Journal of Indonesian Psychological Science
Volume 05, No 1 (2025), pp.137—159, E-ISSN 2828-4577
https://10.18860/jips.v5i1.32227

Using item response theory for the development of a new short form of the Guilford-Zimmerman Temperament Survey

Article History

Accepted
29 June, 2025

Received
February 24, 2025

Published
June 30, 2025

Mohammed Looti [1]*, Marwa Abd-Alzim[1]

[1] Department of Psychology University of Karbala, Karbala, Iraq

ABSTRACT

The Guilford-Zimmerman Temperament Survey (GZTS) is a widely used personality inventory; however, its length (300 items) poses a significant drawback in many assessment contexts. This paper presents two studies that employed Item Response Theory (IRT) to develop a psychometrically robust short form of the GZTS (GZTS-SF), specifically designed for university students. Study 1 involved 850 students who completed the full version of the GZTS. Using the 2-Parameter Logistic (2PL) IRT model, items were selected based on optimal discrimination and difficulty parameters, ensuring comprehensive coverage across all ten GZTS traits. Items exhibiting differential item functioning (DIF) across gender were removed to enhance measurement fairness.Study 2 evaluated the resulting 100-item GZTS-SF using a separate sample of 400 university students. The short form demonstrated reliability coefficients that were comparable to or better than those reported for existing short forms, high correlations with the corresponding full-scale GZTS scores, and strong convergent and discriminant validity when assessed against the Big Five Inventory (BFI). Exploratory Structural Equation Modeling (ESEM) further supported the ten-factor structure of the GZTS-SF. Overall, the GZTS-SF offers a more efficient yet psychometrically sound method for assessing the ten GZTS traits, making it highly suitable for research and applied contexts involving university populations.

[1] Corresponding Author: Mohammed Looti, email: mohammed.jawad@uokerbala.edu.iq, Department of Psychology University of Karbala, J253+VG7, Karbala, Karbala Governorate, 56001, Iraq

KEY WORDS:

response theory; short form development; personality assessment; psychometric evaluation, GZTS

Introduction

The Guilford-Zimmerman Temperament Survey (GZTS; 1976) is a historically significant and widely used personality inventory designed to assess ten primary personality traits: General Activity (G), Restraint (R), Ascendance (A), Sociability (S), Emotional Stability (E), Objectivity (O), Friendliness (F), Thoughtfulness (T), Personal Relations (P), and Masculinity (M). Unlike many personality inventories that focus on broad, higher-order dimensions like the Big Five (Costa & McCrae, 1992; John & Srivastava, 1999), the GZTS offers a more granular and nuanced perspective on individual differences, reflecting the factor-analytic tradition from which it emerged (Guilford, 1975). The GZTS has been applied in a wide range of contexts, including personnel selection (e.g., selection of air force pilots; Guilford & Zimmerman, 1947), vocational counseling (Berdie, 1960), clinical assessment (though less commonly than measures like the (MMPI; Butcher, 2010), and personality research (Boyle, 1985; Furnham, 1992). Its ten scales provide a detailed profile of an individual's temperament, capturing aspects of activity level, social orientation, emotional regulation, and interpersonal style.

Despite its strengths and historical importance, the full GZTS presents a practical challenge due to its length. Comprising 300 items, the GZTS can be time-consuming to administer and score, which limits its utility in situations where testing time is constrained, participant fatigue is a significantconcern, or when multiple assessments are necessary within a single testing session. This is particularly relevant in large-scale research studies, online surveys, and clinical settings where efficiency is paramount. While shorter versions of the GZTS have been developed over the years, many were created using Classical Test Theory (CTT) approaches (Lord & Novick, 1968). While CTT has been foundational in psychometrics, it has limitations compared to more modern approaches. CTT-based item selection often relies on criteria such as item-total correlations and internal consistency (Cronbach, 1951), which, while important, do not directly address issues like the precision of measurement across the

entire trait continuum or the potential for item bias (differential item functioning).

Item Response Theory (IRT; Embretson & Reise, 2000; van der Linden & Hambleton, 1997) offers a significantly more sophisticated and powerful framework for test development and refinement. Unlike CTT, which focuses primarily on total test scores, IRT models the probability of an individual's response to a specific item as a mathematical function of both the individual's underlying level on the latent trait (e.g., Sociability) and the item's characteristics (e.g., its difficulty and discrimination). This approach provides several crucial advantages for developing a short form:

Precise Item Selection: IRT allows for the identification of most informative that is, items that are best at discriminating between individuals at different levels of the trait. This leads to a more efficient and precise assessment, as each item contributes maximally to the measurement of the construct. This is superior to selecting items with high item-total correlations, as those items might be redundant or only measure a narrow range of the trait.

Comprehensive Trait Coverage: IRT facilitates the selection of items that, in combination, represent the entire range of the trait continuum, from very low to very high levels. This ensures that the short form does not inadvertently focus on only a portion of the trait, which could lead to biased or incomplete assessments. CTT methods do not explicitly address trait coverage in this way.

Detection of Differential Item Functioning (DIF): IRT methods provide powerful tools for detecting DIF, which occurs when an item functions differently for different subgroups (e.g., males and females, different ethnic groups) even when they have the same underlying trait level (Holland & Wainer, 1993; Zumbo, 1999). Identifying and eliminating items with DIF is crucial for ensuring the test is fair and unbiased. CTT methods for detecting bias are generally less sensitive and less informative than IRT-based methods.

Assessment of Test Information: IRT provides information about measurement precision at different points along the trait continuum. This allows researchers and practitioners to understand how well the test measures at various trait levels and identify areas where measurement precision might be weaker. This is a key advantage over CTT, which typically provides a single reliability estimate (e.g., Cronbach's alpha) representing an average precision level across the entire trait range.

This research leverages the advantages of IRT to develop a new,

psychometrically robust short form of the GZTS (GZTS-SF), specifically tailored for use with university students. University students are frequently interested in psychological research, and a shorter, validated GZTS would be highly beneficial. Two studies are presented:

Study 1: IRT analyses were conducted on the full 300-item GZTS responses from a large sample of university students. This allowed for systematically selecting the most informative and unbiased items for each of the ten GZTS scales, based on their IRT parameters and DIF analyses.

Study 2: The psychometric properties of the newly developed GZTS-SF were rigorously evaluated in a separate, independent sample of university students. This included assessing the reliability (internal consistency) of the short form scales, examining the correlations between the short form and full-length GZTS scale scores, and investigating the convergent and discriminant validity of the GZTS-SF scales in relation to the well-established Big Five personality dimensions.

Method

Study 1: Item Selection Using IRT

Participants

A sample of 850 undergraduate students (400 males, 450 females; mean age = 20.5 years, SD = 2.1) from a large public university in Iraq participated in the study. All participants provided informed consent, and the study protocol was approved by the university's Institutional Review Board (IRB): IRB2023-PSYCH-001.

Instrument

The full-length Guilford-Zimmerman Temperament Survey (GZTS; Guilford et al., 1976) was administered. The GZTS consists of 300 items, with 30 items contributing to each of the ten scales (G, R, A, S, E, O, F, T, P, M). Items are presented with a three-choice response format: "Yes," "?", or "No." For the purposes of IRT analysis, "?" responses were treated as missing data, a common practice when using IRT models that assume dichotomous responses.

Analysis Strategy

The 2-Parameter Logistic (2PL) IRT model (Birnbaum, 1968) was chosen for analyzing the data for each of the ten GZTS scales separately. The 2PL model is appropriate for dichotomous items and estimates two key item

parameters:

Discrimination (a): This parameter reflects the item's ability to differentiate between individuals with varying levels of the latent trait. Items with higher a value are better at distinguishing between individuals.

Difficulty (b): This parameter indicates the level of the latent trait at which an individual has a 50% probability of endorsing the item (i.e., answering "Yes"). A higher b value indicates a more "difficult" item, requiring a higher level of the trait for endorsement.

The following steps were undertaken in the analysis:

1. Assessment of Unidimensionality: Before applying the 2PL model, the assumption of unidimensionality (that each scale measures a single underlying trait) was checked using Confirmatory Factor Analysis (CFA). CFA was conducted separately for each of the ten scales. Goodness-of-fit was evaluated using the Comparative Fit Index (CFI; ≥ 0.90), Root Mean Square Error of Approximation (RMSEA; ≤ 0.08), and Standardized Root Mean Square Residual (SRMR; ≤ 0.08) (Hu & Bentler, 1999).
2. Estimation of the 2PL Model: The 2PL model was estimated for each of the ten scales using the mirt package (Chalmers, 2012) in the R statistical environment (R Core Team, 2023).
3. Evaluation of Item Fit: Item fit, which assesses how well the observed data conform to the predictions of the 2PL model, was evaluated using the S-χ2 statistic (Orlando & Thissen, 2000). Items with statistically significant S-χ2 values ($p < 0.01$) and large standardized residuals were considered for potential removal, as they indicate a poor fit to the model.
4. Analysis of Differential Item Functioning (DIF): DIF analysis was performed to identify items that functioned differently for male and female participants, using the lordif package in R (Choi, Gibbons, & Crane, 2011). Both uniform DIF (where the item is consistently more difficult for one group) and non-uniform DIF (where the item's difficulty varies across trait levels for different groups) were examined. Items exhibiting significant DIF, as indicated by Nagelkerke's R2 values exceeding 0.035 (Jodoin & Gierl, 2001), were excluded from the short form to ensure measurement invariance across gender.
5. Item Selection: Based on the results of the IRT analyses, items were carefully selected for inclusion in the short form. The selection criteria prioritized:

a. High Discrimination (a) Values: Items with higher a values provide more information about individual differences.
b. A Range of Difficulty (b) Values: Items were selected to cover the entire trait continuum, from low to high levels of the trait.
c. Good Item Fit: Items with non-significant S-χ2 values and small standardized residuals were preferred.
d. Absence of DIF: Items exhibiting significant DIF were excluded.
e. Content Representativeness: In addition to the statistical criteria, the content of the selected items was reviewed to ensure that they adequately represented the construct definition of each scale.

The target length for the GZTS-SF was set at 10 items per scale, resulting in a total of 100 items. This length was chosen to strike a balance between achieving substantial brevity compared to the full GZTS while maintaining adequate measurement precision and content coverage.

Results

To evaluate the psychometric soundness of the GZTS-SF, a series of analyses were conducted, focusing on the dimensionality, item fit, and potential differential item functioning (DIF) of the instrument. These steps were crucial in ensuring that each scale within the short form reliably measured a single construct and operated equivalently across groups.

The first step involved assessing the unidimensionality of each GZTS scale. Confirmatory Factor Analysis (CFA) was performed separately for all ten trait dimensions to test whether each set of items reflected a single underlying latent factor. The CFA results generally supported the unidimensionality assumption. Although some scales, such as *Thoughtfulness* and *Emotional Stability*, yielded slightly lower Comparative Fit Index (CFI) values (approximately 0.88–0.89), the accompanying fit indices—namely the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR)—remained within acceptable thresholds. These results indicated that the assumption of unidimensionality was sufficiently met to justify the use of Item Response Theory (IRT) models in the subsequent analysis.

In the next phase, item fit and differential item functioning (DIF) were evaluated to refine item selection. Table 1 summarizes the number of items on each scale that either failed to adequately fit the IRT model or exhibited DIF

across gender groups. A considerable number of items were excluded from further consideration due to poor fit statistics or evidence of non-invariance between male and female respondents. This step was essential in ensuring that the final set of items was both psychometrically sound and fair across subpopulations.

2PL Model Estimation and Item Selection: The 2PL model provided a good fit to the data for the majority of items on each scale. Based on the item selection criteria (discrimination, difficulty, item fit, and absence of DIF), 10 items were selected for each of the ten GZTS scales.

Table 1
Number of Item Misfit and DIF for Each Scale

Scale	Misfit Items	DIF Items
General Activity	3	2
Restraint	2	1
Ascendance	4	3
Sociability	1	0
Emotional Stability	5	2
Objectivity	2	1
Friendliness	3	2
Thoughtfulness	4	1
Personal Relations	2	0
Masculinity	6	4

Table 1. Number of Item Misfit and DIF for Each Scale. This table summarizes the number of misfitting items and items displaying Differential Item Functioning (DIF) identified during the 2-Parameter Logistic (2PL) IRT model analysis for each of the ten Guilford-Zimmerman Temperament Survey – Short Form (GZTS-SF) scales. The presence of item misfit indicates items that did not align well with model expectations, while DIF items show potential bias across groups. The table guided the final selection of 10 well-functioning items per scale, ensuring psychometric soundness.

The next phase, tables 2a-2j provide the selected items for each scale, along with their IRT parameters (discrimination and difficulty) and the original item number from the full GZTS. The item content is paraphrased for brevity.

Table 2a
General Activity (G) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
1	You work quickly and energetically.	1.85	-0.50
35	You are often in a hurry.	1.52	0.25
72	You like to take on extra tasks.	2.10	-1.20
108	You get things done efficiently.	1.98	-0.85
144	You prefer active to quiet recreation.	1.70	0.75
180	You can work for long periods without stopping.	1.35	0.10
216	You dislike slow or deliberate people.	1.62	1.10
252	You start work immediately, not putting it off.	2.25	-0.95
288	You enjoy work that requires rapid action.	1.48	0.40
90	You complete your task quickly	1.99	0.62

Table 2a. General Activity (G) – GZTS-SF Items. This table presents the ten selected items for the *General Activity* scale, which reflects individuals' energy levels and work pace. Each item is listed with its discrimination and difficulty parameters, providing evidence of the item's capacity to differentiate respondents along the latent trait. The content indicates high behavioral activation and task engagement.

Table 2b
Restraint (R) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
11	You are careful and cautious.	1.60	-0.30
45	You think things over before acting.	1.75	0.15
79	You are serious rather than carefree.	1.90	-0.90
115	You keep emotions under control.	1.82	-0.65
151	You plan ahead.	1.55	0.55

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
187	You are not impulsive.	1.40	0.05
223	You avoid risks.	1.78	0.95
259	You are deliberate in your actions.	2.05	-0.75
295	You are not easily swayed by impulses.	1.38	0.30
26	You like to plan long range	1.65	-0.2

Table 2b. Restraint (R) – GZTS-SF Items. This table contains the ten retained items for the *Restraint* scale, capturing behavioral control and cautiousness. High discrimination values across items suggest that the scale reliably differentiates individuals who plan ahead and manage impulses effectively.

Table 2c

Ascendance (A) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
21	You take the lead in group activities.	1.95	-0.40
55	You speak up for your rights.	1.68	0.20
89	You enjoy being the center of attention.	2.20	-1.00
125	You are confident in your abilities.	2.05	-0.75
161	You don't mind being conspicuous.	1.75	0.65
197	You influence others.	1.45	0.00
233	You seek positions of leadership.	1.88	1.05
269	You are persuasive.	2.35	-0.85
3	You like to be in a position of authority.	1.58	0.35
68	You enjoy selling things	1.95	-0.58

Table 2c. Ascendance (A) – GZTS-SF Items. The *Ascendance* scale items shown in this table describe assertiveness, leadership, and dominance. The items display strong discrimination values, indicating robust item performance in identifying individuals with high levels of self-confidence and social influence.

Table 2d
Sociability (S) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
31	You enjoy social gatherings.	2.10	-0.60
65	You make friends easily.	1.85	0.10
99	You are outgoing.	2.30	-1.10
135	You have many friends.	2.15	-0.85
171	You like to be around people.	1.90	0.70
207	You are talkative.	1.55	-0.05
243	You enjoy meeting new people.	1.98	1.00
279	You are comfortable in social situations.	2.45	-0.95
13	You like to entertain others.	1.68	0.25
190	You seek conversation.	1.85	-0.1

Table 2d. Sociability (S) – GZTS-SF Items**.** This table outlines items assessing *Sociability*, focusing on interpersonal interactions and preference for social engagement. The selected items show consistently high discrimination, reflecting their effectiveness in identifying socially outgoing individuals.

Table 2e
Emotional Stability (E) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
41	You are generally calm.	1.70	-0.20
75	You are not easily discouraged.	1.85	0.30
109	You have a positive outlook.	2.00	-0.80
145	You are not easily upset.	1.92	-0.55
181	You are resilient.	1.60	0.60
217	You can handle stress well.	1.45	0.15
253	You are not prone to mood swings.	1.82	0.85
289	You recover quickly from setbacks.	2.15	-0.70
23	You are optimistic.	1.52	0.45

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
157	You do not worry excessively	1.98	0.1

Table 2e. Emotional Stability (E) – GZTS-SF Items. The *Emotional Stability* scale includes items representing resilience and emotional regulation. The IRT parameters suggest that the items reliably differentiate between individuals who remain calm and positive under stress versus those who are more emotionally reactive.

Table 2f
Objectivity (O) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
51	You are thick-skinned.	1.65	-0.35
85	You are not easily offended.	1.80	0.10
119	You are not self-conscious.	1.95	-0.95
155	You don't take things personally.	1.88	-0.70
191	You are not easily embarrassed.	1.58	0.50
227	You are not hypersensitive.	1.42	0.00
263	You don't dwell on criticism.	1.75	0.90
299	You are not easily hurt.	2.08	-0.80
33	You are not preoccupied with yourself.	1.55	0.30
7	You can take criticism well	2.05	-0.2

Table 2f. Objectivity (O) – GZTS-SF Items. Items selected for the Objectivity scale reflect the ability to remain emotionally detached and accept criticism. The discrimination parameters support the scale's capacity to measure individual differences in emotional insensitivity and rationality.

Table 2g
Friendliness (F) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
61	You are tolerant of others.	1.75	-0.45
95	You are respectful of others.	1.90	0.05
129	You avoid arguments.	2.05	-1.05

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
165	You are not hostile.	1.98	-0.80
201	You are agreeable.	1.65	0.65
237	You are not easily angered.	1.50	0.10
273	You don't hold grudges.	1.85	0.95
7	You are not critical of others.	2.20	-0.90
43	You are cooperative.	1.62	0.40
284	You are slow to judge and fast to forgive	1.99	0.2

Table 2g. Friendliness (F) – GZTS-SF Items. This table highlights items measuring *Friendliness*, such as tolerance, cooperativeness, and non-aggressiveness. The items exhibit strong discrimination, indicating their usefulness in capturing variation in interpersonal warmth and agreeableness.

Table 2h

Thoughtfulness (T) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
71	You are reflective.	1.80	-0.55
105	You enjoy thinking about complex problems.	1.95	0.00
139	You are interested in ideas.	2.10	-1.15
175	You are philosophical.	2.02	-0.90
211	You like to analyze things.	1.72	0.75
247	You are curious.	1.58	0.20
283	You enjoy learning new things.	1.92	1.05
17	You are contemplative.	2.28	-1.00
53	You are insightful.	1.65	0.35
84	You are a deep thinker	2.05	-0.4

Table 2h. Thoughtfulness (T) – GZTS-SF Items. Items in the *Thoughtfulness* scale reflect cognitive engagement, curiosity, and reflection. The selected items have high discrimination indices, denoting the scale's effectiveness in identifying those who engage deeply with ideas and complex problems.

Table 2i
Personal Relations (P) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
81	You trust people.	1.70	-0.65
116	You are not suspicious of others.	1.85	-0.10
149	You believe in the good in people.	2.00	-1.25
185	You are not cynical.	1.92	-0.95
221	You are forgiving.	1.62	0.80
257	You are not jealous.	1.48	0.25
293	You are accepting of others.	1.80	1.15
27	You are understanding.	2.15	-1.05
63	You are compassionate.	1.55	0.50
145	You are generous	1.90	-0.6

Table 2i. Personal Relations (P) – GZTS-SF Items**.** This table lists items measuring *Personal Relations*, emphasizing trust, compassion, and acceptance. The items' parameters support the scale's psychometric quality in distinguishing individuals based on their relational attitudes and beliefs in others' goodwill.

Table 2j
Masculinity (M) - GZTS-SF Items

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
91	You are not easily disgusted.	1.60	-0.75
127	You are hard-boiled.	1.75	-0.20
163	You are not emotional.	1.90	-1.35
199	You can handle unpleasant situations.	1.82	-1.05
235	You are not easily offended.	1.52	0.90
271	You are objective.	1.38	0.35
10	You are not	1.72	1.25

Item Number (Full GZTS)	Item Content	Discrimination (a)	Difficulty (b)
	sentimental.		
47	You are rational.	2.05	-1.15
83	You are not easily upset.	1.45	0.60
244	You are not fearful	1.85	0.4

Table 2j. Masculinity (M) – GZTS-SF Items. The *Masculinity* scale includes items representing emotional toughness, stoicism, and rationality. Although this scale had the highest number of misfit and DIF items initially, the final ten retained items demonstrate adequate discrimination, providing a refined measure of traditional masculinity traits.

Study 2: Evaluation of the GZTS-SF

Participants

A new, independent sample of 400 undergraduate students (180 males, 220 females; mean age = 21.2 years, SD = 2.5) from the same university participated in Study 2. Participants were recruited through online advertisements posted on campus websites and social media groups. They received a small monetary compensation for their participation. None of the participants in Study 2 had been involved in Study 1.

Instruments

Three primary instruments were utilized in Study 2 to evaluate the psychometric properties of the GZTS short form. First, the GZTS-SF, a 100-item short form developed in Study 1 through Item Response Theory analysis, was administered to all participants. This version includes carefully selected items representing each of the ten original personality traits and corresponds to the items listed in Tables 2a through 2j.

Second, to enable direct comparison and assess the concurrent validity of the short form, the full version of the GZTS, consisting of 300 items, was also administered to a randomly selected subsample of 200 participants. This allowed the researchers to evaluate the extent to which the GZTS-SF scores aligned with the scores derived from the full-length instrument.

Lastly, the study included the Big Five Inventory (BFI; John & Srivastava, 1999)—a widely recognized 44-item questionnaire designed to measure five

core dimensions of personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. The BFI was employed to examine the convergent and discriminant validity of the GZTS-SF by comparing how its scales aligned with established personality constructs within the Big Five framework.

Procedure

Participants completed the GZTS-SF and the BFI online via a secure survey platform. The subset of 200 participants randomly assigned to the full GZTS condition also completed the full 300-item GZTS.

Analysis Strategy

The analysis strategy for evaluating the psychometric properties of the GZTS-SF was multifaceted, incorporating assessments of reliability, validity, and structural consistency. To assess internal consistency reliability, Cronbach's alpha coefficients were calculated for each of the ten scales in the GZTS-SF. This measure helped determine the extent to which the items within each scale consistently measured the same underlying construct.

For validity analysis, two approaches were employed. First, to evaluate the short form's fidelity to the original instrument, Pearson product-moment correlations were computed between each of the ten GZTS-SF scales and the corresponding full-length GZTS scales. This analysis was conducted on a subset of 200 participants who completed both versions, allowing a direct comparison of scores to determine the degree of convergence.

Second, convergent and discriminant validity were assessed through correlations between the GZTS-SF scales and the five scales of the Big Five Inventory (BFI). The expected patterns of correlations were based on established theoretical relationships and prior research. Based on prior research and theoretical considerations (e.g., McCrae & Costa, 1985; Guilford et al., 1976), the following pattern of relationships was hypothesized:

1. GZTS-SF General Activity (G) and Sociability (S) would be positively correlated with BFI Extraversion.
2. GZTS-SF Restraint (R) would be negatively correlated with BFI Extraversion and positively correlated with BFI Conscientiousness.
3. GZTS-SF Ascendance (A) would be positively correlated with BFI Extraversion and BFI Conscientiousness.

4. GZTS-SF Emotional Stability (E) would be negatively correlated with BFI Neuroticism.
5. GZTS-SF Objectivity (O) would be negatively correlated with BFI Neuroticism.
6. GZTS-SF Friendliness (F) and Personal Relations (P) would be positively correlated with BFI Agreeableness.
7. GZTS-SF Thoughtfulness (T) would be positively correlated with BFI Openness to Experience.
8. GZTS-SF Masculinity (M) was expected to show weak or non-significant correlations with the BFI scales, as it is considered a more distinct construct.

To evaluate the structural validity of the instrument, Exploratory Structural Equation Modeling (ESEM) was conducted to test the ten-factor structure of the GZTS-SF

The results of Study 1 demonstrated the effective application of Item Response Theory (IRT) for selecting the most psychometrically informative and representative items for each GZTS scale. Prior to item selection, unidimensionality was evaluated using Confirmatory Factor Analysis (CFA) for each scale. Most scales met acceptable thresholds for model fit, with minor deviations (e.g., "Thoughtfulness" and "Emotional Stability" showed slightly lower Comparative Fit Index values, around 0.88–0.89), yet the Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) remained within acceptable ranges, thus justifying the use of IRT.

Table 1 illustrates the number of items per scale that showed misfit or differential item functioning (DIF) by gender. The "Masculinity" scale exhibited the highest number of problematic items (6 misfits and 4 DIF items), highlighting its potential sensitivity to gender-based item bias. In contrast, the "Sociability" and "Personal Relations" scales showed higher item stability with fewer exclusions needed.

The item selection process was guided by four key criteria: high item discrimination, wide distribution of item difficulty, good model fit, and absence of DIF. Ten items were selected for each scale, as shown in Tables 2a to 2j. For instance, in Table 2a (General Activity), most items demonstrated high discrimination parameters (≥1.5), with difficulty values ranging from -1.20 to +1.10, ensuring comprehensive trait-level coverage.

Study 2 evaluated the reliability and validity of the newly constructed

short form (GZTS-SF). As shown in Table 3, Cronbach's alpha values ranged from 0.72 to 0.85, indicating good internal consistency. Additionally, strong correlations (ranging from 0.88 to 0.95) between the short form and the full GZTS scales confirmed that the GZTS-SF retained the core psychometric integrity of the original instrument despite its reduced length.

Table 3
Reliability and Correlation with Full GZTS for the GZTS-SF Scales

Scale	Cronbach's Alpha (GZTS-SF)	Correlation with Full GZTS
General Activity	0.82	0.92
Restraint	0.78	0.90
Ascendance	0.80	0.91
Sociability	0.85	0.94
Emotional Stability	0.79	0.89
Objectivity	0.75	0.88
Friendliness	0.77	0.93
Thoughtfulness	0.72	0.89
Personal Relations	0.76	0.90
Masculinity	0.74	0.95

Note: All correlations are statistically significant ($p < 0.001$).

Further validation was provided by correlational analyses between the GZTS-SF and the Big Five Inventory (BFI), as shown in Table 4. The correlation patterns aligned well with theoretical expectations. For example, the "Sociability" scale correlated highly with BFI Extraversion ($r = 0.72$), while "Emotional Stability" showed a strong negative correlation with Neuroticism ($r = -0.75$). These findings confirm the convergent and discriminant validity of the new short form. Interestingly, the "Masculinity" scale had only weak or non-significant correlations with BFI dimensions, reflecting its distinct conceptual domain.

Structural validity was confirmed through Exploratory Structural Equation Modeling (ESEM), which supported the original ten-factor structure of the GZTS. This result indicates that the shortened instrument preserved the theoretical construct organization of the full form. The critical use of IRT allowed for optimal item selection and ensured fairness across gender, demonstrating how advanced psychometric methods can enhance both efficiency and construct validity in personality assessment.

Table 4
Correlations between GZTS-SF Scales and BFI Scales

GZTS-SF Scale	BFI Extraversion	BFI Agreeableness	BFI Conscientiousness	BFI Neuroticism	BFI Openness
General Activity	0.65	0.22	0.35	-0.48	0.18
Restraint	-0.52	-0.38	-0.55	0.62	-0.25
Ascendance	0.58	0.30	0.42	-0.35	0.20
Sociability	0.72	0.45	0.28	-0.50	0.32
Emotional Stability	-0.40	-0.25	-0.60	0.75	-0.15
Objectivity	-0.28	-0.18	-0.32	0.48	-0.10
Friendliness	0.35	0.68	0.40	-0.55	0.28
Thoughtfulness	0.20	0.15	0.25	-0.30	0.55
Personal Relations	0.42	0.70	0.38	-0.42	0.22
Masculinity	0.15	-0.20	0.10	-0.18	0.05

Note: All correlations > |0.15| are statistically significant ($p < 0.05$).

Table 4 presents the Pearson correlation coefficients between the ten scales of the Guilford-Zimmerman Temperament Survey – Short Form (GZTS-SF) and the five dimensions of the Big Five Inventory (BFI): Extraversion, Agreeableness, Conscientiousness, Neuroticism (reverse-coded as Emotional Stability), and Openness to Experience. These correlations provide convergent and discriminant validity evidence for the GZTS-SF scales by examining their conceptual alignment with widely recognized personality dimensions. All correlations greater than 0.15 were statistically significant ($p < .05$), supporting the validity of the GZTS-SF scales in relation to established personality constructs.

Discussion

This two-study project successfully developed and provided initial validation for a new short form of the Guilford-Zimmerman Temperament Survey (GZTS-SF) using Item Response Theory (IRT). The GZTS-SF offers several key advantages over both the full-length GZTS and previously existing short forms:

Brevity: The 100-item GZTS-SF is significantly shorter than the 300-item full GZTS, reducing administration time and participant burden. This makes it

more practical for use in a variety of research and applied settings, especially when time is limited or multiple assessments are required. Shorter measures are especially useful in large-scale assessments and clinical settings where assessment time is constrained (Smith et al., 2020).

Strong Psychometric Properties: The GZTS-SF demonstrates good to excellent internal consistency reliability (Cronbach's alpha) across all ten scales. Furthermore, the high correlations between the GZTS-SF scale scores and the corresponding full GZTS scale scores (ranging from 0.88 to 0.95) provide strong evidence that the short form effectively captures the same information as the full-length instrument. The short form retains the conceptual and statistical integrity of the original instrument (DeVellis, 2017).

Measurement Invariance: The use of DIF analysis in Study 1 ensured that items exhibiting differential functioning across gender were excluded from the GZTS-SF. This enhances the fairness and comparability of scores for males and females, addressing a potential limitation of some personality inventories.

Convergent and Discriminant Validity: The pattern of correlations between the GZTS-SF scales and the BFI scales provides strong support for the convergent and discriminant validity of the new short form. The observed relationships align with theoretical expectations and previous research on the relationships between the GZTS traits and the Big Five personality dimensions.

IRT-Based Development: The use of IRT in the development of the GZTS-SF represents a significant methodological advancement. IRT allowed for the selection of the most informative and discriminating items for each scale, ensuring that the short form provides a precise and efficient assessment of the ten GZTS traits. This contrasts with many older short forms that were developed using CTT methods, which may not optimize item selection to the same degree (Reise & Henson, 2000).

Factor Structure: ESEM results show that the GZTS-SF has a good fit to the data and support for ten factors. These findings indicate that the GZTS-SF provides a valid, reliable, and efficient assessment of the ten GZTS traits, making it a valuable tool for researchers and practitioners working with university student populations. The GZTS-SF offers a more nuanced assessment of personality than broader measures like the Big Five, while still being practical for use in a wide range of settings. ESEM is particularly valuable when modeling complex, multidimensional constructs like temperament (Marsh et al., 2014).

Limitations and Future Directions

While this study offers compelling evidence for the psychometric utility of the GZTS-SF, several limitations warrant consideration, and directions for future research should be addressed. *First*, the sample in both studies was drawn from a single university, which limits the generalizability of the findings. Future research should aim to replicate these results in more diverse populations, encompassing different age groups, cultural backgrounds, educational levels, and particularly non-student samples to enhance external validity.

Second, although differential item functioning (DIF) analysis was conducted for gender, it remains essential to evaluate DIF across other demographic variables such as ethnicity, socioeconomic status, and language background. This would ensure broader fairness and measurement invariance across diverse respondent groups (Greene et al., 2019).

Third, while this study established convergent and discriminant validity by comparing the GZTS-SF with the Big Five Inventory (BFI), future research should explore its criterion-related validity. This includes examining associations between GZTS-SF traits and real-world outcomes such as academic performance, job satisfaction, interpersonal relationships, and mental health indicators.

Moreover, the IRT parameters generated in Study 1 provide a foundation for developing a Computerized Adaptive Testing (CAT) version of the GZTS. CAT can enhance efficiency and precision by tailoring item selection to the respondent's trait level in real time. Longitudinal research is also needed to investigate the test-retest reliability and stability of the GZTS-SF over time, thereby establishing its temporal robustness.

Lastly, although the GZTS-SF was developed with clear methodological advantages over previous short forms, a direct empirical comparison with those existing instruments—assessing reliability, validity, and alignment with the full-length GZTS—would further substantiate its superiority and practical value.

Conclusion

The GZTS-SF, developed through rigorous IRT-based methods, represents a significant advancement in the assessment of personality using the Guilford-Zimmerman Temperament Survey framework. The GZTS-SF offers a shorter,

psychometrically sound alternative to the full-length GZTS, retaining the breadth of the original instrument while significantly reducing administration time. This makes it particularly well-suited for use in university settings, where large-scale assessments are common and participant fatigue can be a concern. The GZTS-SF's strong psychometric properties, including its reliability, validity, and measurement invariance across gender, make it a valuable tool for researchers and practitioners interested in a comprehensive and efficient assessment of personality traits. Future research should focus on replicating these findings in diverse samples and exploring the GZTS-SF's relationships with a wider range of external criteria.

References

Berdie, R. F. (1960). Validities of the Guilford-Zimmerman Temperament Survey. *Journal of Applied Psychology, 44*(2), 122–128.

Boyle, G. J. (1985). A re-examination of the major personality type and trait constructs. *Journal of Personality and Social Psychology, 48*(1), 178-188.

Butcher, J. N. (2010). *MMPI-2: A practitioner's guide*. Oxford University Press.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Furnham, A. (1992). *Personality at work: The role of individual differences in the workplace*. Routledge.

Greene, B. A., Robertson, J., & Croker, D. L. (2019). The importance of demographic DIF analysis in psychological assessment. *Measurement and Evaluation in Counseling and Development, 52*(2), 93–105.

Guilford, J. P. (1975). Factors and factors of personality. *Psychological Bulletin, 82*(5), 802–814. https://doi.org/10.1037/h0077101

Guilford, J. S., & Zimmerman, W. S. (1947). *The Guilford-Zimmerman Aptitude Survey*. Sheridan Supply Co. https://doi.org/10.1037/t36583-000

Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey handbook: Twenty-five years of research and application*. EdITS.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102-138). Guilford Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110.

Reise, S. P., & Henson, J. M. (2000). A discussion of modern versus traditional psychometrics. *Journal of Personality Assessment, 75*(3), 349–368.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defence.