2026, Volume 16, Issue 1, Article Number 18
DOI: 10.48146/odusobiad.1810609
Received Date: 25.10.2025
Acceptance Date: 05.02.2025
Online Published Date: 26.03.2026

Research Article *Open Access*

Structural and behavioral implications of wording effects: A comprehensive examination via exploratory graph analysis and explanatory item response models[1]

Sinem Demirkol[2]

[2] Ordu University, Faculty Of Education, Ordu, Türkiye

ABSTRACT

The inclusion of positively and negatively worded items in self-report scales can generate additional variance unrelated to the target construct, giving rise to wording effects (WE). The aim of this study is to examine the impact of WE on both the scale structure and respondents' response behavior. The study sample comprised 7,245 students who participated in the PISA 2022 Türkiye. The Grit and Self-Control scales examined in this study comprise 10 items (6 positively worded, 4 negatively worded) and are administered in a five-point Likert format. Data analyses were conducted in two stages. First, the dimensionality of the scales was investigated using the Exploratory Graph Analysis method, an extension of psychometric network analysis. Next, the influence of WE on response behavior was examined using Explanatory Item Response Models (EIRM). The findings indicate that, although these scales were theoretically designed as unidimensional, positively and negatively worded items form distinct dimensions, and positively worded items show higher probabilities of being assigned to higher response categories. Overall, the results suggest that ignoring WE may undermine measurement validity and introduce bias, particularly in comparative studies. Therefore, WE and method effects should be systematically incorporated into modeling and reporting practices. Future research is encouraged to address WE systematically in scale structures, and to collect process data such as response times and rereading indicators.

KEYWORDS

Wording effect, exploratory graph analysis, explanatory item response model, PISA 2022.

İfade etkisinin yapısal ve davranışsal yansımaları: Keşifsel grafik analizi ve açıklayıcı madde tepki modelleriyle bütüncül bir inceleme

ÖZET

Öz-bildirim temelli ölçeklerde olumlu ve olumsuz maddelerin birlikte yer alması, hedef yapıyla ilgisiz ek varyans üreterek ifade etkisi (İE) doğurabilir. Bu çalışmanın amacı, İE'nin hem ölçek yapısı hem de yanıtlayıcı davranışları üzerindeki etkisini incelemektir. Araştırmanın çalışma grubunu PISA 2022 Türkiye uygulamasına katılan 7245 öğrenci oluşturmaktadır. Çalışmada ele alınan Azim ve Öz-kontrol ölçekleri 10 maddeden (6 olumlu, 4 olumsuz) oluşmakta ve beşli Likert formatında uygulanmaktadır. Verilerin analizi iki aşamada yürütülmüştür. Öncelikle psikometrik ağ analizinin bir uzantısı olan keşifsel grafik analiziyle ölçeklerin boyutlanması incelenmiş, ardından İE'nin yanıt davranışı üzerindeki etkisi Açıklayıcı Madde Tepki Modelleri ile analiz edilmiştir. Bulgular, kuramsal olarak tek boyutlu tasarlanan bu ölçeklerde olumlu ve olumsuz maddelerin ayrı boyutlar oluşturduğunu ve olumlu maddelerin onaylanma olasılıklarının daha yüksek olduğunu göstermektedir. Sonuçlar, İE'nin göz ardı edilmesinin ölçüm geçerliğini zayıflatıp özellikle karşılaştırmalı çalışmalarda yanlılık üretebileceğini, bu nedenle İE ve yöntem

[1] Parts of this study were presented as an oral presentation at the 13th International Symposium on Social Studies Education (USBES), Tokat, Türkiye, in 2025.

Cite: Demirkol, S. (2026). Structural and behavioral implications of wording effects: A comprehensive examination via exploratory graph analysis and explanatory item response models. *Ordu University Institute of Social Sciences Journal of Social Sciences Research,* 16(1), Article Number 18. https://doi.org/ 10.48146/odusobiad.1810609

Corresponding Author: dmrklsinem@gmail.com

etkilerinin modelleme ile raporlama aşamalarına sistematik biçimde dâhil edilmesi gerektiğini göstermektedir. Gelecek çalışmalar için, ölçek yapılarında İE'nin sistematik olarak ele alınması, yanıt süresi/yeniden okuma gibi süreç verilerinin toplanması ve madde yönünün incelenmesi önerilir.

ANAHTAR KELİMELER

İfade etkisi, keşifsel grafik analizi, açıklayıcı madde tepki modeli, PISA 2022.

Introduction

Self-report scales are widely used measurement tools in education and psychology. They allow respondents to express their views on a particular phenomenon, situation, or statement and provide response options that represent varying levels of agreement or approval. In this way, individuals contribute to the measurement of the relevant attitude, emotion, or trait by selecting the option that best reflects them (Likert, 1932). Despite their strengths, these instruments also have certain limitations. One such limitation is the presence of method effects. Method effects refer to the portion of observed variance in measurements that can be attributed to the measurement method rather than to the constructs intended to be represented (Williams et al., 1989). In other words, method effects reflect unwanted sources of variance that arise from the instrument, format, context, or respondents' tendencies, beyond the construct of interest (e.g., self-esteem, motivation, anxiety). Although method effects can stem from many sources, one of the most well-known forms is the wording effect. Wording effects (WE) can be defined as the systematic influence of positively versus negatively worded items on responses, producing additional variance unrelated to the intended construct (Podsakoff et al., 2003).

This study aims to examine wording effects (WE) on the psychometric structure and response patterns of the Grit and Self-Control scales included in the PISA 2022 student questionnaire, a large-scale international assessment. The research was conducted in two stages. In the first stage, the dimensional structure of the scales was examined using Exploratory Graph Analysis (EGA), an extension of psychometric network analysis. In the second stage, item parameters were estimated simultaneously within the framework of Explanatory Item Response Models (EIRM). This approach enabled an evaluation of the unique contribution of WE to response probabilities and item-level characteristics.

Method effects

The term method refers to how measurement is carried out—across different levels of abstraction such as item content, scale type, response format, and the administration context (Fiske, 1982). Each of these dimensions may give rise to specific types of systematic bias in measurement. Podsakoff et al. (2003) classified method-related influences into four broad categories: common rater effects, item context effects, measurement context effects, and effects stemming from item characteristics. Common rater effects are methodological distortions that occur when both predictor and criterion variables are reported by the same person, thereby inflating associations between variables due to shared source variance. Subcomponents of this category include consistency motives, social desirability, and transient mood states. Item context effects arise from an item's position within a scale or its relationship with surrounding items. Examples include earlier questions influencing later responses, initial items inducing a particular mood, scale length, and the inclusion of items measuring different constructs within the same section. Measurement context effects originate from the setting and conditions under which measurement occurs. Data collected within the same context may generate spurious associations independent of substantive content. Examples include measuring predictor and criterion variables at the same time, collecting them in the same location, or obtaining them through the same mode (e.g., online-only surveys). Finally, effects due to item characteristics arise from how items are written, the scale format used, or linguistic properties. In other words, respondents may provide systematic answers influenced by the

formal properties of the item rather than their true attitudes. Examples include ambiguous statements, items implying socially desirable responses, repeated use of the same scale format (e.g., Likert) or the same endpoint labels (e.g., "never–always"), and the inclusion of both positively and negatively worded items within the same scale (Podsakoff et al., 2003).

One of the most comprehensive syntheses on method effects was conducted by Cote and Buckley (1987), who reviewed 70 studies from psychology, sociology, marketing, business, and education. Their findings suggested that, in a typical measurement, approximately 26.3% of the variance may be attributable to systematic measurement error such as common method variance. They further reported that this proportion varies substantially across disciplines and by the type of construct measured. Specifically, method-related variance was reported to be around 15% in marketing, rising to 30% in education, 22% in job performance measures, and up to 40% in attitude measures.

Evidence also indicates that method effects can influence relationships between measures. When common method variance is controlled, the average explained variance between two variables has been reported to be about 11%, whereas it is about 35% when it is not controlled (Fuller et al., 1996; Lowe et al., 1996; Podsakoff et al., 2000; Wagner & Gooding, 1987). In addition, Podsakoff et al. (2003) showed that if measures contain common method variance, the observed relationship between predictor and criterion variables may be underestimated by approximately 25%. Taken together, this evidence suggests that method effects may inflate associations in some cases and suppress them in others, thereby increasing the likelihood of both Type I and Type II errors (Campbell & Fiske, 1959; Cote & Buckley, 1987; Podsakoff et al., 2003). Consequently, identifying method effects and modeling them appropriately supports more valid inferences and strengthens the validity of the resulting conclusions.

Wording effects

Likert scales are a widely used and recommended approach for collecting data on attitudes, beliefs, values, and other latent constructs (Peterson, 1994). Nevertheless, this robust framework entails certain methodological risks stemming from respondent behavior. For example, acquiescence bias refers to a tendency to select positively worded items rather than negatively worded ones (Watson, 1992). Similarly, the systematic tendency to prefer extreme categories on rating scales—or to avoid them—regardless of item content is referred to as extreme response style, a potential source of bias in rating data (Greenleaf, 1992). Such systematic response tendencies may lead to over- or underestimation of the true level of the latent construct intended to be measured (Winkler et al., 1982).

A common strategy to reduce systematic response biases in scales is to sample the same construct using both positively and negatively worded items. This approach aims to produce more neutral responses and improve measurement validity by balancing tendencies associated with one-sided wording (DeVellis, 2005). In this context, the desire to eliminate response bias in the scale-development literature (Nunnally, 1967) has supported the inclusion of positive–negative item pairs within the same instrument, with the goal of mitigating the influence of acquiescence (Hinz et al., 2007). The critical assumption underlying this strategy is that positively and negatively worded items measure the same underlying construct (Marsh, 1996). In line with this assumption, responses to negatively worded items are reverse-coded and treated identically to positively worded items (Horan et al., 2003).

Dalal and Carter (2015) conceptualize negatively worded items in two types and emphasize that this distinction may have different implications for the measurement process. The first type comprises oppositely keyed items. Without using linguistic negation, these items target opposite poles of the same construct and consist of statements that are semantically antonymous (e.g., "I study systematically" vs. "I study haphazardly," "I am open to new ideas" vs. "I am closed to new ideas"). The second type is negated normal items, in which a positively worded statement is turned into its direct opposite through logical negation markers such as

"not" (e.g., "I use my time effectively" vs. "I do not use my time effectively," "I arrive at appointments on time" vs. "I do not arrive at appointments on time"). In the literature, the distinction between these two types of items (fully opposite vs. negated normal items) is not always clear, and both are often treated broadly as negatively worded items. In practice, it is commonly assumed that positively and negatively worded items are functionally interchangeable (Lindwall et al., 2012). Under this assumption, a participant's "strongly agree" response to a positively worded item should exhibit mirror symmetry with a "strongly disagree" response to its logically corresponding negatively worded item (Marsh, 1996). However, empirical findings on this issue are inconsistent. Some studies have reported that, even for semantically equivalent item pairs, respondents show a relatively higher tendency to agree with negatively worded items (Holleman, 1999; Schriesheim & Hill, 1981), whereas other studies have found systematically higher mean scores for positively worded items (Weem et al., 2003).

Including negatively worded items in scales forces respondents to answer more carefully and thus encourages more controlled cognitive processing rather than automatic responding (Hinkin, 1995; Idaszak & Drasgow, 1987). Accordingly, negatively worded items can function like cognitive speed bumps, increasing the likelihood that a questionnaire is completed more thoughtfully (Podsakoff et al., 2003). Such a practice may help an instrument capture the target construct more accurately (Anderson & Gerbing, 1988), improve measurement efficiency (Worcester & Burns, 1975), and achieve broader coverage of the construct (Weijters & Baumgartner, 2012).

Research has shown that scales containing both positively and negatively worded items exhibit weaker inter-item correlations than scales composed solely of positively worded items (DiStefano & Motl, 2006) and reduced internal consistency (Lee et al., 2008). Findings regarding the direction of WE are also mixed. Some studies report that negatively worded items produce stronger WE than positively worded items (DiStefano & Motl, 2009; Marsh, 1996), whereas others indicate that the more pronounced effect emerges for positively worded items (Lindwall et al., 2012). Moreover, using positive and negative items together can alter the factor structure of a scale by allowing shared variance unrelated to the target construct to enter the measurement model (Corwyn, 2000; John et al., 2019), and it has been reported that negatively worded items often load on a separate factor (Benson & Hocevar, 1985; Pilotte & Gable, 1990). However, such artificial factors have been shown to disappear when the relevant items are rewritten in a positive form (Harvey et al., 1985; Idaszak & Drasgow, 1987). Collectively, these findings suggest that WE can undermine—and in some cases violate—the intended unidimensional structure of a scale (Cordery & Sevastos, 1993; Hevey et al., 2010).

Purpose of the study

In light of this background, the present study examines whether wording effects (WE) lead to differences in the factor structure of the Grit and Self-Control scales, which include both positively and negatively worded items, and to investigate the impact of WE on individuals' response behavior. To this end, the study addresses the following research questions:

1. Does WE influence the factor/dimensional structure of the Grit and Self-Control scales?
2. Does WE significantly alter respondents' response behavior?

Method

Participants

The participants comprised 7245 students who took part in the PISA 2022 Turkey assessment. Of the participants, 3561 (49.2%) were female and 3684 (50.8%) were male. Turkey participated in PISA 2022 through 196 schools representing 12 regions according to the Turkish Statistical Area Classification (TSAC) Level-1. Regarding school type distribution, regular Anatolian high

schools accounted for the largest share (56%), followed by vocational and technical Anatolian high schools (23%) (MEB, 2022).

Instruments

Data were collected using the Self-Control and Perseverance (Grit/Persistence) scales from the General Social and Emotional Characteristics module of the PISA 2022 student questionnaire. Both scales are grounded in the OECD's Survey on Social and Emotional Skills (SSES) framework and are positioned as subcomponents of conscientiousness within the Big Five personality model. The scales were administered and scored using a within-construct matrix sampling approach (OECD, 2024).

The Self-Control scale (ST309) measures students' ability to resist distractions, refrain from acting on impulses, and stay focused on long-term goals. The scale consists of 10 items (6 positively worded; 4 negatively worded). Students rated each item on a five-point Likert-type scale ranging from "Strongly disagree" to "Strongly agree." The items are provided in Appendix 1.

The Perseverance scale (ST307) assesses students' tendency to complete tasks they start, continue exerting effort when faced with difficulties, and avoid giving up easily. The scale also consists of 10 items (6 positively worded; 4 negatively worded), rated on the same five-point Likert-type scale from "Strongly disagree" to "Strongly agree." The items are provided in Appendix 1.

An item-level WE variable was created by the researchers. Negatively worded items were coded as 0 (reference) and positively worded items as 1. This variable was used as an item-level covariate in the EIRM framework to test whether WE is associated with item location (threshold) parameters. Under this coding scheme, a positive coefficient indicates that positively worded items tend to have higher location/threshold values than negatively worded items (i.e., they require higher latent trait levels to endorse at higher response categories).

Data analysis

Analyses were conducted in two stages. First, EGA was used to identify how items cluster into dimensions. Developed by Golino and Epskamp (2017), EGA provides an alternative to factor-analytic methods for determining dimensionality, drawing on network psychometrics. In this approach, nodes represent items and edges represent partial correlations between items. Edges are shown as green for positive associations and red for negative associations, with thickness proportional to association strength (thicker edges indicate stronger relations). Node colors indicate the dimension to which items belong.

Prior to EGA, data quality and the suitability of the association matrix were evaluated as prerequisites. Specifically, the level and pattern of missing data were examined, and potential issues that could distort correlation estimation—such as items with excessive missingness or severe category pile-ups—were checked. In the EGA procedure, partial correlations among items were first estimated. Next, the graphical LASSO (GLASSO) was applied to shrink small correlations toward zero, preventing spurious connections and yielding a more parsimonious and interpretable network. Finally, the Walktrap algorithm was used to detect clusters of variables, which correspond to dimensions in factor-analytic terms. This process provides both an estimate of the number of dimensions and a visualization of how items cluster (Christensen et al., 2020). EGA was conducted in R (R Core Team, 2025) using the EGAnet package (Golino & Christensen, 2022).

After establishing the dimensional structure, the second step examined the effect of WE on the probability of selecting higher response categories (analogous to item difficulty in achievement tests for Likert-type scales). For this purpose, polytomous EIRM suitable for ordinal item formats

were used, in which item and person parameters are estimated simultaneously (De Boeck & Wilson, 2004).

EIRM extends traditional Item Response Theory (IRT) by integrating measurement (item/person parameters) and explanation (how item and person characteristics influence responses) within a single model (De Boeck & Wilson, 2004). Statistically, EIRM can be expressed within the generalized linear mixed modeling (GLMM) framework: each person's response to each item is treated as a repeated observation, allowing item- and person-level effects to be estimated jointly within one model.

For Likert-type items, EIRM is implemented via polytomous IRT models (e.g., PCM, GRM, RSM). Threshold parameters are specified as functions of item- and/or person-level covariates, enabling an explanation of when and how category transitions shift (Stanke & Bulut, 2019). In this study, prior to fitting EIRM, the response data were converted to long format, with each row representing a person × item response. The explanatory component was then added by allowing thresholds (and/or item parameters) to vary as functions of item and person characteristics. This enables the simultaneous testing of item features (e.g., positive vs. negative wording), person characteristics (e.g., gender, home language), and—if needed—person × item interactions within the same model.

Polytomous EIRM analyses were conducted in R using the eirm package (Bulut, 2021). Because items are nested within persons, the original wide-format dataset (persons as rows, items as columns) was reshaped into long format to support multilevel modeling, such that each row represented a single person–item observation (items were hierarchically nested within persons). In addition, polytomous (Likert-type) responses were decomposed into multiple binary indicators using the polyreformat function in the eirm package. This transformation preserves the original response structure while enabling the estimation of binary response models within the GLMM framework.

Table 1 Transforming multi-category responses into binary responses

Original Answer	Disagree	Neutral	Agree	Strongly agree
Strongly disagree	0	NA	NA	NA
Disagree	1	0	NA	NA
Neutral	NA	1	0	NA
Agree	NA	NA	1	0
Strongly agree	NA	NA	NA	1

To test whether incorporating WE into the model is warranted within the explanatory IRT framework, hierarchical model comparisons were conducted. Specifically, model fit was compared between a baseline model (M_0) that did not include WE and an extended model (M_1) that included WE. Model evaluation was based on AIC, BIC, and log-likelihood (logLik) values. Because the models were nested, a Likelihood Ratio Test (LRT; χ^2, Δdf) was applied. Lower AIC/BIC values, a higher (i.e., less negative) logLik, and a statistically significant $\chi^2(\Delta df)$ were interpreted as evidence that adding WE improved model fit.

Findings

Figure 1 presents the EGA results for the Grit and Self-Control scales. Nodes represent items, and edges represent the relationships among items. Green edges indicate positive associations, whereas red edges indicate negative associations. In addition, thicker edges reflect stronger relationships. Red nodes correspond to positively worded items, and blue nodes correspond to negatively worded items.

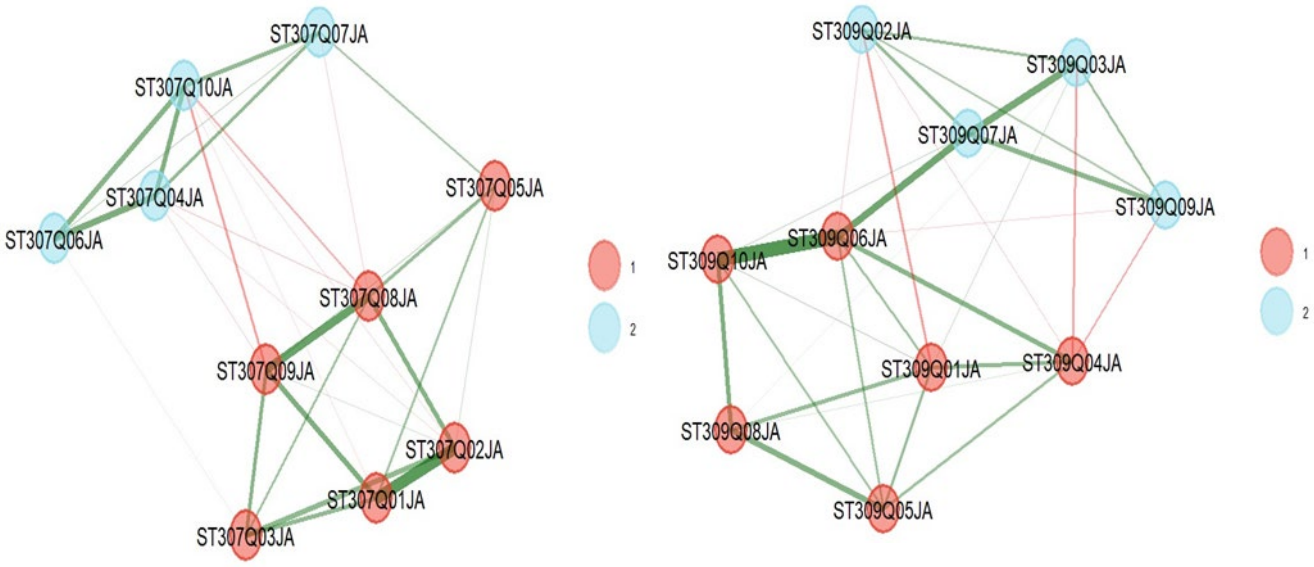


Figure 1 EGA results for the Perseverance (left) and Self-Control (right) scales.

As shown in Figure 1, a two-dimensional structure emerged (red and blue indicators on the right side of each network denote two dimensions), and these dimensions corresponding entirely to WE (positive vs. negative). Within-dimension connections are positive and strong, whereas between-dimension connections are mostly weak. This pattern suggests that WE is substantial in both scales and that items may cluster more by wording than by substantive content. In contrast to the standard PISA scaling approach—where student questionnaire scale scores are typically reported as IRT-based derived variables (e.g., Rasch/PCM-based scaling)—this finding indicates that the assumption of unidimensionality may be questionable for these scales (OECD, 2024).

In the second stage, EIRM models were fitted to examine whether the observed dimensionality was reflected in response behavior. For each scale, a baseline/null model (M0) without WE was specified first, followed by an extended model (M1) that included WE, to evaluate the incremental contribution of WE to model fit.

Table 2 Model fit values for the models

Scale	Models	n	logLik	AIC	BIC	Deviation	$\Delta\chi^2$ (Δdf)	p
Perseverance	M_0	5	−41170	82351	82396	82341	—	—
	M_1	6	−40397	80806	80860	80794	1546.2 (1)	.001
Self-Control	M_0	5	−40347	80704	80749	80694	—	—
	M_1	6	−40107	80226	80280	80214	479.50 (1)	.001

Note. M_0 = base model (no WE); M_1 = extended model (model including WE). Lower AIC/BIC and higher logLik indicate better fit; a significant LRT (χ^2) supports the superiority of M_1 over M_0.

Table 2 shows that including WE in the model significantly improves model fit. Consistent with this result, the fixed effects for models including WE are presented in Table 3.

Table 3 Fixed effects for WE in the perseverance and self-control scales

	Perseverance			Self-control		
Variable	Estimate(SE)	z	p	Estimate(SE)	z	p
Intercept	0.052(0.019)	2.68	.007 **	0.298 (0.024)	12.28	< .001 ***
Wording Effect	0.732(0.019)	38.73	< .001 ***	0.385 (0.018)	21.81	< .001 ***
polycategory_3	−0.294(0.025)	−11.62	< .001 ***	−0.040 (0.030)	−1.32	.187
polycategory_4	−0.277(0.025)	−10.93	< .001 ***	0.034 (0.028)	1.21	.225
polycategory_5	−1.222(0.027)	−45.11	< .001 ***	−1.071 (0.028)	−38.17	< .001 ***

*Note. SE = standard error. p values: *** < .001, ** < .01. The values are on the log-odds scale.*

Polytomous item threshold parameters represent the distance between adjacent response categories, with the reference threshold was defined as the transition from “Strongly disagree” to “Disagree.” Coefficients are expressed on the log-odds scale. A negative coefficient indicates that the odds of selecting the relevant category (relative to the reference) are lower. For the Perseverance scale, the odds of selecting Category 3 (the distance from “Disagree” to “Neither agree nor disagree”), Category 4 (the distance from “Neither agree nor disagree” to “Agree”), and Category 5 (the distance from “Agree” to “Strongly agree”) are lower compared to the reference threshold. The largest decrease occurs for Category 5. For the Self-Control scale, the pattern differs somewhat: relative to the reference threshold, changes in the odds of selecting Categories 3 and 4 are not statistically significant, whereas the odds of selecting Category 5 (from “Agree” to “Strongly agree”) decrease. In other words, relative to the reference threshold (from "Strongly disagree" to "Disagree"), the odds of selecting Categories 3–5 were lower for the Perseverance scale (strongest reduction for Category 5), whereas only Category 5 showed a statistically significant reduction for the Self-Control scale.

WE was statistically significant in both scales. In the Perseverance scale, responses to positively worded items were more likely to fall into higher response categories than responses to negatively worded items (β = 0.732, SE = 0.019, z = 38.73, p < .001, OR=2.08). The odds ratio (OR) indicates that the odds of endorsing a higher category were 2.08 times higher for positively worded items. For example, if the probability of selecting the higher category for a negatively worded item is 0.50, then the corresponding probability for a positively worded item would be approximately 0.68. Similarly, WE was significant for the Self-Control scale (β = 0.385, SE = 0.018, z = 21.81, p < .001). Applying the same transformation, if the probability of selecting the higher category for a negatively worded item is 0.50, then the corresponding probability for a positively worded item would be approximately 0.59.

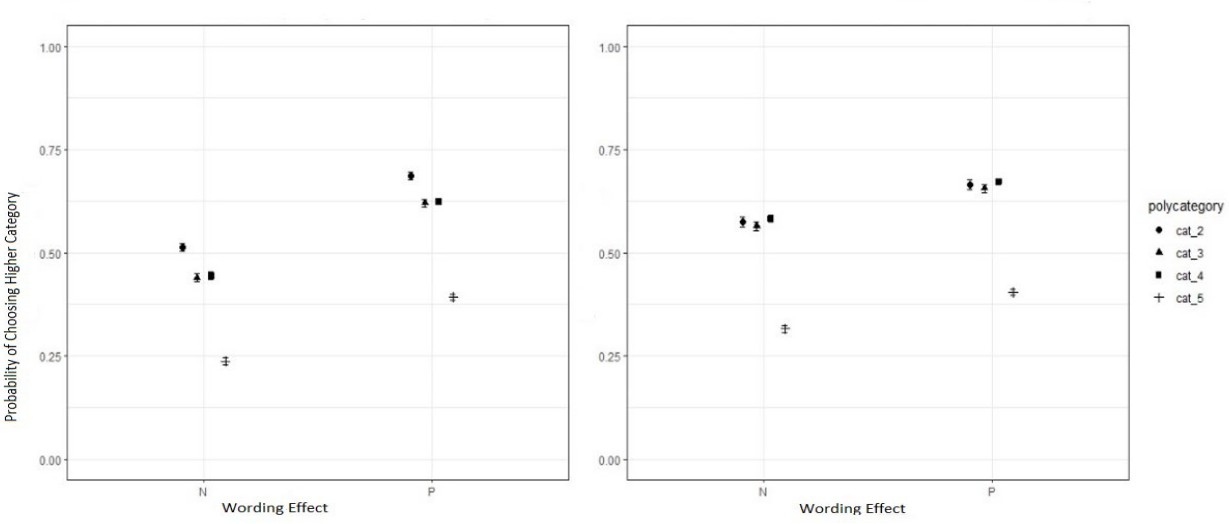


Figure 2 Graphs of wording effects for the Perseverance and Self-Control Scales.

In summary, positively worded items exhibited higher endorsement probabilities than negatively worded items across both scales.

Conclusion and recommendations

The purpose of this study was to examine the impact of WE on both scale structure and respondents’ response behavior. To this end, the dimensionality of the Perseverance and Self-Control scales—each containing both positively and negatively worded items—was first investigated using EGA, an extension of psychometric network analysis. Next, whether this

dimensional pattern was reflected in response behavior was tested using EIRM. In this way, the study provided an integrated evaluation of WE by addressing not only its structural implications but also its behavioral manifestations in responding.

Overall, the findings indicate that positively and negatively worded items in the Perseverance and Self-Control scales form distinct dimensions. The EGA results showed that, in both scales, positive and negative items clustered into separate communities. The EIRM results, which examined whether this structural split translated into response behavior, further demonstrated that WE significantly influenced the probability of selecting higher response categories, with positively worded items showing higher probabilities than negatively worded items. Taken together, the dimensional separation and the systematic differences in category-selection probabilities suggest that, although the Perseverance and Self-Control scales are theoretically designed as unidimensional, WE introduces a systematic method-related influence into scale functioning.

WE has been documented in a wide range of personality and well-being measures, including the Rosenberg Self-Esteem Scale (DiStefano & Motl, 2006), the General Health Questionnaire–12 (Ye, 2009), the Erikson Psychosocial Stage Inventory (Schwartz et al., 2009), the Social Physique Anxiety Scale (Motl et al., 2000), the Callous–Unemotional Traits Inventory (Paiva-Salisbury et al., 2016), the Occupational Personality Scale (McLarnon et al., 2016), the Social Dominance Orientation Scale (Xin & Chi, 2010), and the UCLA Loneliness Scale (Ebesutani et al., 2012). Prior research suggests that positively and negatively worded items can trigger different cognitive processes. Negatively worded items typically require longer processing time than positively worded items (Clark, 1976). Moreover, matching responses to the available options is often more difficult for negative items, and the overall response process tends to take longer (Chessa & Holleman, 2007). Respondents also reread negative items and their response options more frequently than positive items, producing medium-to-large effect sizes (Kamoen et al., 2011). Longer processing time is also an indicator of increased processing complexity (Bassili & Scott, 1996), which may help explain the mechanisms underlying the WE observed in the present study. Collectively, this body of evidence indicates that negatively worded items are more than a mere source of variance; they can systematically alter respondents' cognitive processing styles and, in turn, transform the measurement properties of scales.

The justifications for combining for combining positively and negatively worded items within a single scale are limited (Dalal & Carter, 2015). Therefore, using mixed WE should not be treated as a default solution but rather as a design choice requiring explicit justification in light of the study context and purpose. Consistent with this view, Roszkowski and Soven (2010) and Kam and Fan (2020) reported that some respondents experience greater difficulty in correctly interpreting negatively worded items; as a result, in certain contexts, positively worded items may yield more consistent and reliable results. In this respect, the recommendation by Kam and Fan (2020) is particularly instructive: systematically examining the linguistic and cognitive mechanisms that make negative items difficult may be more productive than debating whether such items should be included. For instance, experimental designs supported by eye-tracking, response time, or comprehension checks can be used to investigate why WE emerges. In addition, decisions about using mixed wording can be informed by the scale–sample–context triad (Alessandri et al., 2010; Kam, 2018). If positive and negative items are to be used together, strategies to mitigate response bias can be planned in advance, such as maintaining a balance between positive and negative items, providing salient warnings and example items in the instructions, limiting item length, and simplifying WE (Roszkowski & Soven, 2010).

This study has several limitations that suggest directions for future research. First, the single cultural context limits generalizability; future studies should examine WE across different cultural contexts. Second, to obtain deeper insight into why WE emerges, future EIRM applications could incorporate interactions between WE and item- and person-level characteristics. For example, the proportion of negative items in the scale, total number of items,

item length, type of negative wording, and person characteristics such as attention, response time, and reading proficiency may be examined within this framework. Finally, because WE was investigated only for the Perseverance and Self-Control scales; its magnitude and direction may vary across domains and should be investigated in other constructs.

Conflict of interest declaration

Conflict of interest

The authors declare that they have no known competing financial interests, institutional affiliations, or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research received no external funding.

Author contributions

Author 1 (Sinem Demirkol): Conceptualization; Methodology; Investigation; Data curation; Formal analysis; Writing—original draft; Writing—review & editing; Visualization.

The author has read and approved the final version of the manuscript.

Data availability statement

The data used in the study can be accessed at https://www.oecd.org/en/data/datasets/pisa-2022-database.html.

Ethics approval and consent to participate

The analyses were conducted using secondary data obtained from an open-access repository. Because the dataset is publicly available and de-identified, and the study involved no direct contact with participants, ethical review and approval were not required

Use of artificial intelligence (AI) tools

During the preparation of this work the author used the Chat GPT in order to improve readability and language. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

TÜRKÇE SÜRÜM

Giriş

Öz bildirim temelli ölçekler, eğitim ve psikoloji alanlarında yaygın kullanılan ölçme araçlarıdır. Katılımcılara belirli bir olguya, duruma veya ifadeye yönelik görüşlerini belirtme olanağı tanır ve farklı düzeylerde katılım ya da onayı temsil eden seçenekler sunar. Böylece bireyler, kendilerine en uygun seçeneği işaretleyerek ilgili tutum, duygu ya da özelliğin ölçülmesine katkı sağlarlar (Likert, 1932). Bu araçların güçlü yönlerine rağmen bazı sınırlılıkları bulunmaktadır. Bu sınırlılıklardan biri de yöntem yanlılığıdır. Yöntem yanlılığı, ölçümlerde gözlenen varyansın, temsil edilmesi amaçlanan yapılardan ziyade ölçme yöntemine atfedilebilen kısmını ifade eder (Williams vd., 1989). Başka bir ifadeyle ölçülmek istenen yapının (ör. benlik saygısı, motivasyon, kaygı) dışında, kullanılan araç, format, bağlam ya da yanıtlayıcıların eğilimlerinden kaynaklanan istenmeyen varyans kaynaklarıdır. Yöntem yanlılığının birçok nedeni olmasına rağmen en bilinen türlerinden biri de ifade etkisidir. İfade etkisi (wording effect, İE), maddelerin olumlu ya da

olumsuz biçimlendirilmesinin yanıtları sistematik biçimde etkilemesi ve ölçülmek istenen yapıyla ilgisiz ek varyans üretmesi olarak tanımlanabilir (Podsakoff vd., 2003).

Bu çalışma, geniş ölçekli uluslararası bir değerlendirme olan PISA 2022 öğrenci anketindeki Azim ve Öz-kontrol ölçeklerinin psikometrik yapısı ve yanıt örüntüleri üzerinde İE'ni incelemeyi amaçlamaktadır. Araştırma iki aşamada yürütülmüştür. İlk aşamada, psikometrik ağ analizinin bir uzantısı olan Keşifsel Grafik Analizi (EGA) ile ölçeklerin boyutsal yapılanması ortaya konmuştur. İkinci aşamada ise Açıklayıcı Madde Tepki Kuramı (Explanatory Item Response Models; EIRM) kapsamında madde parametreleri eşzamanlı olarak kestirilmiştir. Bu sayede İE'nin yanıt olasılıkları ve madde düzeyindeki işleyiş üzerindeki özgül katkısı değerlendirilmiştir.

Yöntem etkisi

Yöntem terimi, ölçmenin hangi biçimde yapıldığına—örneğin madde içeriği, ölçek türü, yanıt formatı ve uygulama bağlamı gibi farklı soyutlama düzeylerine—işaret etmektedir (Fiske, 1982). Bu boyutların her biri, ölçümlerde belirli türde sistematik yanlılıklara zemin hazırlayabilir. Nitekim Podsakoff vd. (2003) tarafından yöntem yanlılığı ortak değerlendirici etkileri, madde bağlamı etkileri, ölçme bağlamı etkileri ve madde özelliklerinden kaynaklanan etkiler olmak üzere dört temel grupta sınıflandırılmıştır. Ortak değerlendirici etkileri, yordayıcı ve ölçüt değişkenlerinin aynı kişi tarafından bildirildiğinde ortaya çıkan ve değişkenler arasındaki ilişkiyi kaynak ortaklığı nedeniyle şişiren yöntemsel sapmadır. Bu türün alt başlıkları arasında tutarlılık güdüsü, sosyal beğenirlik ve geçici duygu durumu yer almaktadır. Madde bağlamı etkileri, bir maddenin ölçek içinde bulunduğu konumdan veya çevresindeki diğer maddelerle olan ilişkisinden doğar. Önceki soruların sonraki cevapları etkilemesi, ölçeğin başındaki soruların yanıtlayıcıda ruh hali oluşturması, ölçek uzunluğu ve farklı yapıları ölçen maddelerin aynı bölümde verilmesi bu kategoriye örnektir. Ölçme bağlamı etkileri, ölçümün gerçekleştirildiği ortam ve koşullardan kaynaklanır. Aynı bağlam içinde toplanan veriler, içerikten bağımsız olarak sahte ilişkiler oluşturabilir. Yordayıcı ve ölçüt değişkenlerinin aynı zamanda ölçülmesi, aynı mekânda toplanması veya aynı araçla (ör. yalnızca çevrim içi anket) elde edilmesi bu kategoriye örnektir. Son olarak ise, madde özelliklerinden kaynaklanan etkiler, ölçekte yer alan maddelerin yazım biçimi, kullanılan ölçek formatı veya dilsel özelliklerinden kaynaklanır. Başka bir deyişle, yanıtlayıcı gerçek tutumundan bağımsız olarak, maddenin biçimsel özelliklerinden etkilenerek sistematik yanıt verebilir. Belirsiz ifadeler, sosyal açıdan arzu edilen yanıtları ima eden maddeler, sürekli aynı ölçek formatının (ör. Likert) veya aynı uç ifadelerin (ör. "asla–her zaman") kullanılması ve olumlu/olumsuz yönde yazılan maddelerin birlikte bulunması bu kategoriye örnek olarak verilebilir (Podsakoff vd., 2003).

Yöntem yanlılığı ile ilgili yürütülen en kapsamlı çalışmalardan biri, psikoloji/sosyoloji, pazarlama, işletme ve eğitim alanlarından 70 çalışmayı derleyen Cote ve Buckley (1987) tarafından gerçekleştirilmiştir. Araştırmacıların bulguları, tipik bir ölçümde varyansın yaklaşık %26,3'ünün ortak yöntem sapmaları gibi sistematik ölçüm hatalarına atfedilebileceğini göstermiştir. Ayrıca bu oranın disiplin ve ölçülen yapı türüne göre anlamlı biçimde değiştiği rapor edilmiştir. Araştırmacılar, pazarlama alanında yöntem kaynaklı varyansın yaklaşık %15 düzeyinde seyrettiğini; eğitim alanında %30'a, iş performansı ölçümlerinde %22'ye ve tutum ölçümlerinde ise %40'a kadar çıktığını belirtmiştir.

Yöntem yanlılığının ölçümler arası ilişkileri de etkileyebildiğine dair kanıtlar mevcuttur; nitekim ortak yöntem varyansı kontrol edildiğinde iki değişken arasındaki açıklanan varyansın ortalama %11, kontrol edilmediğinde ise ortalama %35 düzeyinde olduğu raporlanmıştır (Fuller vd., 1996; Lowe vd., 1996; Podsakoff vd., 2000; Wagner ve Gooding, 1987). Buna ek olarak Podsakoff vd., (2003) ölçümler ortak yöntem varyansı içeriyorsa, yordayıcı ile kriter değişkeni arasındaki gözlenen ilişkinin yaklaşık %25 oranında düşük tahmin edilebileceğini göstermiştir. Bu genel tablo, yöntem etkilerinin değişkenler arasındaki ilişkileri kimi durumlarda şişirip kimi durumlarda bastırabildiğini ve buna bağlı olarak Tip I ile Tip II hata olasılıklarını artırabildiğini düşündürmektedir (Campbell ve Fiske, 1959; Cote ve Buckley, 1987; Podsakoff vd., 2003). Sonuç

olarak, yöntem etkilerini belirlemek ve uygun biçimde modellemek, sağlıklı çıkarımları destekleyerek elde edilen sonuçların geçerliğini güçlendirir.

İfade etkisi

Çok maddeli Likert ölçekleri tutumlar, inançlar, değerler ve diğer gizil yapılar hakkında veri toplamak için yaygın ve önerilen bir yaklaşımdır (Peterson, 1994). Bununla birlikte bu güçlü çerçeve, yanıtlayıcı davranışından kaynaklanan bazı yöntemsel riskler taşır. Örneğin, onay eğilimi/yanlılığı (acquiescence bias), olumsuz maddeler yerine olumlu maddeleri seçme eğilimini ifade eder (Watson, 1992). Benzer biçimde, madde içeriğinden bağımsız olarak derecelendirme ölçeklerinin uç kategorilerini sistematik biçimde tercih etme ya da bu kategorilerden kaçınma eğilimi, aşırı yanıt stili (extreme response style) olarak adlandırılır ve derecelendirme verilerinde potansiyel bir yanlılık kaynağıdır (Greenleaf, 1992). Bu tür sistematik yanıt eğilimleri, ölçülmek istenen örtük yapının gerçek değerinin olduğundan daha yüksek ya da daha düşük tahmin edilmesine neden olabilir (Winkler vd., 1982).

Ölçeklerde sistematik yanıt yanlılıklarını azaltmanın yaygın bir yolu, aynı yapıyı hem olumlu hem de olumsuz biçimde formüle edilmiş maddelerle örneklemektir. Bu yaklaşım, tek yönlü ifadelerden kaynaklanan eğilimleri dengeleyerek daha tarafsız yanıtlar üretmeyi ve ölçümlerin geçerliğini artırmayı amaçlar (DeVellis, 2005). Bu bağlamda, ölçek geliştirme literatüründeki yanıt yanlılığını ortadan kaldırma isteği (Nunnally, 1967), olumlu–olumsuz madde eşleşmelerinin aynı araçta yer almasını desteklemiş, böylelikle onay eğilimlerinin etkisinin azaltılması hedeflenmiştir (Hinz vd., 2007). Bu stratejinin dayandığı kritik varsayım ise olumlu ve olumsuz maddelerin aynı temel yapıyı ölçtüğüdür (Marsh, 1996). Bu varsayım doğrultusunda, negatif maddelere verilen yanıtlar ters puanlanır ve pozitif maddelerle aynı şekilde işlenir (Horan vd., 2003).

Dalal ve Carter (2015), olumsuz maddeleri iki türde ele alır ve bu ayrımın ölçme sürecindeki sonuçları farklılaştırabileceğini vurgular. İlk tür, taban tabana zıt maddelerdir. Dilsel olumsuzluk kullanılmadan aynı yapının karşıt kutuplarını hedefleyen, anlamca birbirinin karşıtı ifadelerden oluşur (ör., "Sistematik çalışırım" ile "Gelişigüzel çalışırım", "Yeni fikirlere açığım" ile "Yeni fikirlere kapalıyım"). İkinci tür ise olumsuzlanmış normal maddelerdir. Olumlu bir ifadenin "değil" ya da -ma/-me gibi mantıksal olumsuzluk ekleriyle doğrudan tersi hâline getirilmesi de söz konusudur (ör., "Zamanımı etkin kullanırım" ile "Zamanımı etkin kullanmam", "Randevulara zamanında giderim" ile "Randevulara zamanında gitmem"). Literatürde bu iki tür madde (birbirine tamamen zıt maddeler ve olumsuzlanmış normal maddeler) arasındaki ayrım her zaman net değildir ve her iki tür de genel anlamda olumsuz ifade edilmiş maddeler olarak kabul edilir. Uygulamada, olumlu ve olumsuz maddelerin işlevsel olarak birbirlerinin yerine geçebileceği varsayılır (Lindwall vd., 2012). Bu varsayıma göre, bir katılımcının olumlu bir maddeye "kesinlikle katılıyorum" demesi, mantıksal karşılığı olan olumsuz maddeye "kesinlikle katılmıyorum" demesiyle ayna simetrisi göstermelidir (Marsh, 1996). Ne var ki bu konudaki bulgular tutarlı değildir. Bazı çalışmalar, anlamca eşdeğer madde çiftlerinde katılımcıların olumsuz maddelere onay verme eğiliminin görece daha yüksek olduğunu bildirirken (Holleman, 1999; Schriesheim ve Hill, 1981), diğer çalışmalar olumlu maddelerde ortalama puanların sistematik olarak daha yüksek olduğunu raporlamıştır (Weem vd., 2003)

Olumsuz maddelerin ölçeklere dahil edilmesi, katılımcıları daha dikkatli yanıt vermeye zorlar ve böylece otomatik yerine daha kontrollü bilişsel işlemeyi teşvik eder (Hinkin, 1995; Idaszak ve Drasgow, 1987). Dolayısıyla olumsuz maddeler adeta birer bilişsel hız tümseği gibi işlev görerek anketin daha özenli biçimde tamamlanma ihtimalini artırır (Podsakoff vd., 2003). Böyle bir uygulama, hem ölçme aracının hedef özelliği daha doğru yakalamasını (Anderson ve Gerbing, 1988) hem ölçüm etkinliğini yükseltmesini (Worcester ve Burns, 1975) hem de hedef özelliğin daha iyi bir kapsamını yakalamasını sağlayabilir (Weijters ve Baumgartner, 2012).

Araştırmalar, olumlu ve olumsuz maddelerin birlikte yer aldığı ölçeklerde maddeler arası korelasyonların, yalnızca olumlu maddelerden oluşan ölçeklere kıyasla daha zayıf olduğunu (DiStefano ve Motl, 2006) ve bu iki tür maddenin birlikte kullanımının ölçeğin iç tutarlılığını

azalttığını göstermektedir (Lee vd., 2008). İE'nin yönü konusunda bulgular da çelişkilidir. Bazı çalışmalar olumsuz maddelerin olumlu maddelere kıyasla daha güçlü İE ürettiğini bildirirken (DiStefano ve Motl, 2009; Marsh, 1996), diğer çalışmalar daha belirgin etkinin olumlu maddelerde ortaya çıktığını göstermiştir (Lindwall vd., 2012). Ayrıca olumlu ve olumsuz maddelerin birlikte kullanımı, hedeflenen özelliğe ilgisi olmayan ortak varyansın ölçüme sızması yoluyla ölçeğin faktör yapısını da dönüştürebildiği ortaya konulmuş (Corwyn, 2000; John vd., 2019), olumsuz yönde yazılmış maddelerin sıklıkla ayrı bir faktöre yüklendiği rapor edilmiştir (Benson ve Hocevar, 1985; Pilotte ve Gable, 1990). Bununla birlikte bu tür yapay faktörlerin ilgili maddeler olumlu biçimde yeniden yazıldığında ortadan kalkabildiği gösterilmiştir (Harvey vd., 1985; Idaszak ve Drasgow, 1987). Bu bulgular İE'nin ölçeğin varsayılan tek boyutlu yapısını zedeleyerek, kimi durumlarda ihlal etmesine yol açabildiğini ortaya koymaktadır (Cordery ve Sevastos, 1993; Hevey vd., 2010).

Araştırmanın amacı

Tüm bu bilgiler ışığında bu çalışmanın amacı olumlu ve olumsuz maddelerden oluşan Azim ve Öz-kontrol ölçeklerinde İE'nin faktör yapısında bir farklılığa yol açıp açmadığını incelemek ve bireylerin yanıt davranışları üzerindeki etkisini araştırmaktır. Bu amaçla aşağıdaki sorulara yanıt aranmıştır.

1. Azim ve Öz-kontrol ölçeklerinin faktör/boyut yapısı İE'den etkilenmekte midir?
2. İE yanıtlayıcıların yanıt davranışlarını anlamlı biçimde değiştiriyor mu?

Yöntem

Çalışma grubu

Bu araştırmanın çalışma grubunu, PISA 2022 Türkiye uygulamasına katılan toplam 7245 öğrenci oluşturmaktadır. Katılımcıların 3561'i (%49,2) kız, 3684'ü (%50,8) erkektir. Türkiye, PISA 2022'ye Türkiye İstatistiki Bölge Birimleri Sınıflandırması (İBBS) Düzey-1 kapsamındaki 12 bölgeyi temsil eden 196 okul aracılığıyla katılmıştır. Okul türü dağılımında en büyük pay Anadolu liselerindeyken (%56) bunu mesleki ve teknik Anadolu liseleri (%23) izlemektedir (MEB, 2022).

Veri toplama aracı

Bu araştırmada veri toplama aracı olarak, PISA 2022 öğrenci anketinde yer alan "Genel Sosyal ve Duygusal Özellikler" modülündeki Öz-kontrol (self-control) ve Azim (perseverance) ölçekleri kullanılmıştır. Her iki ölçek, OECD'nin Sosyal ve Duygusal Beceriler Araştırması çerçevesine dayanmaktadır ve Beş Faktör Kişilik Modeli'nde sorumluluk (conscientiousness) boyutunun alt bileşenleri olarak konumlanır. Ölçekler, yapı-içi matris örnekleme (within-construct matrix sampling) yaklaşımıyla uygulanmış ve puanlanmıştır (OECD, 2024).

Öz-Kontrol ölçeği (ST309), öğrencilerin dikkat dağıtıcı uyaranlardan uzak durabilme, ani dürtülere kapılmadan hareket etme ve uzun vadeli hedeflerine odaklanabilme becerilerini ölçmeyi amaçlamaktadır. Ölçek toplam 10 maddeden oluşmakta olup bunların 6'sı olumlu, 4'ü olumsuz yönde ifade edilmiştir. Öğrenciler, her bir maddeyi "Kesinlikle katılmıyorum" ile "Kesinlikle katılıyorum" arasında değişen beşli Likert tipi derecelendirme ölçeği üzerinde değerlendirmiştir. Ölçeğe ait maddeler Ek 1 de verilmiştir.

Azim/istikrar ölçeği (ST307) ise öğrencilerin başladıkları görevleri tamamlama, zorluklar karşısında çaba göstermeye devam etme ve kolayca vazgeçmeme eğilimlerini ölçmektedir. Ölçek toplam 10 maddeden oluşmakta olup bunların 6'sı olumlu, 4'ü olumsuz yönde ifade edilmiştir. Öğrenciler, her bir maddeyi "Kesinlikle katılmıyorum" ile "Kesinlikle katılıyorum" arasında değişen beşli Likert tipi derecelendirme ölçeği üzerinde değerlendirmiştir. Ölçeğe ait maddeler Ek 1 de verilmiştir.

Madde düzeyinde İE değişkeni araştırmacı tarafından oluşturulmuştur. Olumsuz maddeler = 0 (referans), olumlu maddeler = 1 olarak kodlanmıştır. Bu değişken, EIRM çerçevesinde madde konumu parametreleri üzerindeki İE'yi sınamak üzere kovaryat olarak kullanılmıştır. Bu kodlama ile katsayıların pozitif olması, olumlu maddelerin olumsuzlara kıyasla daha yüksek konum/eşik (daha 'zor onaylanma/kabul edilme') eğilimine işaret eder.

Verilerin analizi

Verilerin analizi iki aşamada yürütülmüştür. İlk olarak maddelerin hangi boyutlarda kümelendiğini belirlemek için Keşifsel Grafik Analizi (Exploratory Graph Analysis, EGA) kullanılmıştır. Golino ve Epskamp (2017) tarafından geliştirilen EGA, ağ psikometrisine dayalı olarak faktör belirleme yöntemlerine alternatif bir yaklaşım sunmaktadır. Bu yöntemde düğümler maddeleri, kenarlar ise maddeler arasındaki kısmi korelasyonları temsil etmektedir. Maddeler arasındaki pozitif ilişkiler yeşil çizgilerle, negatif ilişkiler ise kırmızı çizgilerle ifade edilir. Çizgi kalınlığı ilişki gücüyle orantılıdır, ilişki arttıkça kalınlaşır, zayıfladıkça incelir. Ayrıca düğümlerin renkleri maddelerin hangi boyuta ait olduğunu gösterir.

EGA analizlerinin sağlıklı biçimde yürütülebilmesi için veri kalitesi ve ilişki matrisinin uygunluğu ön koşul olarak değerlendirilmiştir. Bu kapsamda eksik verinin düzeyi ve örüntüsü incelenmiş, aşırı düzeyde eksikliği bulunan maddeler ile tek bir kategoriye yığılma gibi korelasyon kestirimini bozabilecek durumlar kontrol edilmiştir. EGA sürecinde önce maddeler arasındaki kısmi korelasyonlar hesaplanmıştır. Daha sonra GLASSO (Grafiksel En Küçük Mutlak Küçültme ve Seçim Operatörü) ile küçük korelasyonların sahte bağlar oluşturması engellenerek daha sade ve yorumlanabilir bir ağ modeli elde edilmiştir. Son aşamada Walktrap algoritması ile değişkenler arasındaki kümelenmeler belirlenmiştir. Bu kümeler faktör analizindeki boyutlara karşılık gelmektedir. Böylece hem faktör sayısı kestirilmiş hem de maddelerin hangi boyutlarda kümelendiği görselleştirilmiştir (Christensen vd., 2020). EGA analizleri R (R Core Team, 2025) ortamında *EGAnet* paketi (Golino ve Christensen, 2022) kullanılarak yürütülmüştür.

Ölçeklere ait boyut yapısı belirlendikten sonra ikinci adımda İE'nin maddenin kabul edilme olasılığı üzerindeki etkisi incelenmiştir (Likert tipi ölçeklerde kabul edilme veya onaylanma olasılığı başarı testindeki madde güçlüğü ile eşdeğerdir). Bunun için madde ve birey parametrelerinin eş zamanlı olarak kestirildiği Likert tipi sıralı madde formatına uygun olan polikotom Açıklayıcı Madde Tepki Kuramı (Explanatory Item Response Models, EIRM) kullanılmıştır (De Boeck ve Wilson, 2004).

EIRM, geleneksel Madde Tepki Kuramı'nın (MTK) yalnızca ölçmeye odaklanan yaklaşımını genişleterek ölçme (madde/kişi parametreleri) ile açıklamayı (madde ve kişi özelliklerinin yanıtları nasıl etkilediği) aynı modelde birleştirir (De Boeck ve Wilson, 2004). Bu yaklaşım, istatistiksel olarak Genelleştirilmiş Doğrusal Karma Modeller (GLMM) çerçevesinde ifade edilir: Her bireyin her bir maddeye verdiği yanıt, tekrarlı ölçüm olarak ele alınır ve böylece madde ve kişi düzeyindeki etkiler tek bir model içinde birlikte tahmin edilebilir.

Likert tipi maddelerde EIRM, polikotom MTK modelleri (örn. PCM, GRM, RSM) üzerinden uygulanır. Eşik (threshold) parametreleri madde ve/veya kişi kovaryatlarının bir fonksiyonu olarak yazılarak kategori geçişlerinin hangi koşullarda kaydığı açıklanır (Stanke ve Bulut, 2019). Bu çalışmada EIRM analizine başlamadan önce yanıt verisi uzun formata dönüştürülerek her satır kişi × madde yanıtını temsil edecek şekilde düzenlenmiştir. Ardından açıklayıcı kısım eklenerek, eşiklerin (ve/veya madde parametrelerinin) madde ve kişi özelliklerinin bir fonksiyonu olarak değişmesine izin verilmiştir. Bu sayede madde özellikleri (ör. maddenin olumlu/olumsuz ifadesi), kişi özellikleri (ör. cinsiyet, ana dil) ve gerekirse kişi×madde etkileşimleri aynı modelde test edilebilir.

Polikotom EIRM analizleri R da bulunan *eirm* (Bulut, 2021) paketi kullanılarak gerçekleştirilmiştir. Maddeler kişiler içinde kümelendiğinden ham yanıt verisi, kişi–satır/madde–sütun yapısındaki, geniş formattan uzun formata dönüştürülmüştür. Böylece her satır tek bir kişi–madde gözlemini

temsil edecek, maddeler de kişiler içinde iç içe (hiyerarşik) yer alacak şekilde çok düzeyli modellere uygun hale getirilmiştir. Ayrıca polikotom (Likert tipi) yanıtlar, eirm paketindeki *polyreformat* işlevi kullanılarak birden çok ikili göstergeye ayrıştırılmış, bu sayede özgün yanıt yapısı korunarak GLMM çerçevesinde ikili tepki modelleri kurulabilmiştir.

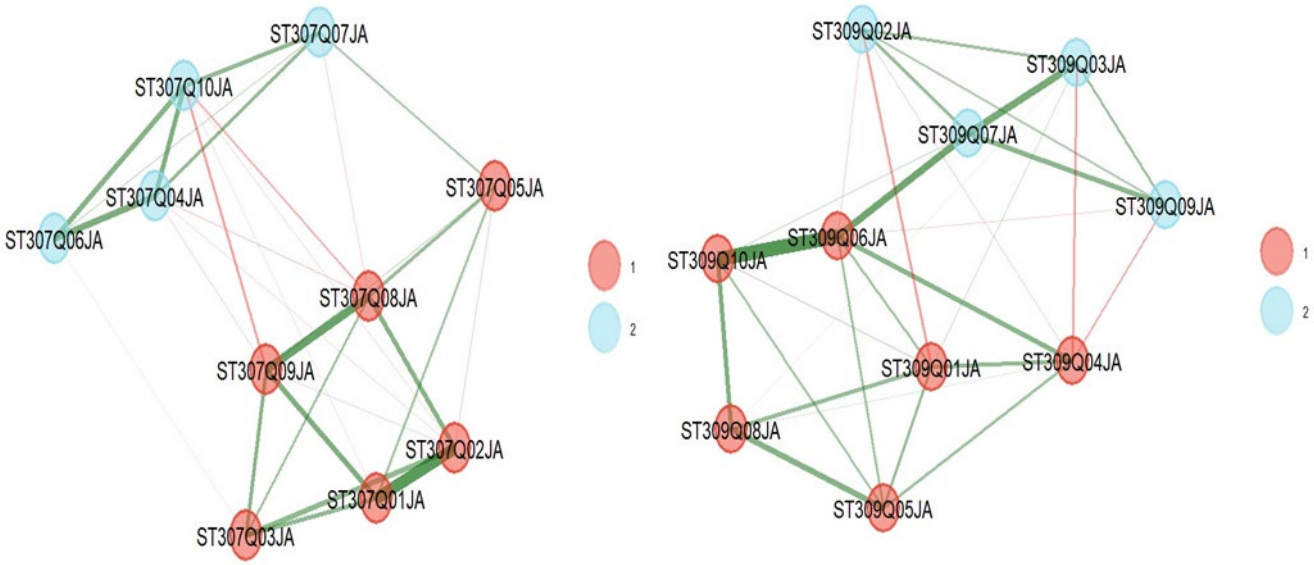
Tablo 1 Çok kategorili yanıtların ikili yanıta dönüştürülmesi

Orijinal Cevap	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum
Kesinlikle katılmıyorum	0	NA	NA	NA
Katılmıyorum	1	0	NA	NA
Kararsızım	NA	1	0	NA
Katılıyorum	NA	NA	1	0
Kesinlikle katılıyorum	NA	NA	NA	1

Açıklayıcı MTK çerçevesinde İE'ni modele dâhil etmenin anlamlı olup olmadığını sınamak için hiyerarşik model karşılaştırmaları yapılmıştır. Bunun için Temel model (M_0): İfade etkisini içermeyen model ve Genişletilmiş model (M_1): İfade etkisini içeren model uyum açısından karşılaştırılmıştır. Değerlendirmede AIC, BIC ve log-likelihood (logLik) değerleri incelenmiş, modeller iç içe (nested) geçtiğinden olasılık oranı testi (Likelihood-Ratio Test, LRT; χ^2, Δdf) uygulanmıştır. Daha düşük AIC/BIC, daha yüksek (daha az negatif) logLik ve anlamlı $\chi^2(\Delta df)$, İE'nin modele eklenmesinin model uyumu iyileştirdiği biçiminde yorumlanmıştır.

Bulgular

Şekil 1, Azim ve Öz-Kontrol ölçekleri için EGA sonuçlarını göstermektedir. Düğümler maddeleri, kenarlar ise maddeler arasındaki ilişkileri göstermektedir. Yeşil çizgiler pozitif, kırmızı çizgiler ise negatif ilişkileri göstermektedir. Ayrıca kenarlar kalınlaştıkça ilişki gücü artmaktadır. Kırmızı düğümler olumlu, mavi düğümler olumsuz maddelerdir.



Şekil 1 Azim (soldaki) ve Öz kontrol (sağdaki) ölçeklerine ait EGA sonuçları

Şekil 1'de iki boyut gözlenmiştir (şekillerin sağındaki kırmızı ve mavi göstergeler iki boyut olduğunu göstermektedir) ve bu boyutlar tamamen maddelerin ifade ediliş türüne (olumlu–olumsuz) karşılık gelmektedir. Boyutlar içi bağlantılar pozitif ve güçlü, boyutlar arası bağlantılar ise çoğunlukla zayıftır. Bu desenler, her iki ölçekte de İE'nin güçlü olabileceğini ve maddelerin içerikten çok ifade biçimine göre kümelenebildiğini göstermektedir. Bu bulgu, PISA'da öğrenci anketi ölçek puanlarının genellikle IRT temelli türetilmiş değişkenler olarak raporlandığı standart

ölçekleme yaklaşımıyla (ör. Rasch/PCM temelli ölçekleme) karşılaştırıldığında, tek boyut varsayımının bu ölçeklerde sınırlı kalabileceğine işaret etmektedir (OECD, 2024)

İkinci aşamada, elde edilen boyutlanmanın yanıt davranışlarına yansıyıp yansımadığını incelemek için EIRM modelleri kurulmuştur. Her iki ölçek için İE'nin modele anlamlı katkısını değerlendirmek üzere, önce İE'ni içermeyen temel/boş model (M0) tanımlandı, ardından İE'ni içeren genişletilmiş model (M1) kuruldu.

Tablo 2 Kurulan modellere ait model uyum değerleri

Ölçek	Modeller	n	logLik	AIC	BIC	Sapma	$\Delta\chi^2$ (Δdf)	p
Azim	M_0	5	−41170	82351	82396	82341	—	—
	M_1	6	−40397	80806	80860	80794	1546.2 (1)	.001
Öz-Kontrol	M_0	5	−40347	80704	80749	80694	—	—
	M_1	6	−40107	80226	80280	80214	479.50 (1)	.001

Not. M_0 = temel model (ifade etkisi yok); M_1 = genişletilmiş model (ifade etkisi dahil edilmiş model). Daha düşük AIC/BIC ve daha yüksek (daha az negatif) logLik daha iyi uyuma işaret eder; anlamlı LRT (χ^2) M_1'in M_0'a göre üstünlüğünü destekler.

Tablo 2 incelendiğinde, İE'nin modele dahil edilmesinin model uyumu anlamlı ölçüde iyileştirdiği görülmektedir. Bu sonuç doğrultusunda İE'ni içeren modellere ait sabit etkiler Tablo 3 de verilmiştir.

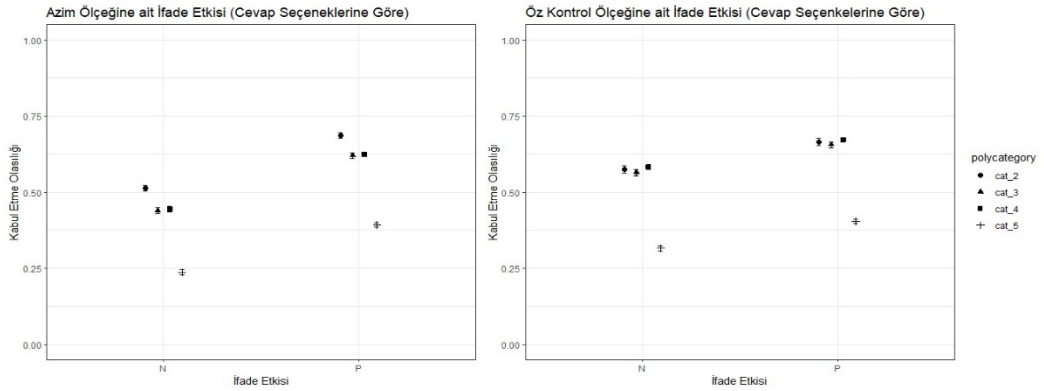
Tablo 3 Azim ve öz kontrol ölçeklerinde ortaya çıkan ifade etkisine ait sabit etkiler

	Azim			Öz kontrol		
Değişken	Tahmin (SH)	z	p	Tahmin (SH)	z	p
Kesişim	0.052(0.019)	2.68	.007 **	0.298 (0.024)	12.28	< .001 ***
İfade etkisi	0.732(0.019)	38.73	< .001 ***	0.385 (0.018)	21.81	< .001 ***
polykategori_3	−0.294(0.025)	−11.62	< .001 ***	−0.040 (0.030)	−1.32	.187
polykategori_4	−0.277(0.025)	−10.93	< .001 ***	0.034 (0.028)	1.21	.225
polykategori_5	−1.222(0.027)	−45.11	< .001 ***	−1.071 (0.028)	−38.17	< .001 ***

*Not. B = tahmin katsayısı (Estimate); SH = standart hata. p değerleri: *** < .001, ** < .01. elde edilen değerler log-odds ölçeğindedir.*

Polykategori değişkendeki eşik parametreleri, komşu kategoriler arasındaki mesafeyi temsil etmektedir. Referans eşik 'Kesinlikle katılmıyorum'dan 'katılmıyorum'a geçiş olarak belirlenmiştir. Katsayılar log-odds ölçeğindedir. Negatif değer, ilgili kategoriyi seçme olasılık oranının (odds) referansa göre daha düşük olduğunu gösterir. Buna göre, Azim ölçeğinde referans konumuna göre kategori 3 (katılmıyorumdan kararsıza olan uzaklık), kategori 4 (kararsızımdan katılıyoruma olan uzaklık) ve kategori 5 (katılıyorumdan kesinlikle katılıyoruma olan uzaklık) seçme olasılıklarının daha düşük olduğu görülmektedir. En fazla düşüş ise kategori 5'de meydana gelmektedir. Öz kontrol ölçeğinde ise durum biraz daha farklıdır. Öğrencilerin referans konumuna göre kategori 3 ve 4'ü seçme olasılıklarındaki artış ya da azalış anlamlı değilken kategori 5'i (katılıyorumdan kesinlikle katılıyorumu) seçme olasılıkları düşüş göstermiştir. Başka bir ifadeyle, 'Kesinlikle katılmıyorum'dan 'Katılmıyorum'a eşiği referans alındığında, Azim ölçeğinde Kategori 3–5'in seçilme olasılık oranlarının daha düşük olduğu (en fazla Kategori 5), Öz kontrol ölçeğinde ise yalnızca Kategori 5'te anlamlı bir düşüşün meydana geldiği görülmüştür.

İE her iki ölçekte de istatistiksel olarak anlamlı bulunmuştur. Azim ölçeğinde olumsuz ifadelere kıyasla olumlu ifadelere verilen yanıtların kabul edilme/onaylanma olasılıklarının daha yüksek olduğu görülmüştür (β = 0.732, SE = 0.019, z = 38.73, p < .001). Azim ölçeği için lojit ölçeğindeki 0.732 değerinin üsseli alındığında (exp(0.732)) elde edilen 2.08 değeri odds oranını verir. Odds oranı, bir maddenin doğru yanıtlanma olasılığının yanlış yanıtlanma olasılığına oranıdır. Olumsuz bir maddenin kabul edilme olasılığı 0.50 ise, olumlu bir maddenin kabul oranı yaklaşık 0.68 olacaktır. Benzer şekilde, Öz Kontrol ölçeğinde İE anlamlıdır (β = 0.385, SE = 0.018, z = 21.81, p < .001). Öz kontrol ölçeği için (Azim ölçeği için yapılan işlemler tekrarlandığında) olumsuz bir maddenin kabul edilme olasılığı 0.50 ise, olumlu bir maddenin kabul oranı yaklaşık 0.59 olacaktır.



Şekil 2 Azim ve Öz-kontrol Ölçeği için ifade etkilerine ait grafikler

Elde edilen sonuçlar ve grafik incelendiğinde her iki ölçekte de olumlu maddelerin kabul olasılıklarının olumsuz maddelere göre daha yüksek olduğu görülmüştür.

Sonuç ve öneriler

Bu çalışmanın amacı, İE'nin ölçek yapısı ve yanıtlayıcı davranışları üzerindeki etkisini incelemektir. Bu kapsamda öncelikle psikometrik ağ analizinin bir uzantısı olan EGA ile olumlu ve olumsuz maddelerin beraber bulunduğu Azim ve Öz-kontrol ölçeklerin boyutlanması incelenmiş, ardından elde edilen bu boyutlanmanın yanıt davranışları üzerindeki etkisi EIRM ile test edilmiştir. Böylece, İE'nin yalnızca ölçme yapısındaki etkisi değil, bu etkinin yanıt davranışlarına yansıması bütüncül biçimde değerlendirilmiştir

Elde edilen bulgular, Azim ve Öz-kontrol ölçeklerinde yer alan olumlu ve olumsuz maddelerin farklı boyutlar oluşturabildiğini göstermektedir. Nitekim EGA sonuçları, her iki ölçekte de olumlu ve olumsuz maddelerin ayrı kümelerde toplandığını ortaya koymuştur. Bu yapısal ayrışmanın yanıt davranışına yansımasını inceleyen EIRM bulguları ise, İE'nin maddelerin onaylanma olasılıklarını anlamlı biçimde etkilediğini, olumlu maddelerin, olumsuzlara kıyasla daha yüksek onay olasılıklarına sahip olduğunu göstermektedir. Boyutsal ayrışma ile onay olasılıklarındaki fark birlikte değerlendirildiğinde, Öz-kontrol ve Azim ölçekleri kuramsal olarak tek boyutlu tasarlansa da, İE'nin ölçeğin işleyişine sistematik bir yöntemsel etki kattığı şeklinde yorumlanabilir.

İE çok sayıda kişilik ve iyi oluş ölçeğinde belgelenmiştir. Bunlar arasında Rosenberg Benlik Saygısı Ölçeği (DiStefano ve Motl, 2006), Genel Sağlık Ölçeği–12 (Ye, 2009), Erikson Psikososyal Evre Envanteri (Schwartz vd., 2009), Sosyal Fizik Kaygı Ölçeği (Motl vd., 2000), Duygusuz-Duygusuz Özellik Envanteri (Paiva-Salisbury vd, 2016), Mesleki Kişilik Ölçeği (McLarnon vd., 2016), Sosyal Baskınlık Yönelimi Ölçeği (Xin ve Chi, 2010) ve Yalnızlık Ölçeği (Ebesutani vd., 2012) yer almaktadır. Yapılan çalışmalar olumlu ve olumsuz maddelerin farklı bilişsel süreçleri tetiklediğini göstermektedir. Olumsuz maddeler, olumlu maddelere kıyasla daha fazla işlem süresi gerektirmektedir (Clark, 1976). Ayrıca bu maddelerde yanıtların seçeneklerle eşleştirilmesi daha güç olmakta ve süreç genel olarak daha uzun sürmektedir (Chessa ve Holleman, 2007). Katılımcılar olumsuz maddelerin soru ve yanıt seçeneklerini olumlu maddelere göre daha sık yeniden okumakta, bu da orta ya da büyük etki büyüklükleriyle sonuçlanmaktadır (Kamoen vd., 2011). Daha uzun işlem süresi, artan işlem karmaşıklığının da bir göstergesidir (Bassili ve Scott, 1996) ve bu artış, bu çalışmada gözlenen İE'nin mekanizmasını açıklamaya yardımcı olmaktadır. Bu kanıtlar, olumsuz maddelerin tek başına bir varyans kaynağı olmaktan öte, yanıtlayıcıların bilişsel işleme tarzlarını sistematik olarak değiştirebildiğini ve dolayısıyla ölçeklerin ölçme özelliklerini dönüştürebildiğini göstermektedir.

Bir yapıyı ölçmek üzere hazırlanmış bir ölçekte olumlu ve olumsuz maddelerin birlikte kullanılmasına yönelik gerekçelerin sayısı sınırlıdır (Dalal ve Carter, 2015). Bu nedenle olumlu-

olumsuz maddelerin bir arada kullanılması varsayılan bir çözüm olarak değil, bağlam ve amaç doğrultusunda açıkça gerekçelendirilmesi gereken bir tasarım tercihi olarak görülmelidir. Nitekim Roszkowski ve Soven (2010) ile Kam ve Fan (2020), bazı katılımcıların olumsuz maddeleri doğru anlamlandırmada daha fazla güçlük yaşadığını, bu nedenle belirli bağlamlarda olumlu maddelerin daha tutarlı ve güvenilir sonuçlar üretebildiğini raporlamışlardır. Bu bağlamda Kam ve Fan (2020)'ın önerisi özellikle yol göstericidir. Olumsuz maddelerin anlaşılmasını güçleştiren dilsel ve bilişsel mekanizmaların sistematik olarak incelenmesi, bu tür maddelerin ölçeğe dahil edilip edilmemesini tartışmaktan daha faydalı olabilir. Örneğin, göz izleme, yanıt süresi veya anlama denetimleriyle desteklenen deneysel tasarımlarla İE'nin neden ortaya çıktığı incelenebilir. Ayrıca ölçek tasarımında, olumlu ve olumsuz maddelerin birlikte kullanımına ilişkin karar ölçek-örneklem-bağlam üçlüsünü dikkate alınarak da verilebilir (Alessandri vd., 2010; Kam, 2018). Olumlu-olumsuz maddeler bir arada kulanılacaksa yanıtlayıcıların tepki yanlılığını azaltmaya dönük stratejiler baştan planlanabilir, olumlu-olumsuz madde dengesinin korunması, yönergelerde görünür uyarılar ve örnek maddeler, madde uzunluğunun sınırlandırılması ve madde ifadelerinin sadeleştirilmesi gibi önlemler uygulanabilir (Roszkowski ve Soven, 2010).

Bu çalışmanın bazı sınırlılıkları bulunmaktadır ve buna bağlı olarak gelecekteki araştırmalar için çeşitli öneriler sunulabilir. Öncelikle, örneklemin tek bir kültürel/ulusal bağlamdan gelmesi bulguların genellenebilirliğini sınırlandırmaktadır. İleride yapılacak olan çalışmalarda farklı kültürlerden gelen öğrencilere ait verilerde İE incelenebilir. Ayrıca İE'nin neden ortaya çıktığına dair kapsamlı bir içgörü sağlamak amacıyla EIRM modellerine birey ve madde özelliklerinin İE ile etkileşimi eklenerek eş zamanlı olarak incelenebilir. Örneğin ölçekte bulunan olumsuz madde oranı, ölçekteki toplam madde sayısı, madde uzunluğu, olumsuz madde türü veya birey özellikleri olarak dikkat düzeyi, yanıtlama süresi, okuma yeterliliği gibi özellikler bu bağlamda ele alınabilir. Bu çalışmada İE Öz-kontrol ve Azim ölçeklerinde incelenmiştir. Alana bağlı olarak da bu etkinin büyüklüğü veya yönü değişebilir. Bu sebeple farklı bağlamlara ait ölçeklerde de İE incelenebilir.

Çıkar çatışması beyanı

Çıkar çatışması

Yazar(lar), bu çalışmada sunulan araştırmayı etkileyebilecek herhangi bir bilinen finansal çıkar, kurumsal bağlantı ya da kişisel ilişki bulunmadığını beyan eder.

Finansman

Bu araştırma için dış finansman desteği alınmamıştır.

Yazar katkıları

Yazar 1 (Sinem Demirkol): Kavramsallaştırma; Yöntem; Araştırma/İnceleme; Veri düzenleme; Biçimsel analiz; Özgün taslak yazımı; Yazım—gözden geçirme ve düzenleme; Görselleştirme.

Veri erişilebilirliği

Çalışmada kullanılan verilere https://www.oecd.org/en/data/datasets/pisa-2022-database.html adresinden erişilebilir.

Etik kurul onayı ve katılımcı onamı

Analizler, açık erişimli bir veri havuzundan elde edilen ikincil veriler kullanılarak gerçekleştirilmiştir. Veri seti kamuya açık ve kimlik bilgileri gizlenmiş olduğundan ve çalışma katılımcılarla doğrudan temas gerektirmediğinden, etik inceleme ve onay gerekli olmamıştır.

Yapay zekâ (YZ) araçlarının kullanımı

Bu çalışmanın hazırlanması sırasında yazar, okunabilirliği ve dili iyileştirmek için Chat GPT'yi kullanmıştır. Bu araç/hizmeti kullandıktan sonra yazar, içeriği gerektiği gibi gözden geçirip düzenlemiş ve yayının içeriği konusunda tüm sorumluluğu üstlenmiştir.

Kaynaklar

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P. M., Barbaranelli, C., Medda, E., Nisticò, L., Stazi, M. A., & Caprara, G. V. (2010). Much more than model fitting? Evidence for the heritability of method effect associated with positively worded items of the Life Orientation Test–Revised. *Structural Equation Modeling, 17*(4), 642–653. https://doi.org/10.1080/10705511.2010.510064

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Bassili, J. N., & Scott, S. B. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*(3), 390–399. https://doi.org/10.1086/297760

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales on elementary school children. *Journal of Educational Measurement, 22*(3), 231–240.

Bulut, H. C., & Bulut, O. (2022). Item wording effects in self-report measures and reading achievement: Does removing careless respondents help? *Studies in Educational Evaluation, 72*, 101126. https://doi.org/10.1016/j.stueduc.2022.101126

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses* (R package version 0.3.0). https://CRAN.R-project.org/package=eirm https://doi.org/10.5281/zenodo.4556285

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology, 21*(2), 203–225. https://doi.org/10.1002/acp.1337

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality, 34*(6), 1095–1108.

Clark, H. H. (1976). *Semantics and comprehension*. The Hague, Netherlands: Mouton.

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised Job Diagnostic Survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78*(1), 141–143. https://doi.org/10.1037/0021-9010.78.1.141

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*(4), 357–379.

Cote, J. A., & Buckley, R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24*(3), 315–318.

Dalal, D. K., & Carter, N. T. (2015). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 112–132). New York, NY: Routledge.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

DeVellis, R. F. (2005). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440–464.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 46*(3), 309–313. https://doi.org/10.1016/j.paid.2008.10.020

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., & Young, J. (2012). The importance of modeling method effects: Resolving the (uni)dimensionality of the Loneliness Questionnaire. *Journal of Personality Assessment*. https://doi.org/10.1080/00223891.2011.627967

Fiske, D. (1982). Convergent–discriminant validation in measurement and research strategies. In D. Brinberg & L. Kidder (Eds.), *New directions for methodology of social and behavioral science: Forms of validity in research*. San Francisco, CA: Jossey-Bass.

Fuller, J. B., Patterson, C. E. P., Hester, K., & Stringer, S. Y. (1996). A quantitative review of research on charismatic leadership. *Psychological Reports, 78*(1), 271–287.

Golino, H. F., & Christensen, A. P. (2022). *EGAnet: Exploratory graph analysis—A framework for estimating the number of dimensions in multivariate data using network psychometrics* (R package). https://CRAN.R-project.org/package=EGAnet

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE, 12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*(3), 328–351. https://doi.org/10.1086/269326

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology, 70*(3), 461–468.

Hevey, D., Pertl, M., Thomas, K., Maher, L., Craig, A., & Ní Chuinnéagain, S. (2010). Consideration of Future Consequences Scale: Confirmatory factor analysis. *Personality and Individual Differences, 48*(5), 654–657. https://doi.org/10.1016/j.paid.2010.01.006

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988.

Hinz, A., D. M., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social Medicine, 4*, 1–9.

Holleman, B. (1999). Wording effects in survey research: Using meta-analysis to explain the forbid/allow asymmetry. *Journal of Quantitative Linguistics, 6*(1), 29–40. https://doi.org/10.1076/jqul.6.1.29.4145

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10*(3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

Idaszak, J., & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology, 72*(1), 69–74.

John, T., Kulas, O., Rachael, K., & Lindsey, K. (2019). Confound it!: Social desirability and the "reverse-scoring" method effect. *European Journal of Psychological Assessment, 35*(6), 855–867. https://doi.org/10.1027/1015-5759/a000459

Kam, C. C. S. (2018). Why do we still have an impoverished understanding of the item wording effect? An empirical examination. *Sociological Methods & Research, 47*(3), 574–597. https://doi.org/10.1177/0049124115626177

Kam, C. C. S., & Fan, X. (2020). Investigating response heterogeneity in the context of positively and negatively worded items by using factor mixture modeling. *Organizational Research Methods, 23*(2), 322–341. https://doi.org/10.1177/1094428118790371

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes, 48*(5), 355–385. https://doi.org/10.1080/0163853X.2011.578910

Lee, P. H., Chang, L. I., & Ravens-Sieberer, U. (2008). Psychometric evaluation of the Taiwanese version of the Kiddo-KINDL generic children's health-related quality of life instrument. *Quality of Life Research, 17*(4), 603–611. https://doi.org/10.1007/s11136-008-9328-3

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment, 94*(2), 196–204. https://doi.org/10.1080/00223891.2011.645936

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1–55.

Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformational and transactional leadership: A meta-analytic review of the MLQ literature. *Leadership Quarterly, 7*(3), 385–425.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

MEB. (2022). *PISA 2022 OECD ülke raporu: Türkiye* (ISBN: 978-975-11-7448-2).

McLarnon, M. J., Goffin, R. D., Schneider, T. J., & Johnston, N. G. (2016). To be or not to be: Exploring the nature of positively and negatively keyed personality items in high-stakes testing. *Journal of Personality Assessment, 98*(5), 480–490. https://doi.org/10.1080/00223891.2016.1170691

Motl, R. W., Conroy, D. E., & Horan, P. M. (2000). The Social Physique Anxiety Scale: An example of the potential consequences of negatively worded items in factorial validity studies. *Journal of Applied Measurement, 1*(4), 327–345.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

OECD. (2024). *PISA 2022 Technical Report*. Paris: OECD Publishing. https://doi.org/10.1787/01820d6d-en

Paiva-Salisbury, M. L., Gill, A. D., & Stickle, T. R. (2017). Isolating Trait and Method Variance in the Measurement of Callous and Unemotional Traits. Assessment, 24(6), 763–771. https://doi.org/10.1177/1073191115624546

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(September), 381–391.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*(3), 603–610.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational citizenship behavior: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management, 26*(3), 513–563.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 31*(1), 113–130. https://doi.org/10.1080/02602930802618344

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101–1114. https://doi.org/10.1177/001316448104100420

Schwartz, S. J., Zamboanga, B. L., Wang, W., & Olthuis, J. V. (2009). Measuring identity from an Eriksonian perspective: Two sides of the same coin? *Journal of Personality Assessment, 91*(2), 143–154. https://doi.org/10.1080/00223890802634266

Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education, 6*(2), 259–278. https://doi.org/10.21449/ijate.515085

Wagner, J. A., III, & Gooding, R. Z. (1987). Shared influence and organizational behavior: A meta-analysis of situational variables expected to moderate participation–outcome relationships. *Academy of Management Journal, 30*(3), 524–541.

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness. *Sociological Methods & Research, 21*(1), 52–88. https://doi.org/10.1177/0049124192021001003

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education, 28*(6), 587–607. https://doi.org/10.1080/0260293032000130234

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*(5), 737–747.

Williams, L. J., Cote, J. A., & Buckley, M. R. (1989). Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology, 74*(3), 462–468. https://doi.org/10.1037/0021-9010.74.3.462

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67*(5), 555–561. https://doi.org/10.1037/0021-9010.67.5.555

Worcester, R. M., & Burns, T. R. (1975). Statistical examination of relative precision of verbal scales. *Journal of the Market Research Society, 17*, 181–197.

Xin, Z., & Chi, L. (2010). Wording effect leads to a controversy over the construct of the Social Dominance Orientation Scale. *The Journal of Psychology, 144*(5), 473–488. https://doi.org/10.1080/00223980.2010.496672

Ye, S. (2009). Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences, 46*(2), 197–201. https://doi.org/10.1016/j.paid.2008.09.027

Zeng, B., Wen, H., & Zhang, J. (2020). How does the valence of wording affect features of a scale? The method effects in the Undergraduate Learning Burnout Scale. *Frontiers in Psychology, 11*, 585179. https://doi.org/10.3389/fpsyg.2020.585179