009). A grade response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 641-662. https://doi.org/10.1348/000711008X377745
Ferrando, P. J. (2013). A general linear framework for modeling continuous responses with error in persons and items. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(4), 150-161. https://doi.org/10.1177/0146621613497532
Ferrando, P. J. (2014a). A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement, 38*(2),166-183. https://doi.org/10.1177/0146621613497532
Ferrando, P. J. (2014b). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behavioral Research*, *49*(4), 390-405, https://doi.org/10.1080/00273171.2014.911074
Ferrando, P. J. (2016). An IRT modeling approach for assessing item and person discrimination in binary personality responses. *Applied Psychological Measurement*, *40*(3), 218-232. https://doi.org/10.1177/0146621615622633
Ferrando, P. J. (2019). A comprehensive IRT approach for modeling binary, graded, and continuous responses with error in persons and items. *Applied Psychological Measurement*, *43*(5), 339-359. https://doi.org/10.1177/0146621618817779
Fiske, D. W. (1968). Items and persons: Formal duals and psychological differences. *Multivariate Behavioral Research, 3,* 393-401. https://doi.org/10.1207/ s15327906mbr0304_2
International Test Commission. (2012). *International guidelines on quality control in scoring, test analysis, and reporting of test scores.* Retrieved from www.intestcom.org.
LaHuis, D. M., Barnes, T., Hakoyama, S., Blackmore, C., & Hartman, M. J. (2017). Measuring traitedness with person reliabilities parameters. *Personality and Individual Differences*, *109,* 111-116. https://doi.org/10.1016/j.paid.2016.12.034
LaHuis, D., Bryant-Lees, K., Hakoyama, S., Barnes, T., & Wiemann, A. (2018). A comparison of procedures for estimating person reliability parameters in the graded response model. *Journal of Educational Measurement*, *55*, 421-432. https://doi.org/10.1111/jedm.12186
Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 109-131). Academic Press.
Lubbe, D., & Schuster, C. (2017). The graded response differential discrimination model accounting for extreme response style. *Multivariate Behavioral Research*, *52*, 616-629. https://doi.org/10.1080/00273171.2017.1350561
Lubbe, D., & Schuster, C. (2020). A scaled threshold model for measuring extreme response style. *Journal of Educational and Behavioral Statistics*, *45*(1), 86-107. https://doi.org/10.3102/1076998619859541
Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Muthén & Muthén. (Original work published 1998).
Navarro-Gonzalez, D., & Ferrando, J. P. (2019). *InDisc: Obtaining and estimating unidimensional and multidimensional IRT dual models.* R package manual. https://cran.r-project.org/package=InDisc
Oei, T. P., Evans, L., & Crook, G. M. (1990). Utility and validity of the STAI with anxiety disorder patients. *British Journal of Clinical Psychology*, *29*(4), 429-432. https://doi.org/10.1111/j.2044-8260.1990.tb00906.x
Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory models*. Sage Publications.
Öner, N. (1997). *Türkiye'de Kullanılan Psikolojik Testler* [Psychological tests used in Turkey]. Boğaziçi Üniversitesi Matbaası.
Öner, N., & Le Compte, A. (1985). *Durumluk-Sürekli Kaygı Envanteri el kitabı* [State-Trait Anxiety Inventory manual]. Boğaziçi Üniversitesi Matbaası.
Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*(1), 143-151. https://doi.org/10.1037/0022-3514.65.1.143
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(Suppl. 1), 1-97. https://doi.org/10.1007/BF03372160
Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y).* Consulting Psychologists Press.
Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 622-663. https://doi.org/10.1111/j.1467-6494.1988.tb00905.x

Warner, M. B. (2005). Personality traits, traitedness, and disorders: Towards an enhanced understanding of trait -disorder relationships. Retrieved from https://www.proquest.com/dissertations-theses/personality-traits-traitedness-disorders-towards/docview/305375848/se-2?accountid=11054

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x.