

**Nida Ayaz¹, Sayyeda Jawerya Ayaz²**

1. Lecturer, Department of Education, Virtual University, Khyber Pakhtunkhwa, Pakistan.

2. MPhil Scholar, Department of English, Abdul Wali Khan University Mardan Khyber Pakhtunkhwa, Pakistan.

Cite This Article: Ayaz, N. & Ayaz, S. J. (2026). Developing and Validating a Student-Rated Instructional Intelligence Scale for University Teachers in Khyber Pakhtunkhwa, Pakistan. *Journal of Social Sciences Research & Policy*. 4 (01), 1013-1021.

DOI: <https://doi.org/10.71327/jssrp.41.1013.1021>

ISSN: 3006-6557 (Online)

ISSN: 3006-6549 (Print)

Vol. 4, No. 1 (2026)

Pages: 1013-1021

Key Words:

Instructional intelligence; student

Evaluation of teaching; scale development

Principal component analysis; reliability

Higher education

Corresponding Author:**Nida Ayaz**Email: nida.ayaz@vu.edu.pk**License:**

Abstract: Evaluation by students is one of the valuable means to assess teaching in universities, and surprisingly there has been only a few psychometric analyses of student ratings measures which are based on the notion of instructional intelligence. This study investigated the internal structure, reliability, and item performance of a 30-item student rating scale containing six hypothesized factors: tactics, skills, organizers, strategies, concepts, and classroom management. The research design employed was cross-sectional survey and followed the principle of scale development. Data were collected from 400 students (250 from a public university and 150 from a private university) in Khyber Pakhtunkhwa, Pakistan. The sampling adequacy was satisfactory ($KMO = .897$; Bartlett's $\chi^2(435, N = 400) = 7,991.42, p < .001$). Twenty-seven items loaded above .50 on their corresponding component. Fourteen items had secondary loadings above .30. Internal consistency was high (Cronbach's $\alpha = .936$). There were no differences between sector ($d = 0.07$) or gender ($d = -0.02$) groups regarding total scores. The structure was interpretable mostly. Revision is required for items IO2, IO4, IO5, and ST1. The next steps are confirmatory and invariance analyses on independent sample.

Introduction

Student evaluations of teaching are widely used in higher education to collect feedback on teaching quality. Student ratings provide useful evidence when the instrument measures clearly defined teaching dimension reliably and validly (Marsh, 1984; Marsh & Roche, 1997; Spoor et al., 2013). Instructional Intelligence is a theory of teaching designed by Bennett & Rolheiser (2001) and refined by Bennett (2002, 2010). Instructional intelligence considers teaching as an intentional synthesis of the science of instruction (researched-based tactics, skills and strategies) and its artful adaptation to certain learners through careful consideration of lesson organizers, curriculum concepts and classroom management (Bennett & Rolheiser, 2001; Bennett, 2010). The theory is consistent with the idea of differentiated instruction stating that teachers must adapt teaching content and processes to the diversity of their students (Tomlinson, 2001), and the body of research on instructional strategies connected to achievement gains (Marzano et al., 2001).

Well-known measuring devices reveal that teaching quality is multi-dimensional. One example would be the Students' Evaluations of Educational Quality (SEEQ; Marsh, 1982), and the Course Experience Questionnaire (Ramsden, 1991). Both are generic in nature. No particular model of teaching is used in developing these tools. Therefore, the feedback received by teachers is generalized.

In Pakistan, the universities gather students' feedback through the quality assurance processes.

The instructional intelligence provides a model-based approach to the measurement of teaching. A measure developed from this model would allow teachers to receive specific feedback regarding tactics, skills, strategies, organizers, concepts, and classroom management techniques.

The current study aims to fill this gap. The aim of the current study was to analyze the internal structure, reliability and item characteristics of a 30-item scale on student-rated instructional intelligence at two universities in Khyber Pakhtunkhwa, Pakistan.

Research Questions

Therefore, this study was intended to answer four research questions:

1. **RQ1.** Do responses to the 30-item scale justify the hypothesized six dimensional structures?
2. **RQ2.** Are the scores on the total scale reliable?
3. **RQ3.** What are the problematic items and how can they be improved?
4. **RQ4.** Do total scores differ by university sectors or student gender?

Sector and gender were chosen as an initial analysis for group difference before any measurement invariance analysis.

Conceptual Framework

The scale measures six constructs that are based on the instructional intelligence framework (Bennett & Rolheiser, 2001; Bennett, 2010), consisting of five items in each dimension and thus within the range suggested (three to five items per factor) for obtaining stable factors (Fabrigar et al., 1999; Costello & Osborne, 2005). The six dimensions have been described in the following manner. Instructional tactics involve certain techniques which are used to organize the process of instruction. These include the use of diagrams, cooperative structures, and think time. Instructional skills refer to various teaching actions and include questioning, gaining the attention of students, waiting time, and providing feedback. Instructional strategies refer to more general approaches to instruction which involve both tactics and skills to achieve an instructional objective. Instructional organizers are those ways in which the teacher organizes his/her lesson. This includes setting the objectives for the lesson, organization of content, evaluation at the end of the lesson, and higher order activities. Instructional concepts refer to the major concepts of the curriculum based on which the instructional objectives are formulated.

Table 1. Hypothesized dimensions and items distribution.

Dimension	Cod e	Items	Illustrative content
Instructional tactics	IT	1-5	Diagrams, cooperative structures, individual thinking time, sharing views.
Instructional skills	IS	6-10	Questioning, attention, wait time, feedback.
Instructional organizers	IO	11-15	Lesson organization, goal setting, end-of-lesson evaluation, higher-order tasks.
Instructional strategies	ST	16-20	Learning resources, learning tasks, graphs, concept association with diagrams, learning styles of students.
Instructional concepts	IC	21-25	Student responsibility, involving all students, motivation, positive task atmosphere, tasks assigned as punishment.

Dimension	Cod e	Items	Illustrative content
Classroom management	CM	26-30	Reducing disruptive behavior, disruptive/aggressive students, pleasant resolution of minor problems, U-shaped seating, small groups.

Methodology

Research design

A cross-sectional survey research design has been used for a psychometric validation study. The research design adheres to the sequential nature of scale development procedure (item generation, expert rating, administration, dimensionality, and reliability analysis) described in literature (DeVellis, 2017; Worthington and Whittaker, 2006; Boateng et al., 2018) and is described according to the measurement-property framework of the COSMIN initiative (Mokkink et al., 2010). Evidence for only two measurement properties (internal structure and internal consistency) is provided; all other properties are listed as limitations.

Setting, population and sampling

The data collection was carried out in two universities (one public, Abdul Wali Khan University Mardan, and one private, Northern University) of Khyber Pakhtunkhwa, Pakistan. Stratified random sampling with proportional allocation has been employed for its advantage of reducing sampling error where the target population includes recognizable strata (Cochran, 1977). Finally, the last sample included N = 400 people: 250 students from the public university (150 males, 100 females) and 150 students from the private university (90 males, 60 females). [INSERT enrolment numbers for each university and data collection periods; see pre-submission notes.]

Sample Adequacy. The ratio of 400 respondents to 30 items (13.3:1) is higher than the recommended minimum of 10 cases per variable required for conducting factor analysis (Hair et al., 2019). According to Comrey and Lee (1992), 300 respondents represent a good sample. The component stability also relies on the loading value and the number of markers per component, both of which are favourable in this case (Guadagnoli & Velicer, 1988).

Instrument

Item Generation. Items were generated based on the instructional intelligence literature (Bennett & Rolheiser, 2001; Bennett, 2010) and were expressed as brief student perspective statements (e.g., "Our teachers ..."). Item development guidelines state that it is recommended to anchor items in a well-defined construct (Clark & Watson, 1995; DeVellis, 2017). Redundant items were eliminated, and 30 items remained in 6 subscales of five items.

Content validity. The survey draft has been evaluated by [number] specialists and underwent piloting prior to administration. [ADD item-level content validity index and the agreement procedure. Consult Lawshe, 1975; Lynn, 1986; Polit & Beck, 2006.]

Response format. Responses were measured using a five-point Likert-type scale (Likert, 1932) ranging from 1 = strongly agree to 5 = strongly disagree.

Procedure and ethics

The questionnaires were distributed to volunteer participants in classes ensuring anonymity and confidentiality, which is consistent with ethical guidelines for educational research (British Educational Research Association, 2018).

Data analysis

All analyses were conducted in SPSS. The data analysis strategy followed standard procedures described in the literature for each step. 1. Factorability. The KMO index and the Bartlett's test of sphericity were employed (Bartlett, 1954; Kaiser, 1974). KMO values of around .80 are considered good (Kaiser, 1974), while a significant Bartlett test indicates that the correlation matrix is not the identity matrix (Tabachnick & Fidell, 2019).

Extraction and Rotation

The principal component extraction with Varimax rotation and Kaiser Normalization were chosen (Kaiser, 1958). Component extraction is a data reduction procedure, rather than a common-factor solution, while the oblique factor rotation is commonly favored in case of expected correlation between dimensions (Fabrigar et al., 1999; Costello & Osborne, 2005). Thus, the findings are viewed as preliminary structural validation.

Number of components

Component retention was determined according to the eigenvalue-greater-than-one rule (Kaiser, 1960), the screen test (Cattell, 1966) and, if possible, parallel analysis (Horn, 1965).

Loading criteria.

Loadings of .50 or greater were considered practically significant, and those of .30 or greater as the lowest acceptable loading for a sample of this size (Hair et al., 2019). If an item showed cross-loading, then it would have loaded at .30 or above on a second component (Costello & Osborne, 2005).

Reliability

Internal consistency was assessed using Cronbach's alpha (Cronbach, 1951) and alpha-if-item-deleted. Interpretation of alpha values was based on the George and Mallery (2003) bands, but acknowledging that extremely high values indicate redundant items (Streiner, 2003; Tavakol & Dennick, 2011).

Group comparison

The independent-samples t-test was used to compare scores between sectors and genders. Parametric analyses of sums of Likert-type data are statistically robust in large samples (Norman, 2010). Cohen's d was calculated using means and standard deviations from published sources (Cohen, 1988; Lakens, 2013).

Results

Descriptive statistics

The data included 400 students (public university N=250; private university N=150; males N=240; females N=160). On the average, the items had a mean score of 2.429 (range 2.212 to 3.048; variance of item means = .047). For the total scale score, the mean score was 72.86 (SD = 24.237, variance = 587.414).

Factorability of the correlation matrix (Research Question 1)

The Kaiser-Meyer-Olkin index was found to be .897 – a meritorious value (Kaiser, 1974). Given that the test for sphericity was significant (chi-square (435) = 7991.42, $p < .001$), the correlation matrix is appropriate for factor analysis (Bartlett, 1954; Tabachnick & Fidell, 2019).

Component structure (Research Questions 1 and 3)

The Principal Component Analysis with Varimax rotation converged after 16 iterations and yielded six rotated components. The results are presented in Table 2, which gives the rotated component loadings below .30.

Table 2. Rotated component matrix (Varimax with Kaiser Normalization), summarized from the thesis output.

Component dimension	intended	Items and primary loadings	Secondary loadings of .30 or more
1. Tactics		IT1 .714; IT2 .789; IT3 .831; IT4 .719; IT5 .618	IT5 on C2 .460
2. Skills		IS1 .545; IS2 .624; IS3 .686; IS4 .716; IS5 .687	IS1 on C3 .511; IS2 on C3 .422; IS5 on C5 .313
3. Organizers		IO1 .737; IO2 .776; IO3 .783; IO4 .450; IO5 .470	IO2 on C6 -.431; IO3 on C5 .560
4. Strategies		ST2 .588; ST3 .746; ST4 .714; ST5 .776 (ST1 .410)	ST1 on C5 .660; ST2 on C5 .448; ST3 on C5 .310
5. Concepts		IC1 .696; IC2 .516; IC3 .528; IC4 .719; IC5 .760	IC2 on C1 .347, C6 .419; IC3 on C1 .401, C4 .341; IC4 on C6 .443; IC5 on C6 .374
6. Classroom management		CM1 .791; CM2 .822; CM3 .790; CM4 .742; CM5 .551	CM5 on C4 .320

The results include three important observations. First, all hypothesized dimensions were the main component of its own items, except ST1, which had a higher loadings (.660) on Component 5 than on Component 4 (.410). Second, 27 out of 30 items met the .50 loading criterion on its own component (Hair et al., 2019); the only items that did not were IO4 (.450), IO5 (.470) and ST1 (.410). Third, 14 out of 30 items have secondary loadings of .30 or above (IT5, IS1, IS2, IS5, IO2, IO3, ST1, ST2, ST3, IC2, IC3, IC4, IC5, CM5), which means there is some overlap between adjacent dimensions.

Reliability (RQ2, RQ3)

The Cronbach's alpha for the 30-item scale was .936 (.943 if calculated using standardized items), which is very good based on conventional cut-offs (George & Mallery, 2003). The alpha-if-item-deleted ranged from .924 (IT5) to .944 (IO2). Only IO2 deletion would increase alpha above the alpha for the total scale, and this item had a negative secondary factor load (.431 for Factor 6), thus making it the first candidate for revision or exclusion. The most impactful item on alpha was IT5 because deleting it resulted in the lowest alpha (.924). When alpha is higher than .90, it may mean that there are redundant items in the scale (Streiner, 2003; Tavakol & Dennick, 2011); however, as alpha presupposes equal weights of items, McDonald's omega is suggested as its alternative (McNeish, 2018).

4.5 Group comparisons – preliminary analyses (RQ4)

No significant differences were found between scores for different sectors (M public = 73.49, SD = 25.113; M private = 71.81, SD = 22.571; $t(397) = 0.672$, $p = .502$). The effect size calculated based on these statistics was negligible ($d = 0.07$) (Cohen, 1988; Lakens, 2013). No significant differences were also found between scores for genders (M male = 72.68, SD = 25.102; M female = 73.28, SD = 22.897; $t(398) = -0.243$, $p = .808$; $d = -0.02$). The lack of mean differences does not mean that the measure performs equally in different groups, as this needs to be tested through measurement invariance (Putnick & Bornstein, 2016).

Discussion

These findings provide initial evidence in favor of a multidimensional and internally consistent student-rated measure of instructional intelligence. The fact that both the Kaiser-Guttman criterion was met and

the Bartlett Test showed significance means that there is sufficient shared variance for structural analyses to take place (Kaiser, 1974; Bartlett, 1954). The six factors largely replicate the dimensions generated from the model (Bennett & Rolheiser, 2001). This aligns with the larger body of knowledge regarding the student evaluation process, which notes that teaching evaluations are multidimensional as opposed to monolithic (Marsh, 1984; Marsh & Roche, 1997).

It is important to give cross-loadings more than their statistical significance. Instructional intelligence is based on the integration of tactics, skills, strategies, organizers and concepts (Bennett & Rolheiser, 2001; Bennett, 2010). Students cannot clearly demarcate them in terms of observation of lessons. Thus, it is possible that the cross-loading of the strategy and concepts (ST1-ST3 on Component 5) is true, but it could also be an effect of how the questions were written. It is recommended that cognitive interviews be conducted with students when revising the test (DeVellis, 2017). While the very good alpha facilitates reliability of scores for research purposes, a value above .90 in 30 items suggests some redundancies, so a reduced version may be considered without any loss of content (Streiner, 2003; Tavakol & Dennick, 2011). While the negligible percentage of the sector category and the differences in gender are noteworthy, the results cannot be generalized to the overall public and private universities due to conflation of sectors and institutions because only one institution represents each sector.

Implications

The scale is useful for formative feedback and professional discussion of teaching as long as all necessary cautions regarding student evaluation of teaching are taken into account (Marsh & Roche, 1997; Sporeen et al., 2013).

Limitations and Future Research

- Analysis method. Principal component analysis using orthogonal rotation does not take into account measurement error and dimensions of one teaching model are expected to be correlated; exploratory factor analysis with oblique rotation should be considered (Fabrigar et al., 1999; Costello & Osborne, 2005).
- There is no confirmatory test. The structure was investigated and not confirmed in another sample independently; confirmatory factor analysis with appropriate criteria must be performed (Brown, 2015; Hu & Bentler, 1999).
- Evidence of validity is inadequate. Convergent, discriminant, criterion-related and test-retest evidence was not gathered (Messick, 1995; Mokkink et al., 2010).
- Common method variance. The data were all obtained from one self-reporting source in one point in time (Podsakoff et al., 2003).
- Measurement level. The Likert-type items were assumed to be approximately at the interval level, which is valid for large sums of such items, yet it is still good to check that by ordinal methods (Norman, 2010).
- Context. Only two institutions from one province have been involved in this study; multilocational replication and invariance tests are suggested (Putnick & Bornstein, 2016).

Conclusions

The study provides tentative support for a six-factor structure of a student-rated 30-item scale measuring instructional intelligence with very good internal consistency. The items IO2, IO4, IO5, and ST1 require improvement. Confirmation of construct validity, convergence, and invariance evidence from independent samples prior to practical application of the instrument is advised (AERA et al., 2014).

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.
- Bartlett, M. S. (1954). A note on the multiplying factors for various chi-square approximations. *Journal of the Royal Statistical Society: Series B*, 16(2), 296–298. <https://doi.org/10.1111/j.2517-6161.1954.tb00174.x>
- Bennett, B. (2002). Instructionally intelligent ... socially smart. *Orbit*, 32(4), 1–5.
- Bennett, B. (2010). The artful science of instructional integration. In R. J. Marzano (Ed.), *On excellence in teaching* (pp. 65–91). Solution Tree Press.
- Bennett, B., & Rolheiser, C. (2001). *Beyond Monet: The artful science of instructional integration*. Bookation.
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, Article 149. <https://doi.org/10.3389/fpubh.2018.00149>
- British Educational Research Association. (2018). *Ethical guidelines for educational research* (4th ed.).
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
- Clark, L. A., & 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>
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10(7), 1–9. <https://doi.org/10.7275/jvj1-4868>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, 30(1), 92–101. <https://doi.org/10.5334/irsp.82>
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.
- Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265–275. <https://doi.org/10.1037/0033-2909.103.2.265>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187–200. <https://doi.org/10.1007/BF02289233>
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–151. <https://doi.org/10.1177/001316446002000116>
- Kaiser, H. F. (1974). An index of factor simplicity. *Psychometrika*, 39(1), 31–36. <https://doi.org/10.1007/BF02291575>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1–55.
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–385. <https://doi.org/10.1097/00006199-198611000-00017>
- Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52(1), 77–95. <https://doi.org/10.1111/j.2044-8279.1982.tb02505.x>
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707–754. <https://doi.org/10.1037/0022-0663.76.5.707>
- Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, 52(11), 1187–1197. <https://doi.org/10.1037/0003-066X.52.11.1187>
- Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. ASCD.
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, 19(4), 539–549. <https://doi.org/10.1007/s11136-010-9606-8>
- Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5), 625–632. <https://doi.org/10.1007/s10459-010-9222-y>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>

- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497. <https://doi.org/10.1002/nur.20147>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education*, 16(2), 129–150. <https://doi.org/10.1080/03075079112331382944>
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642. <https://doi.org/10.3102/0034654313496870>
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18
- Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Tomlinson, C. A. (2001). *How to differentiate instruction in mixed-ability classrooms* (2nd ed.). ASCD.
- Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806–838. <https://doi.org/10.1177/0011000006288127>