International Journal of Economic Practices and Theories (IJEPT)
ISSN: 2247-7225
Volume 2026, Issue 1

Internal Consistency and Item-Level Validity of a Six-Construct Leadership-Attributes Scale

G K Ashwin[1], Dr. Suresh Chandra Padhy[2]

[1]Research Scholar, Department of Management, Poornima University, Jaipur (India)

[2]President (Vice Chancellor), Poornima University, Jaipur (India)

Abstract

Psychometric instruments used in comparative educational research are frequently deployed in a main study on the strength of pilot-study reliability alone, with little scrutiny of whether that reliability replicates at scale. This paper reports such a case. A pilot study of a 30-item, six-construct leadership-attributes questionnaire (self-efficacy, resilience, emotional intelligence, cognitive competence, social preference, perceived leadership behaviour) reported Cronbach's alpha of 0.60–0.91 across the six constructs. When the same instrument was administered to a substantially larger main sample (N = 450; 259 Military School and 191 Non-Military School students), internal consistency collapsed: alpha ranged from −0.09 to 0.05 across all six constructs, and inter-item/inter-construct correlations were uniformly weak. This reliability failure is reported in full, some potential explanations are explored (contextual item interpretation, variability in adolescent responses, and construct multi-dimensionality), and it is shown that the findings of an analysis of the items (not the composite scores) are interpretable and theoretically coherent even if the composite score is not. We assert that item-level reporting should be used as a must, and not an optional add-on, to the composite reliability when composite reliability is near the margin or unknown at the time of analysis of the main study.

Keywords: Cronbach's alpha; internal consistency; scale validation; item-level analysis; psychometrics; adolescent self-report

1. Introduction

Reliability the degree to which a set of items measures a single underlying construct consistently is a precondition for interpreting any composite score derived from a multi-item scale. In practice, however, many comparative educational-psychology studies establish reliability only once, at the pilot stage, and then proceed to treat composite construct scores from a larger main study as if that pilot-stage reliability still holds. This is rarely tested explicitly, and even more rarely reported when it fails.

This paper documents a case in which it did fail, markedly. A six-construct, 30-item leadership-attributes questionnaire was piloted with acceptable-to-excellent internal consistency (Cronbach's $\alpha = 0.60–0.91$) and then administered to a much larger main sample of 450 secondary-school students. In the main sample, internal consistency for every one of the six constructs fell to near zero or below. We treat this not as a failure to be concealed but as a substantive methodological finding in its own right, with implications for how comparative educational-psychology research should be designed, analysed, and reported.

The purpose of the paper is to have two objectives. The first is diagnostic: to describe the reliability collapse as accurately as possible, as much as the available data can support, and then to evaluate candidate explanations for the reliability collapse against the evidence, not make them up. The second is constructive; to demonstrate that the breakdown of the reliability of composites does not negate the usefulness of the data, and that a careful item-level analysis can yield a theoretically coherent, substantively interpretable pattern of findings.

2. Background: Reliability as a Precondition for Composite Scoring

2.1 Classical Test Theory and the Logic of Alpha

Classical test theory models an observed score as the sum of a true score and random measurement error. For a multi-item scale intended to measure a single construct, each item is treated as a parallel, imperfect indicator of that same true score, and the composite (typically the sum or mean of the items) is assumed to average out item-specific error while retaining the shared true-score signal. Cronbach's alpha is an operationalisation of this: It is a function of the size of the composite and the amount of inter item covariance as a proportion of the total variance; it is a measure of the proportion of variance in the composite that is due to this common signal and not due to item specific noise. Internal consistency reliability is then measured by the proportion of variance that can be accounted for by a common factor (usually represented by Cronbach's alpha) as opposed to variance caused by noise (chance) for each item in a composite score. The α value close to zero means that the items in the putative scale don't covary as a scale in the sample under study respondents are not answering the items as if they are measuring a common underling trait. A composite score that is constructed from items such as these merges together a variety of response patterns, and thus loses any clear substantive interpretation. In contrast, every item maintains a more interpretable meaning – a difference in the scale as a whole does not affect a difference in a single item with a clear statement. According to conventional standards (Nunnally & Bernstein, 1994), $\alpha \geq .70$ is an acceptable level for research while $\alpha \geq .90$ is necessary for individual-level decision making, and $\alpha < .50$ is generally unacceptable for any purpose. A second diagnostic clue is one that is not merely "low" but negative: If alpha is negative, on average, the items within the putative scale tend to be negatively related, rather than positively related, meaning that they tend to be equally (or even more!) negatively correlated than positively, which is unlikely to be attributable to any single underlying trait, however weakly expressed, but is very different from a merely "low but positive" alpha (Bland & Altman, 1997). But there is no shortage of critics to Alpha. However, Schmitt (1996) and Taber (2018) have warned that often alpha is reported and interpreted without regard to whether the alpha is warranted given the conditions under which it can be a valid index, which is that items must be essential tau equivalent, and that a low alpha does not necessarily indicate low item quality. In essence, this is the diagnostic approach that we take in this paper. The main-study result reported here was a substantial outlier in the context of a meta-analysis of more than 200 alpha coefficients reported in the organisational and social-science literature by Peterson (1994) – it was not a disappointing, but typical, value.

2.2 Pilot Reliability as a Hypothesis, Not a Guarantee

Of course, the reliability data gathered in the pilot phase are generally from small samples, and sometimes convenience samples. If the alpha range is 0.60-0.91, it is a good reason to move on to a main study but it is an educated guess as to how the instrument will behave in the main study and not a best guarantee. Reliability is a characteristic of scores of a sample, not a fixed quality of an instrument; materially different samples (in size, heterogeneity, age range, institutional settings, etc., and response style) can yield different alpha values for the same items (Schmidt & Hunter, 1996). The hypothesis presented here should be retested (not assumed!) when the data from the main study become available, and the outcome of the retesting process should be reported — including when the data indicate as emphatically as here that the hypothesis is false.

3. Method

3.1 Instrument and Sample

This instrument consisted of six subscales with 5 items each: Self-Efficacy (SE), Resilience (RE), Emotional Intelligence (EI), Cognitive Competence (CC), Social Preference (SP) and Perceived Leadership Behaviour (LP) which were rated on a five point Likert scale. The main study sample consisted of 450 students (259 Military School, 191 Non-Military School; Class IX to XII - Ages 13-18).

3.2 Analysis

Standard procedures were used to compute Cronbach's alpha with 95% CI's for each of the 6 five item subscale scores and the total 30 item scale score (Kline, 2000). Pearson correlation coefficients were calculated between

the six composite construct scores to determine if there were any significant correlations between the constructs even if the subscales were not internally consistent. Item-level descriptive statistics (mean, standard deviation, skewness, and kurtosis) were computed for all thirty items, separately by school type, to evaluate whether restricted variability, ceiling effects, or non-normal response distributions could account for the reliability collapse. Friedman's test was used to compare the within-group rank ordering of the six composite constructs, separately for each school type, as a supplementary check on whether the two groups show systematically different response patterns across constructs, and chi-square tests of association (Fleiss, 1981) examined item-by-demographic relationships reported in the companion paper. Independent-samples t-tests were then run at the individual-item level (rather than on composite scores) to determine whether substantively interpretable, statistically defensible findings remained available despite the composite-level reliability failure. All analyses were conducted following conventional procedures for behavioural-science data (Field, 2013).

4. Results

4.1 Internal Consistency

Table 1. Main-study Cronbach's alpha by construct, compared against pilot-study range of 0.60–0.91.

Construct	α (main study)	95% CI	Rating
Self-efficacy	-0.044	[-0.205, 0.101]	Unacceptable
Resilience	0.046	[-0.100, 0.179]	Unacceptable
Emotional Intelligence	-0.029	[-0.188, 0.114]	Unacceptable
Cognitive Competence	0.034	[-0.114, 0.169]	Unacceptable
Social Preference	-0.093	[-0.262, 0.059]	Unacceptable
Perceived Leadership Behaviour	-0.063	[-0.227, 0.085]	Unacceptable
All 30 items combined	-0.065	—	Unacceptable

Every construct returned an alpha at or below 0.05 in the main sample, with several constructs returning small negative values a result that can occur when item covariances are, on balance, as likely to be negative as positive, the opposite of what a coherent scale requires. The full 30-item set fared no better ($\alpha = -0.065$).

4.2 Inter-Construct Correlation

Pearson correlations among the six composite construct scores were uniformly weak (most $|r| < 0.10$). The only statistically significant pair was Emotional Intelligence × Social Preference ($r = -0.125$, $p = 0.008$). Table 2 reports the full 6×6 inter-construct correlation matrix.

Table 2. Full inter-construct Pearson correlation matrix (N = 450).

Construct	SE	RE	EI	CC	SP	LP
Self-efficacy	1.00	0.03	-0.09	0.01	0.01	0.02
Resilience	0.03	1.00	0.01	0.00	-0.00	-0.01

Construct	SE	RE	EI	CC	SP	LP
Emotional Intelligence	-0.09	0.01	1.00	-0.01	-0.13	0.02
Cognitive Competence	0.01	0.00	-0.01	1.00	0.01	-0.02
Social Preference	0.01	-0.00	-0.13	0.01	1.00	0.02
Perceived Leadership Behaviour	0.02	-0.01	0.02	-0.02	0.02	1.00

This near-zero, uniformly weak matrix rules out one candidate explanation for the reliability failure namely, that the six constructs had collapsed into one or two broader underlying factors not anticipated by the original six-construct model. Instead, the constructs behave as largely independent dimensions, and the unreliability appears to originate within, not between, the five-item subscales.

4.3 Item-Level Distributional Characteristics

A low or negative alpha can in principle arise from at least two distinct sources: items that are not measuring a common construct (the substantive concern of this paper), or items that are measuring a common construct but with so little response variability — for example, a strong ceiling effect, where nearly all respondents select the top one or two scale points — that true covariance cannot be detected statistically. To evaluate the second possibility, Table 3 reports descriptive statistics for all thirty items.

Table 3. Item-level descriptive statistics, full sample (N = 450).

Item	Construct	M	SD	Skew	Kurtosis
SE1	Self-efficacy	3.88	0.99	-0.75	0.32
SE2	Self-efficacy	4.15	0.81	-0.79	0.72
SE3	Self-efficacy	4.10	0.94	-1.20	1.49
SE4	Self-efficacy	3.78	1.06	-0.74	0.11
SE5	Self-efficacy	4.19	0.80	-1.19	2.06
RE1	Resilience	4.16	0.82	-1.00	1.35
RE2	Resilience	3.94	0.94	-0.63	-0.02
RE3	Resilience	4.20	0.85	-1.28	2.33
RE4	Resilience	4.10	0.79	-0.49	-0.16
RE5	Resilience	3.90	0.91	-0.42	-0.27
EI1	Emotional Intelligence	3.91	0.99	-0.98	0.92

Item	Construct	M	SD	Skew	Kurtosis
EI2	Emotional Intelligence	4.09	0.97	-1.21	1.31
EI3	Emotional Intelligence	4.02	0.99	-0.96	0.60
EI4	Emotional Intelligence	4.02	0.89	-0.78	0.52
EI5	Emotional Intelligence	4.21	0.77	-0.99	1.50
CC1	Cognitive Competence	3.87	0.89	-0.23	-0.60
CC2	Cognitive Competence	4.07	0.86	-0.77	0.51
CC3	Cognitive Competence	4.16	0.78	-0.68	0.35
CC4	Cognitive Competence	4.14	0.81	-0.74	0.38
CC5	Cognitive Competence	3.96	0.96	-0.80	0.46
SP1	Social Preference	4.01	0.90	-0.86	0.76
SP2	Social Preference	4.20	0.81	-1.27	2.61
SP3	Social Preference	4.14	0.82	-1.10	1.86
SP4	Social Preference	3.92	0.95	-0.83	0.57
SP5	Social Preference	4.22	0.77	-0.93	1.04
LP1	Perceived Leadership Behaviour	3.99	0.91	-0.65	0.10
LP2	Perceived Leadership Behaviour	3.97	0.98	-0.90	0.47
LP3	Perceived Leadership Behaviour	4.03	0.95	-1.07	1.17
LP4	Perceived Leadership Behaviour	4.00	0.86	-0.77	0.71
LP5	Perceived Leadership Behaviour	4.08	0.88	-0.84	0.67

The standard deviations for the 30 items are all between about 0.85 and 1.10 on the five-point scale, which is quite a bit of variation, and nowhere near the near zero range that would suggest a true ceiling or floor effect. Skewness

values are moderate and all negative (as is common with positively worded self report items) and values of kurtosis have no extreme deviations from normality. This pattern eliminates the alternative interpretation of the restricted-variability: all the respondents did not give the same answer. The reliability collapse has to come from the interrelatedness of the respondents' responses to the various items in the same construct and not from the amount of variance in any given item.

4.4 Within-Group Construct Ranking (Friedman Test)

As a further, independent check on response patterns, Friedman's test compared the within-group rank ordering of the six composite constructs. Among Military students, the six constructs were statistically indistinguishable in rank ($\chi^2(5) = 1.469$, $p = 0.917$); among Non-Military students, the ranking was highly significant ($\chi^2(5) = 21.325$, $p = < .001$), with Social Preference rated clearly highest. This divergence shows that the two groups are not responding identically to the instrument as a whole: Non-Military students differentiate sharply among the six constructs in a way Military students do not, even though neither group's within-construct items cohere well enough to support a reliable composite. This is consistent with — though does not on its own prove — a response-style difference between the two groups, one of the candidate explanations discussed below.

4.5 Item-Level Findings Despite Composite Unreliability

Despite the composite-level reliability failure, item-level independent-samples t-tests produced a coherent, interpretable pattern: 10 of 30 items showed a statistically significant Military vs. Non-Military difference, with effect sizes predominantly small and a clear substantive clustering — significant items skewed toward action-oriented, task-facing behaviours (persistence, optimism, decisive action, inspiring others), while the few items favouring the comparison group concerned reflective or adaptive judgement (belief in achieving goals, learning agility). This clustering would not be expected if the item-level results were simply noise from an incoherent scale; it indicates that individual, well-worded items can retain valid, interpretable signal even when the composite scale they were designed to form does not.

5. Discussion

5.1 Candidate Explanations for the Reliability Collapse

Four candidate explanations merit consideration, evaluated below against the evidence gathered in Section 4.

Contextual item interpretation

Adolescent respondents may interpret superficially similar items — for example, two Self-Efficacy items addressing different task domains such as academic performance versus physical endurance — through different frames of reference, reducing their covariance even though each individually captures a meaningful judgement. This explanation is consistent with the item-level distributional findings in Section 4.3: items within a construct show broadly similar means and spreads individually, yet fail to covary, which is exactly the signature one would expect if respondents are answering each item against a different personal reference point rather than a shared underlying trait.

Response-style heterogeneity by institutional context

An institutional, disciplinary administration context, present for the Military subsample specifically, may introduce differential social-desirability or extreme-response patterns that are not uniform across items within a subscale. The Friedman-test divergence reported in Section 4.4 — a flat, non-significant construct ranking among Military students versus a sharply differentiated ranking among Non-Military students — is directly consistent with this explanation: it suggests the two groups are not engaging with the instrument in the same way, which would also depress the pooled, whole-sample alpha even if each subgroup, examined separately, showed somewhat more internal coherence.

Construct multi-dimensionality

What the pilot treated as a single five-item construct may in fact tap two or more narrower sub-dimensions in a more diverse main-study sample, each internally consistent but not consistent with one another. For example, a five-item Self-Efficacy subscale that blends academic, physical, and social self-efficacy items could show healthy alpha in a pilot sample drawn from a single, relatively homogeneous classroom, yet collapse once administered across a more diverse population in which these sub-domains no longer move together. This study did not conduct the exploratory or confirmatory factor analysis needed to test this explanation directly (see Limitations) — standard practice for which would include a sampling-adequacy check (Kaiser, 1974) ahead of factor extraction, following the multivariate procedures set out in texts such as Hair, Black, Babin, and Anderson (2010) — but it remains, in our assessment, the most parsimonious explanation given the uniformly weak inter-construct correlations reported in Section 4.2, which rule out a between-construct collapse but say nothing about a within-construct split.

Sample heterogeneity

A pilot sample is typically smaller and more homogeneous than a main sample; the same items can show higher apparent consistency in a narrower sample purely as a function of restricted variability in who is answering, even when the items' true covariance structure is unchanged. The jump from a single-digit pilot sample to a 450-student, two-school-type main sample in this study is large enough that some degree of reliability attenuation would be expected on this basis alone, though the magnitude of the collapse observed here — from acceptable-to-excellent to negative — is larger than sample-heterogeneity alone would typically predict.

5.2 Implications for Practice

Here we draw three practical implications: First of all, main study reliability must always be calculated and reported before using composite scores for interpretation or publication, pilot-study results are NOT the same as main study results, and a pilot alpha is only a hypothesis. Second, item-level analysis should not be understood as a "last resort" method of analysis, but as a viable and often more meaningful unit of analysis as well — in this case, item-level testing found a theoretically coherent and interpretable pattern that would have been buried in or even completely lost when analyzing at the composite level. Third, instrument developers need to plan for a comprehensive psychometric re-validation of their instrument when the instrument is used at a very different or larger scale and/or using a very different population than when it was originally pilot tested (not conducted here). One additional implication is offered from the group-level Friedman divergence presented in Section 4.4: When two or more groups are compared as the data may vary in institutional context and/or in response style, reliability should be reported separately per group, not pooled across all groups. A pooled alpha can mask a pattern in which one subgroup responds relatively coherently and another does not, and in the presence of such a pattern, the pooled figure under-represents the better-behaved subgroup and over-represents the problem.

5.3 Recommendations for Future Instrument Validation

- Conduct exploratory factor analysis on the main-study data (not merely the pilot) before relying on any composite score, to test whether the intended six-construct structure is recovered, and confirmatory factor analysis in an independent sample once a structure is proposed.
- Report reliability separately by subgroup (school type, gender) in addition to the pooled figure, given the response-style divergence identified here.
- Conduct cognitive interviews with a small sample of respondents from the target population before finalising item wording, to detect contextual-interpretation problems of the kind discussed in Section 5.1 before, rather than after, main-study data collection.
- Where a pilot sample is substantially smaller or more homogeneous than the intended main sample, treat the pilot alpha as a lower bound on required sample diversity, not as evidence of expected main-study reliability, and budget for an interim reliability check on an early tranche of main-study data before full data collection is complete.

- Retain and report item-level descriptive statistics (as in Table 3) as standard practice, since they provide an inexpensive first check on whether a reliability failure is attributable to restricted variability or must be attributed to a deeper structural cause.

5.4 Implications for the Broader Field

This case has implications beyond the specific instrument examined here. Comparative educational-psychology research on Indian secondary-school populations frequently adapts or newly develops multi-item Likert scales for constructs such as those examined in this study, often validated only at a small pilot stage before large-scale deployment. If reliability collapses of this magnitude are more widespread than currently documented — and the absence of routine main-study reliability reporting in much of this literature means the true prevalence is simply unknown — then a non-trivial share of published composite-score findings in this literature may rest on scales whose main-study reliability was never actually verified. We do not claim this is the case for any specific prior study, only that the practice of verifying and reporting main-study reliability, demonstrated as straightforward to do here, should become routine rather than exceptional.

5.5 Limitations

- The study did not use exploratory factor analysis or confirmatory factor analysis to determine if this arrangement of items on the six constructs (as well as other possible arrangements) better fits the data from the primary study.
- This study did not include cognitive interviewing of the adolescent respondents, and this approach is suggested to further clarify how the individual items were interpreted, and could be included in future revision of the instrument.
- The reliability finding does not extend to other instruments, or other samples, and should not be viewed as a general statement that Likert-type scales of leadership attributes are unreliable in general.
- The Friedman-test divergence presented in Section 4.4 is a supportive, but not definitive, indication that the reliability collapse might be due to response style.

6. Conclusion

A 30-item, 6-construct leadership-attributes instrument had acceptable reliabilities for the pilot study ($\alpha = 0.60 – 0.91$; $N = 120$), but these reliabilities were not good for the 450 students on the main study ($\alpha = -0.09$ to 0.05). However, this was found to be unlikely to be due to limited response variability at the item level (based on item-level descriptive statistics) and was not due to a collapse into fewer, broader factors as was indicated by the uniformly low inter-construct correlation matrix, leaving the most likely explanation for this as either an unintended collapse into fewer, broader factors (or an unintentionally interpreted context in the items) or to response style differences between the two school types, or to an undetected sub-dimensionality of the constructs, which is supported by the Friedman-test differences that were found between the two groups. We do not consider the failure to reach reliability to be a reason to dismiss the analysis, however, but argue that sticking with the item-level analysis still allows a coherent theoretically interpretable pattern of results to emerge. Reliability as a whole study should be calculated and reported as a standard practice before moving on to interpreting the composite score; reliability should be reported by each subscale if there are several subscales because some groups may have different contexts or response styles; and reliability should be performed as a regular robustness check of the data in comparative research in the field of educational psychology, not as an afterthought when reliability is poor.

References

[1] Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. BMJ, 314(7080), 572.

[2] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334.

[3] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

[4] DeVellis, R. F. (2016). Scale development: Theory and applications (4th ed.). Sage Publications.

[5] Field, A. (2013). Discovering statistics using IBM SPSS statistics (4th ed.). Sage Publications.

[6] Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd ed.). Wiley.

[7] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (7th ed.). Pearson.

[8] Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31–36.

[9] Kline, P. (2000). The handbook of psychological testing (2nd ed.). Routledge.

[10] Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

[11] Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. Journal of Consumer Research, 21(2), 381–391.

[12] Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. Psychological Methods, 1(2), 199–223.

[13] Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Assessment, 8(4), 350–353.

[14] Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. Journal of Personality Assessment, 80(1), 99–103.

[15] Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273–1296.

[16] Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55.