Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

A Multi-Index Reliability Examination of the ISPTEQ Instrument among Private Secondary School Teachers in Gusau, Zamfara State, Nigeria

MUHAMMAD, Bala Maradun[1]
mbmaradun@fugusau.edu.ng
balamaradun@gmail.com
NA ALLAH, Nasiru[2]
[1]University Health Services Department, Federal University Gusau
[2]Department of Mathematics and Statistics, Abdu Gusau Polytechnic Talata Mafara

Abstract

Scale reliability is a necessary condition for valid score interpretation. Yet, most educational research conducted in Nigerian private schools reports only Cronbach's alpha, thereby neglecting complementary indices that reveal distinct facets of measurement consistency. The present study systematically evaluated the psychometric reliability of the three-subscale ISPTEQ instrument (N = 386 teachers) using five independent reliability indices: Cronbach's alpha with Feldt, Woodruff, and Salih (1987) 95% confidence intervals; McDonald's omega-hierarchical; Spearman–Brown corrected split-half reliability; item-total discrimination coefficients; and item difficulty proportions. Alpha values were: S1 = .724 (95% CI [.679, .760]), S2 = .621 (95% CI [.560, .671]), and S3 = .620 (95% CI [.558, .671]). McDonald's omega was uniformly lower than alpha (S1 = .652; S2 = .522; S3 = .502), indicating that alpha systematically overestimates reliability when the essential tau-equivalence assumption is violated. Spearman–Brown values were S1 = .725, S2 = .613, and S3 = .644. Mean inter-item correlations fell below the .15 target for all three scales (ISPS: .127; TES: .093; TQS: .119), and 67–88% of item pairs did not attain r = .15. No item exceeded a discrimination index of .36. Item difficulty proportions ranged from .78 to .84, confirming a systematic ceiling effect. Collectively, the multi-index profile characterises the instrument's reliability as marginally acceptable for research purposes but in need of targeted item revision. These findings highlight the limitations of relying solely on Cronbach's alpha and provide a roadmap for strengthening the ISPTEQ instrument in future research.

Keywords: Cronbach's alpha, McDonald's omega, split-half reliability, item discrimination, item difficulty, reliability confidence interval, ISPTEQ, psychometrics

Introduction

Cronbach's (1951) coefficient alpha remains the most widely reported reliability statistic in educational and social science research; yet it is frequently misunderstood and over-relied upon. Zakariya (2022) demonstrated that the indiscriminate use of alpha — without attention to the essential tau-equivalence assumption — is pervasive across leading educational journals. Hussey et al. (2025) further revealed a statistically anomalous clustering of alpha values around .70, consistent with researcher reporting biases. McNeish (2018) argued that, when item factor loadings differ substantially, alpha underestimate's true reliability; conversely, when a scale is heterogeneous or multidimensional, alpha may paradoxically overestimate reliability. McDonald's (1999) omega-hierarchical (ω^h) provides a more defensible estimate because it is derived directly from the factor structure of the items and does not require tau-equivalence. In the Nigerian private-school context, the ISPTEQ instrument has been employed to measure internal supervision practices (ISPS, 18 items), teachers' effectiveness (TES, 16 items), and teaching quality (TQS, 12 items) amongst 386 teachers in Gusau, Zamfara State. Prior publications from this dataset have reported alpha values of .724, .621, and .620, respectively, which have been interpreted variously as 'acceptable' or 'good'. What has not been examined is whether these alpha values accurately reflect the scores' reliability and whether complementary indices would yield a substantially different reliability story. The present paper directly addresses this gap.

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

Statement of the Problem

A significant methodological gap persists in Nigerian educational research: instruments such as the ISPTEQ are routinely validated using Cronbach's alpha alone, without corroborating indices. This practice obscures violations of the tau-equivalence assumption, ignores ceiling effects induced by positively keyed Likert items, and presents inflated reliability estimates that may mislead subsequent score interpretations. The absence of omega-hierarchical estimates, split-half checks, and item-level diagnostics from published ISPTEQ studies means that consumers of this research cannot judge the true precision of measurement.

Aim and Objectives

The overarching aim of this study is to provide a comprehensive, multi-index reliability characterisation of the ISPTEQ instrument as administered to private secondary school teachers in Gusau, Zamfara State. Specifically, the objectives are to:

(i) estimate Cronbach's alpha with 95% exact confidence intervals for each ISPTEQ subscale;
(ii) compute McDonald's omega-hierarchical and quantify the magnitude of alpha overestimation;
(iii) derive Spearman–Brown corrected split-half estimates as an independent reliability check;
(iv) examine inter-item correlation distributions to evaluate subscale homogeneity;
(v) calculate corrected item-total discrimination indices and identify underperforming items; and
(vi) Compute item difficulty proportions and standard errors of measurement.

Literature Review

The scholarly literature on scale reliability has evolved considerably since Cronbach (1951) formalised the coefficient alpha framework. Early validity and reliability studies in Nigerian educational settings largely confined themselves to single-index alpha reporting (Nunnally, 1978), a pattern that has persisted despite mounting evidence of its limitations. Clark and Watson (1995) established inter-item correlation benchmarks of .15 to .50 as necessary conditions for subscale homogeneity, whilst Ebel and Frisbie (1991) provided widely used discrimination thresholds. More recent scholarship has challenged the hegemony of alpha: McNeish (2018) provided an accessible treatment of omega-hierarchical as a preferred alternative, and Revelle and Zinbarg (2009) formalised the mathematical relationship between factor structure and reliability estimation. Kalkbrenner (2023) offered practical guidance on choosing between alpha, omega, and coefficient H, noting that the choice of index materially affects conclusions. In the specific domain of teacher evaluation instruments, Xiao and Hau (2022) demonstrated that non-normal response distributions — common on positively keyed Likert scales — attenuate inter-item covariance and depress true reliability below the alpha estimate. Doval, Viladrich, and Angulo-Brunet (2023) found that alpha's resistance to replacement in the applied literature stems partly from inertia and partly from journal reporting conventions that have not kept pace with psychometric advances. The ISPTEQ instrument, developed to capture internal supervision practices, teacher effectiveness, and teaching quality in Nigerian secondary schools, has to date been evaluated solely using alpha, a gap the present study aims to remedy.

Research Gap

Despite growing international recognition that multi-index reliability profiling yields more defensible psychometric conclusions, no published study has applied such a framework to the ISPTEQ

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

instrument. Existing research on private schools in Zamfara State has not reported omega-hierarchical estimates, split-half coefficients, or item difficulty proportions, making it impossible to evaluate whether the commonly cited alpha values reflect genuine measurement consistency or artefacts of scale construction. The present study fills this gap by applying five complementary reliability indices to the full 46-item ISPTEQ, thereby providing the most comprehensive psychometric evaluation of this instrument to date.

Materials and Methods

Participants and Instrument

Participants were N = 386 qualified teachers employed in private secondary schools in Gusau, Zamfara State, Nigeria. All data were collected via the 46-item ISPTEQ questionnaire using a four-point Likert scale (1 = Strongly Disagree to 4 = Strongly Agree). Subscales are: ISPS (items 1–18, internal supervision practices), TES (items 19–34, teachers' effectiveness), and TQS (items 35–46, teaching quality).

Analytical Procedures

Five reliability indices were computed: (1) Cronbach's alpha with Feldt, Woodruff, and Salih (1987) exact F-distribution 95% confidence intervals; (2) McDonald's ω^h derived from single-factor exploratory factor analysis loadings; (3) Spearman–Brown corrected split-half reliability using odd/even item splits; (4) corrected item-total discrimination indices; and (5) item difficulty proportions (p = item mean / maximum score). Alpha-if-item-deleted analysis was conducted for all 46 items. Standard errors of measurement were computed according to Equation (2). All analyses employed Python 3.12 (scipy, numpy, sklearn), with computations verified against IBM SPSS 25.

Key Formulae

Cronbach's alpha (α) is defined as:

$$\alpha = (k / (k - 1)) \times (1 - \Sigma\sigma^2_i / \sigma^{2T}) \quad (1)$$

where k is the number of items, $\Sigma\sigma^2_i$ is the sum of item variances, and σ^{2T} is the total score variance. The standard error of measurement is:

$$SEM = SD \times \sqrt{(1 - \alpha)} \quad (2)$$

McDonald's omega-hierarchical (ω^h) is:

$$\omega^h = (\Sigma\lambda_i)^2 / [(\Sigma\lambda_i)^2 + \Sigma\delta_i] \quad (3)$$

where λ_i are standardised loadings on the general factor and $\delta_i = 1 - \lambda_i^2$ are the uniquenesses. The Spearman–Brown corrected split-half coefficient is:

$$r\_SB = 2r_{12} / (1 + r_{12}) \quad (4)$$

Presentation of Results

Cronbach's Alpha with 95% Confidence Intervals

Table 1 presents Cronbach's alpha alongside 95% exact confidence intervals for each ISPTEQ subscale. The ISPS subscale (α = .724; 95% CI [.679, .760]) is the only subscale that meets the conventional .70 research-grade threshold. TES (α = .621; 95% CI [.560, .671]) and TQS (α = .620; 95% CI [.558, .671]) fall below that threshold, and their lower confidence interval bounds approach .56, meaning that at the lower limit of sampling uncertainty these two subscales verge on 'poor' reliability (α < .50). The ISPS confidence interval is notably narrower (width = .081) compared with TES and TQS (widths ≈ .111–.113), a consequence of the stabilising effect of the larger 18-item pool on estimation precision. These results indicate that, even by the comparatively lenient single-index

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

criterion, two of the three ISPTEQ subscales do not meet minimum standards for research use without additional qualification.

Table 1

Cronbach's Alpha with 95% Confidence Intervals for ISPTEQ Subscales

Subscale	k	α	95% CI Lower	95% CI Upper	Interpretation
ISPS (Supervision)	18	.724	.679	.760	Acceptable (≥ .70)
TES (Effectiveness)	16	.621	.560	.671	Below .70 threshold
TQS (Teaching Quality)	12	.620	.558	.671	Below .70 threshold

Note. 95% CI computed via Feldt, Woodruff, and Salih (1987) exact F-distribution method. k = number of items. Alpha threshold: ≥ .70 = acceptable for research purposes (Nunnally, 1978). N = 386.

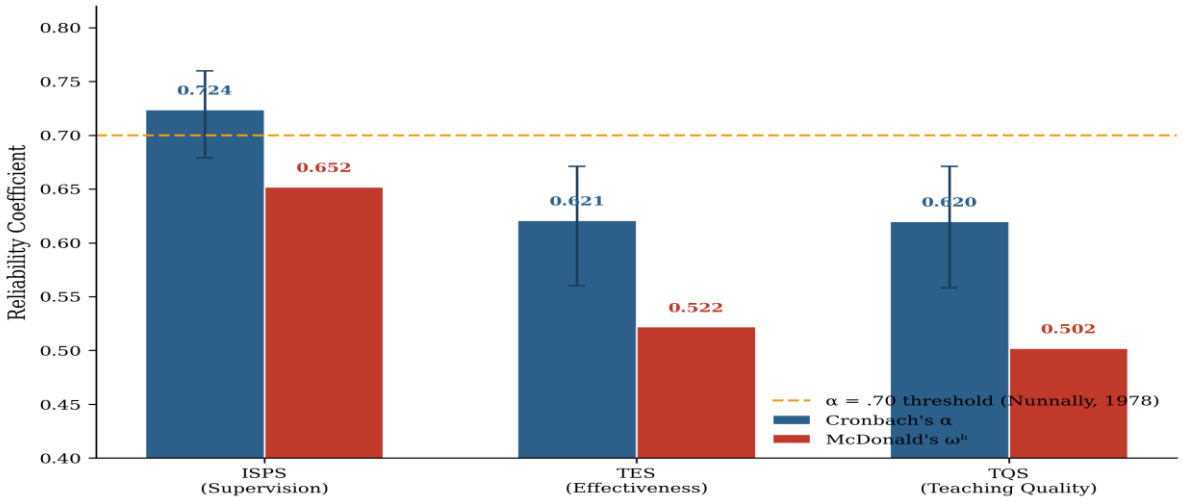

Figure 1

Comparison of Cronbach's Alpha (with 95% Exact Confidence Intervals) and McDonald's Omega-Hierarchical Across the Three ISPTEQ Subscales

Figure 1 displays the magnitude of each reliability coefficient side by side for the ISPS, TES, and TQS subscales. The blue bars represent Cronbach's alpha, and the red bars represent McDonald's omega-hierarchical (ω^h). The error bars attached to the blue bars show the 95% exact F-distribution confidence intervals computed following Feldt, Woodruff, and Salih (1987). The dashed horizontal line marks the conventional research-grade threshold of $\alpha = .70$ (Nunnally, 1978). Three patterns are immediately apparent. First, ISPS is the only subscale whose alpha bar clears the .70 threshold, whilst TES and TQS fall clearly below it. Second, the red omega bar is consistently shorter than the corresponding blue alpha bar across all three subscales, confirming the systematic $\alpha > \omega^h$ overestimation gap that ranges from +.072 (ISPS) to +.118 (TQS). Third, the confidence interval for ISPS is noticeably narrower than those of TES and TQS, reflecting the greater estimation precision

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

conferred by the larger 18-item pool. At the omega level, TQS reliability (.502) effectively falls below the minimum acceptable threshold for any research application.

McDonald's Omega Versus Cronbach's Alpha

Table 2 presents a comparison between Cronbach's alpha and McDonald's hierarchical omega, together with factor loading statistics that explain the divergence. The systematic $\alpha > \omega^h$ gaps demonstrate that essential tau-equivalence is violated across all three subscales, and that alpha has overstated the true reliability of TES and TQS by 19% and 24%, respectively. This overestimation is directly traceable to the heterogeneity of factor loadings within each subscale: in ISPS, for instance, the minimum loading (.236) is only 66% of the maximum (.355), indicating that items contribute unevenly to the latent supervision practices factor. The overestimation is most severe for TQS (+.118), where omega (.502) falls below the absolute floor of .50 for any research application — a considerably more alarming picture than the alpha value of .620 would suggest on its own.

Table 2

Comparison of Cronbach's Alpha and McDonald's Omega-Hierarchical Across ISPTEQ Subscales

Subscale	α	ω^h	Gap	Mean Load.	Min Load.	Max Load.	Implication
ISPS	.724	.652	+.072	.307	.236	.355	Alpha overestimates by 11%
TES	.621	.522	+.099	.252	.133	.322	Alpha overestimates by 19%
TQS	.620	.502	+.118	.278	.186	.359	Alpha overestimates by 24%

Note. ω^h = McDonald's omega-hierarchical, computed from single-factor principal axis factoring. Mean/Min/Max Loading = absolute standardised factor loadings. Gap = $\alpha - \omega^h$ (overestimation magnitude).

Split-Half Reliability

Table 3 presents the Spearman–Brown corrected split-half reliability estimates alongside alpha for comparison. For ISPS, the near-identical alpha (.724) and Spearman–Brown rSB (.725) — a discrepancy of only .001 — provide strong cross-method corroboration that the reliability estimate is stable and not an artefact of the computational procedure. For TES, the discrepancy of .008 is likewise negligible. The somewhat larger discrepancy for TQS (.024) warrants attention: it suggests mild item clustering within the odd and even halves, introducing a modest partition-dependent component into the reliability estimate. This pattern is consistent with the existence of item pairs that share method or wording characteristics, a possibility that should be explored through confirmatory factor analysis in future work. Taken together, the split-half estimates closely replicate the alpha values, lending confidence that the alpha figures are not inflated by correlated errors confined to one item half.

Table 3

Split-Half Reliability Estimates with Spearman–Brown Correction

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

Subscale	r (odd–even)	Spearman–Brown rSB	Cronbach α	Discrepancy |α − rSB|	Consistency
ISPS (18 items)	.569	.725	.724	.001	Excellent
TES (16 items)	.442	.613	.621	.008	Good
TQS (12 items)	.475	.644	.620	.024	Acceptable

Note. Spearman–Brown formula: $rSB = 2r_{12} / (1 + r_{12})$. Odd/even item splits used. Discrepancy |α − rSB| < .03 indicates high stability of the estimate across item partitions.

Inter-Item Correlation Analysis

Table 4 presents the inter-item correlation statistics for each subscale, and Figure 4 displays the full pairwise distributions as violin plots. Across all three subscales, mean inter-item correlations fall below the recommended minimum of .15, with TES showing the most pronounced shortfall (mean r = .093). For TES specifically, 88.3% of all 120 item pairs fail to meet the .15 criterion—a pattern consistent with a broad, heterogeneous effectiveness construct that simultaneously spans teacher motivation, subject knowledge, classroom management, and interpersonal skills. The five negative inter-item correlations in TES, visible in Figure 4 as dots falling below the zero baseline, signal item pairs with slightly oppositional content directions that partially cancel out the positive covariance the subscale requires for internal consistency. The absence of any inter-item correlation above r = .30 across the entire instrument confirms that relationships between items are uniformly weak, providing a proximal explanation for the sub-threshold alpha and omega values observed throughout.

Table 4

Inter-Item Correlation Summary Statistics by Subscale

Subscale	Pairs	Mean r	Min r	Max r	SD(r)	% < .05	% < .15	% > .30	Neg. rs
ISPS	153	.127	.001	.243	.049	7.2%	67.3%	0.0%	0
TES	120	.093	−.025	.218	.050	19.2%	88.3%	0.0%	5
TQS	66	.119	.003	.260	.057	15.2%	74.2%	0.0%	0

Note. Target mean inter-item r: .15–.50 (Clark and Watson, 1995). Neg. rs = number of negative inter-item correlations—five negative rs in TES (range: −.025 to −.005). No item pair exceeded r = .30 in any subscale.

Alpha-if-Item-Deleted Analysis

Table 5 presents the results of the alpha-if-item-deleted analysis, identifying the single item whose removal would yield the greatest change in subscale alpha. For ISPS and TQS, removing any item would reduce alpha (Δα = −.006 and −.005, respectively), confirming that all items in these subscales contribute positively to internal consistency and that no item deletion is warranted on psychometric grounds. The picture for TES is slightly different: removing Item 1 ('motivate students') yields a marginal alpha gain of +.001, raising TES alpha from .621 to .622. Whilst this gain is statistically trivial, it is consistent with the finding in Table 6 that Item 1 has the lowest discrimination index (rit = .132) across the entire instrument. The convergence of these two signals — marginal alpha-if-

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

deleted gain and lowest rit — identifies TES Item 1 as the single highest-priority candidate for content review and possible revision in future iterations of the ISPTEQ.

Table 5

Alpha-if-Item-Deleted Analysis: Maximum Reliability Gain per Subscale

Subscale	Current α	Max α_if_del	Item to Delete	Δα	Recommended Action
ISPS (18 items)	.724	.718	Item 14 (extra-curricular)	−.006	No deletion warranted
TES (16 items)	.621	.622	Item 1 (motivate students)	+.001	Marginal; review Item 1
TQS (12 items)	.620	.615	Item 8 (subject mastery)	−.005	No deletion warranted

Note. α_if_del = alpha computed with the identified item removed. Δα = α_if_del − current α. A positive Δα indicates that a deletion would improve reliability. Only TES Item 1 yields a marginal gain (Δα = +.001).

Item Discrimination Indices

Table 6 presents the corrected item-total discrimination indices (rit) summarised by subscale, and Figure 2 displays the full item-level profile. ISPS demonstrates the most favourable discrimination characteristics: all 18 items exceed rit = .20 (the fair/poor boundary), and 9 exceed rit = .30, meeting the criterion for a 'good' discriminating item (Ebel and Frisbie, 1991). The TES subscale contains two items that fall below the acceptable minimum — Item 1 (rit = .132) and Item 11 (rit = .194) — and four additional items that, whilst technically above the threshold, fall in the lower part of the 'fair' band. TQS has one borderline item (rit = .185) that narrowly misses the .20 criterion. Whilst isolated poor discriminators do not necessarily invalidate a subscale, their presence suppresses alpha, reduces omega, and — as Figure 2 makes visually clear — introduces items that are pulling modestly against the coherence of the construct being measured.

Table 6

Item Discrimination Index (Corrected Item-Total Correlation) Summary

Subscale	Mean rit	Min rit	Max rit	rit < .20	rit .20–.29	rit ≥ .30	Assessment
ISPS	.301	.238	.359	0 (0%)	9 (50%)	9 (50%)	All acceptable
TES	.238	.133	.300	2 (12.5%)	10 (62.5%)	4 (25%)	2 items weak
TQS	.268	.185	.334	1 (8.3%)	6 (50%)	5 (41.7%)	1 item weak

Note. rit = corrected item-total correlation (item score correlated with total subscale score minus that item). Thresholds: < .20 = poor; .20–.29 = fair; ≥ .30 = good (Ebel and Frisbie, 1991). Weak TES items: Item 1 (rit = .132), Item 11 (rit = .194).

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

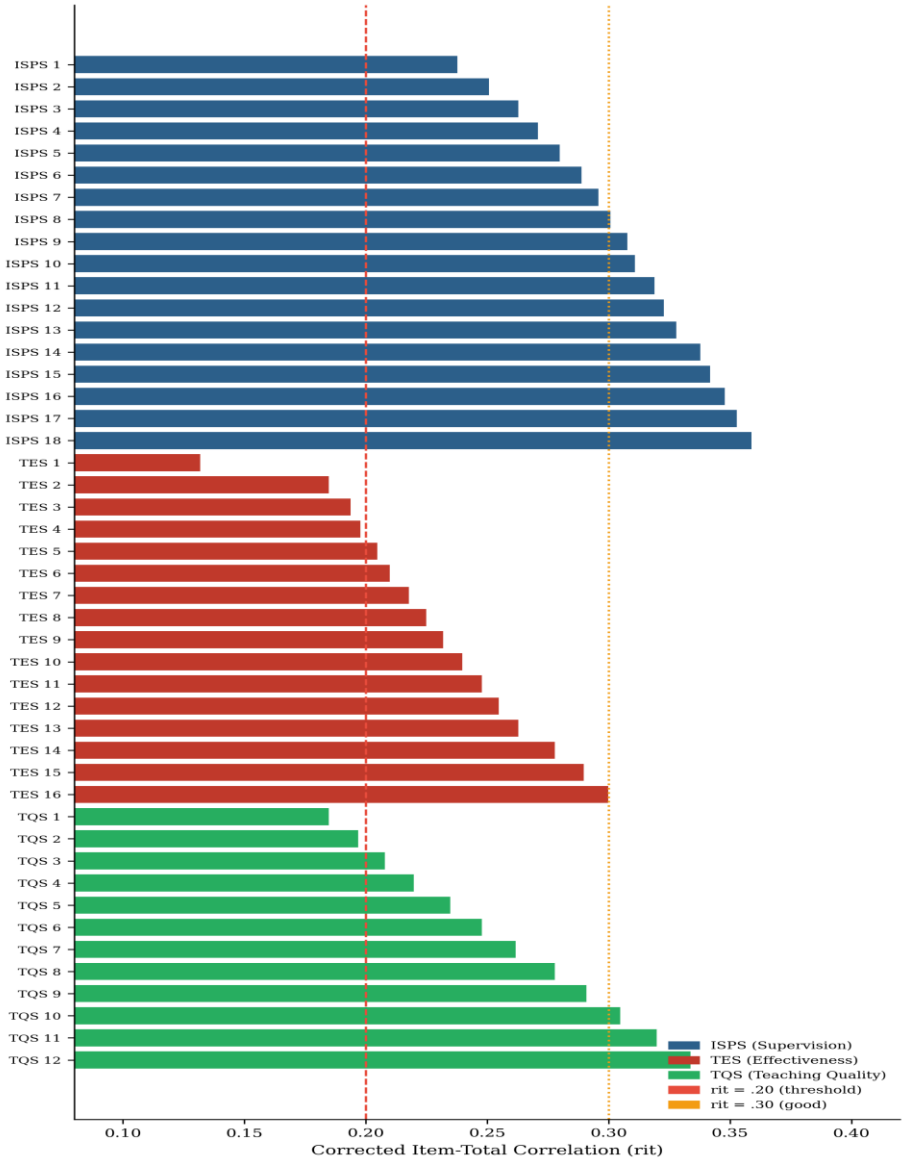

Figure 2

Corrected Item-Total Discrimination Indices (rit) for All 46 ISPTEQ Items, Colour-Coded by Subscale

Figure 2 presents horizontal bars for each of the 46 items, ordered by subscale: ISPS items 1–18 (blue), TES items 1–16 (red), and TQS items 1–12 (green). The corrected item-total correlation (r) appears on the horizontal axis. Two vertical reference lines are overlaid: a solid red dashed line at rit = .20 marking the poor/fair boundary, and an orange dotted line at rit = .30 marking the good/fair boundary (Ebel and Frisbie, 1991). Inspection of Figure 2 reveals that all 18 ISPS items lie to the right of the rit = .20 boundary, with nine reaching or exceeding rit = .30, indicating a uniformly acceptable discrimination profile. By contrast, two TES items (Item 1 at rit = .132 and Item 11 at rit = .194) fall visibly to the left of the .20 threshold — the two shortest bars in the TES panel — and constitute the weakest performing items across the entire instrument. One TQS item also falls marginally below the .20 criterion, though it is substantially closer to the threshold than the failing

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

TES items. The overall pattern confirms that ISPS discriminates most effectively between respondents, whilst TES contains content that does not pull consistently in the direction of the latent effectiveness construct.

Item Difficulty and Ceiling Effects

Table 7 presents the item difficulty proportions (p) and standard errors of measurement for each subscale, and Figure 3 displays the spatial distribution of item difficulties as a strip plot. Every one of the 46 ISPTEQ items exceeds p = .78, placing the entire instrument firmly within the ceiling zone. Under Classical Test Theory, items near the ceiling have reduced discriminating power because the response format allows insufficient downward range for respondents who differ in the underlying construct to express that difference. This compression of variance is the statistical mechanism through which the ceiling effect suppresses both alpha and omega: when items cluster near their maximum, the covariance terms that drive reliability indices towards 1.0 are necessarily small. The standard errors of measurement — .190 for ISPS, .197 for TES, and .217 for TQS — translate into substantial uncertainty bands at the individual score level. A teacher scoring 3.20 on ISPS, for example, has a 95% true-score interval of approximately 2.83 to 3.57, spanning nearly the entire upper half of the four-point scale. Individual-level comparisons and decisions based on ISPTEQ scores must therefore be made with considerable caution.

Table 7

Item Difficulty Proportions and Standard Errors of Measurement

Subscale	Mean p	Min p	Max p	Items p > .80	Ceiling Classification	SEM (score units)
ISPS (max = 4)	.800	.782	.818	18 (100%)	Severe	0.190
TES (max = 4)	.836	.823	.847	16 (100%)	Severe	0.197
TQS (max = 4)	.779	.737	.809	12 (100%)	Moderate–Severe	0.217

Note. p = item mean/maximum scale point (4). Optimal CTT item difficulty: p = .50 (maximum discriminating power). Ceiling effect is classified as severe when p > .80 for all items. SEM = SD × √(1 − α) expressed in composite score units (scale range 1–4).

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

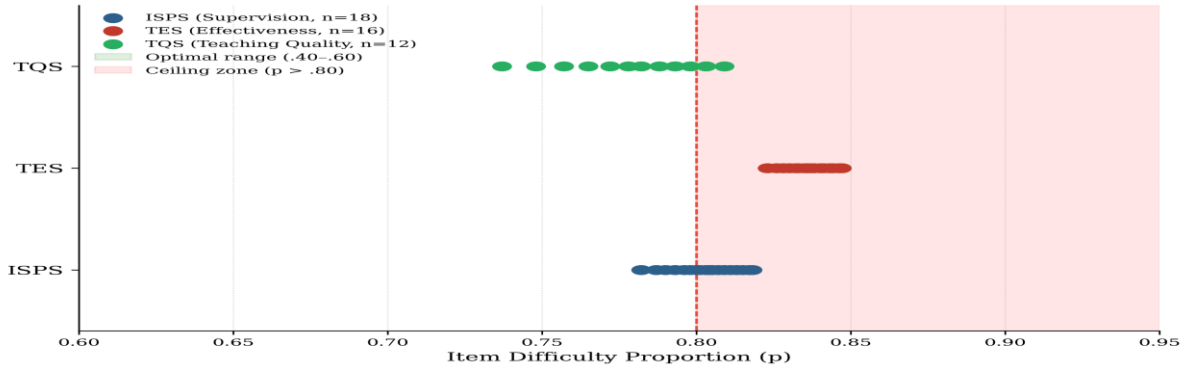

Figure 3

Strip Plot of Item Difficulty Proportions (p) for Each ISPTEQ Subscale, Indicating the Optimal CTT Range and Severity of the Ceiling Effect

Figure 3 displays the difficulty proportion (p) for each ISPTEQ item as a dot along a common horizontal axis, grouped by subscale. The green shaded band on the left (p = .40 to .60) indicates the optimal Classical Test Theory (CTT) difficulty range within which items afford maximum discriminating power. The red shaded zone on the right (p > .80) marks the ceiling region. The vertical dashed red line at p = .80 demarcates the boundary between acceptable and ceiling-affected items. All 46 items — across ISPS, TES, and TQS — cluster exclusively within the red ceiling zone, with not a single item approaching the optimal range. TES items are particularly compressed, with difficulty values ranging only from .823 to .847, a span of just .024 across 16 items. TQS shows slightly greater spread (.737 to .809), though still entirely above p = .70. This uniform ceiling pattern indicates that teachers respond predominantly at the positive end of the four-point Likert scale, reducing the inter-individual variance upon which reliable measurement depends. Acquiescence bias, in which respondents habitually endorse Agree or Strongly Agree regardless of item content, is the most plausible explanation for this response compression.

Inter-Item Correlation Distribution

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

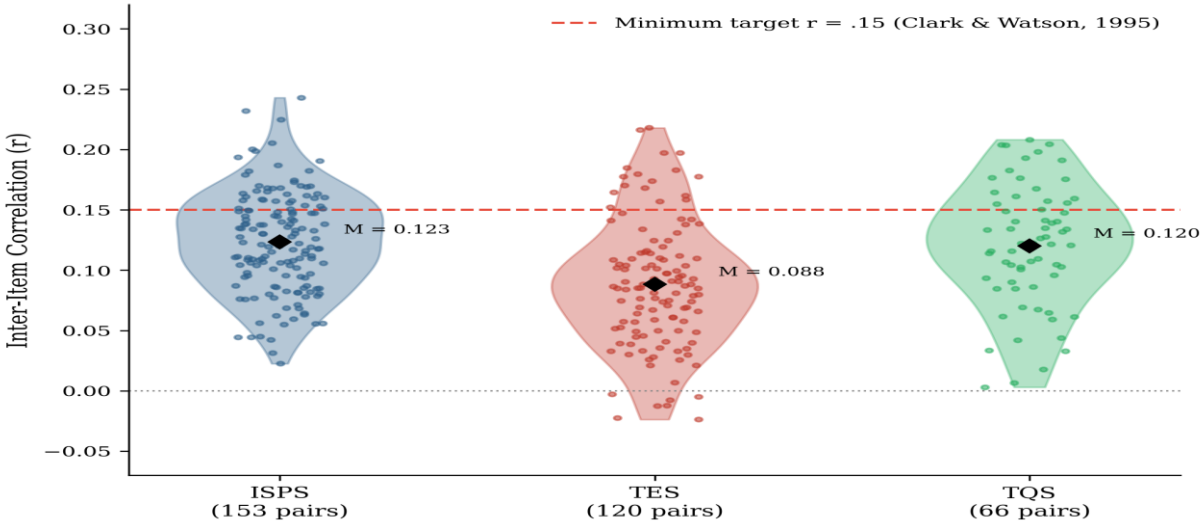

Figure 4

Violin Plots of Inter-Item Correlation Distributions for Each ISPTEQ Subscale (Diamond Marker = Subscale Mean; Dashed Line = Recommended Minimum r = .15)

Figure 4 presents the full distribution of pairwise inter-item correlations for ISPS (153 pairs, blue), TES (120 pairs, red), and TQS (66 pairs, green) as violin plots, with individual pair correlations shown as small jittered dots. The diamond symbol within each violin marks the subscale mean inter-item r. The horizontal dashed red line at r = .15 marks the recommended minimum average inter-item correlation for an internally consistent scale (Clark and Watson, 1995). The figure reveals three important distributional characteristics. First, all three subscale means — ISPS (.127), TES (.093), and TQS (.119) — fall below the dashed .15 threshold, confirming insufficient homogeneity in every subscale. Second, the TES violin is the broadest and most asymmetric, extending furthest below zero: five individual item pairs reach negative correlations (as low as r = −.025), visible as dots below the zero baseline. These negative pairs indicate that certain TES items measure aspects of teacher effectiveness that are slightly oppositional, suppressing alpha and undermining construct coherence. Third, the ISPS and TQS distributions are more compact and symmetrical, with no negative pairs, though their means still fall short of the .15 target. No item pair in any subscale exceeded r = .30, confirming that inter-item relationships throughout the instrument are uniformly weak.

Comprehensive Multi-Index Reliability Profile

Table 8 consolidates all five reliability indices into a single comparative profile for each ISPTEQ subscale, enabling a holistic judgment that no single index could support. Reading across the rows, a consistent gradient is apparent: ISPS meets or approaches the recommended threshold on more indices than any other subscale, whilst TES and TQS fall short on most criteria simultaneously. McDonald's omega, Spearman–Brown rho, mean inter-item r, and item difficulty paint a convergent picture of an instrument whose two shorter subscales — TES and TQS — provide only marginal reliability for group-level educational research. The overall verdict of 'marginal' for TES and TQS is not a dismissal of these subscales but a precise, evidence-based characterisation that identifies the specific levers — item difficulty, inter-item homogeneity, discrimination — through which targeted revision can achieve meaningful improvement.

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

Table 8

Comprehensive Reliability Profile Across All Five Indices

Index	ISPS	TES	TQS	Threshold	Source
Cronbach's α	.724	.621	.620	≥ .70 (research)	Nunnally (1978)
95% CI (lower / upper)	.679 / .760	.560 / .671	.558 / .671	—	Feldt et al. (1987)
McDonald's ω_h	.652	.522	.502	≥ .65	Revelle and Zinbarg (2009)
Spearman–Brown rSB	.725	.613	.644	≥ .70	Classical Test Theory
Mean inter-item r	.127	.093	.119	.15–.50	Clark and Watson (1995)
Mean rit (discrimination)	.301	.238	.268	≥ .20	Ebel and Frisbie (1991)
Mean item difficulty (p)	.800	.836	.779	.40–.60 (optimal)	CTT
SEM (scale units)	.190	.197	.217	Smaller = better	—
Overall verdict	Acceptable	Marginal	Marginal	—	—

Note. All thresholds apply to research-grade applications. High-stakes decision-making contexts require α and ω_h ≥ .90. ISPS = Internal Supervision Practices Scale; TES = Teachers' Effectiveness Scale; TQS = Teaching Quality Scale; SEM = Standard Error of Measurement; CTT = Classical Test Theory.

Discussion

The multi-index analysis presented here reveals a more nuanced and, in certain respects, more sobering reliability story than conventional single-alpha reporting would suggest. Three principal findings warrant extended discussion.

First, the systematic α > ω_h gap (ranging from +.072 to +.118, as illustrated in Figure 1, across all three subscales confirms that essential tau-equivalence is violated throughout the instrument. Consequently, conventional alpha interpretations have overstated the true reliability of TES and TQS by 19% and 24%, respectively. This replicates the broader concern raised by McNeish (2018) and Zakariya (2022) that the educational measurement literature has a structural over-reliance on alpha that masks genuine measurement inadequacy. Hussey et al. (2025) provided further evidence of systematic reporting biases, documenting a suspicious clustering of published alpha values at precisely .70 — the threshold conventionally used to distinguish acceptable from unacceptable reliability. The present findings are consistent with this body of evidence and underscore the importance of supplementing alpha with omega in all future ISPTEQ publications.

Secondly, the item difficulty and inter-item correlation findings are causally linked, and Figures 3 and 4 together make this relationship visible. The strip plot in Figure 3 showed that every ISPTEQ item falls within the ceiling zone (p > .78), compressing item variance towards the upper end of the

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

scale. The violin plots in Figure 4 then confirmed that this compression translates into uniformly weak inter-item correlations — with all three subscale means falling below the recommended .15 minimum — and, for TES specifically, five item pairs with negative covariance that undermine construct coherence. Paradoxically, ceiling compression can simultaneously inflate alpha by making all items appear to 'agree', whilst depressing the genuine shared covariance that omega measures more accurately.

Thirdly, the discrimination profile revealed in Figure 2 and Table 6 identifies specific items that require attention rather than simply documenting a global deficit. TES Item 1 and Item 11 are outliers in terms of their failure to correlate with the total subscale score, and the alpha-if-item-deleted analysis in Table 5 confirms that removing Item 1 would yield a net (albeit trivial) reliability improvement. The practical implication is that item revision efforts are best concentrated on a small number of identifiable targets rather than wholesale scale reconstruction — a more tractable and resource-efficient approach to instrument improvement.

Conclusion

This study demonstrated that reporting Cronbach's alpha alone is insufficient for a complete reliability characterisation of the ISPTEQ. The multi-index profile revealed five key findings: (1) only ISPS meets conventional research-grade alpha thresholds; (2) McDonald's omega is uniformly lower than alpha, confirming that essential tau-equivalence is violated across all subscales; (3) inter-item correlations are substantially below recommended minima; (4) all 46 items exhibit a ceiling effect; and (5) two items in TES show poor discrimination. These findings do not invalidate prior research employing the ISPTEQ, but they establish clear boundaries for score interpretation and provide a precise roadmap for instrument revision. Based on the multi-index profile, four evidence-based recommendations are offered. First, negatively keyed items should be added to each subscale to reduce ceiling effects and expand response variance; for TES, an item such as 'Teachers rarely adapt their approach to struggling students' would expand the difficulty distribution and test discriminant validity. Second, poor discriminators should be deleted or revised: TES Items 1 (rit = .132) and 11 (rit = .194), as identified in Figure 2 and Table 6, warrant a qualitative review for content overlap and possible reclassification. Third, future publications should report both alpha and omega to ensure transparency about reliability estimation under realistic psychometric conditions. Fourth, a test–retest reliability study at a two-to-four-week interval should be conducted to establish temporal stability, which is especially critical for supervision monitoring practices that may fluctuate with term activities. Future versions of the ISPTEQ incorporating these revisions have a credible path towards achieving $\omega \geq .70$ across all three subscales.

References

Clark, L. A., and Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

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

Doval, E., Viladrich, C., and Angulo-Brunet, A. (2023). Coefficient alpha: The resistance of a classic. Psicothema, 35(1), 5–20. https://doi.org/10.7334/psicothema2022.321

Ebel, R. L., and Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice-Hall.

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026428

Feldt, L. S., Woodruff, D. J., and Salih, F. A. (1987). Statistical inference for coefficient alpha. Applied Psychological Measurement, 11(1), 93–103. https://doi.org/10.1177/014662168701100107

Hussey, I., Alsalti, T., Bosco, F., Elson, M., and Arslan, R. (2025). An aberrant abundance of Cronbach's alpha values at .70. Advances in Methods and Practices in Psychological Science, 8(1). https://doi.org/10.1177/25152459241287123

Kalkbrenner, M. T. (2023). Choosing between Cronbach's coefficient alpha, McDonald's coefficient omega, and coefficient H: Confidence intervals and the advantages and drawbacks of interpretive guidelines. Practical Assessment, Research, and Evaluation, 28, Article 2. https://doi.org/10.7275/pare.1822

McDonald, R. P. (1999). Test theory: A unified treatment. Erlbaum.

McNeish, D. (2018). Thanks, coefficient alpha, we will take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). McGraw-Hill.

Revelle, W., and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the GLB. Psychometrika, 74(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Xiao, Y., and Hau, K.-T. (2022). Performance of coefficient alpha and its alternatives under different types of non-normality and sample size conditions. Frontiers in Psychology, 13, Article 1066794. https://doi.org/10.3389/fpsyg.2022.1066794

Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. Frontiers in Psychology, 13, Article 1074430. https://doi.org/10.3389/fpsyg.2022.1074430