

**DEVELOPMENT OF A DIAGNOSTIC TEST INSTRUMENT TO MEASURE
HIGHER ORDER THINKING SKILLS ON RESPIRATORY SYSTEM MATERIAL
BASED ON SUSTAINABLE HEALTH ISSUES**

Syafri Musthofa¹, Nur Kuswanti², Mochammad Ichsan³

¹²³Department of Biology Education / State University of Surabaya

E-mail: 1syafri.22050@mhs.unesa.ac.id, 2nurkuswanti@unesa.ac.id,

3mochammadichsan@unesa.ac.id

ABSTRACT

The persistently low achievement of Higher Order Thinking Skills (HOTS) among Indonesian students, as reflected in PISA and TIMSS results, is partly attributable to the absence of test instruments specifically designed to diagnose weaknesses in analyzing (C4), evaluating (C5), and creating (C6), rather than merely identifying factual misconceptions. This study aimed to develop a HOTS diagnostic test instrument on the respiratory system topic contextualized within sustainable health issues (PM2.5 air pollutin, smoking-related pathophysiology, and communicable respiratory diseases), and to examine its content validity, construct validity, reliability, and the resulting HOTS profile of Grade XI students at Nation Star Academy Surabaya. This Research and Development (R&D) study employed the ADDIE model and produced three parallel instrument sets (Set A, B, and C), each comprising six items (two multiple-choice items for C4, two multiple choice multiple answer items for C5, and two structured essay items for C6) that synthesized the Merdeka Curriculum Fase F Learning Outcomes with the Cambridge AS & A Level Biology Learning Outcomes. A limited field trial was conducted with 50 Grade XI Science students. Results showed very strong content validity (mean Aiken's V = 0.907). Construct validity, examined through corrected item-total correlations, sub-scale internal consistency, and exploratory factor analysis, provided initial but not fully confirmed evidence: the creating dimension (C6) proved the most empirically stable (item-total r up to 0.781), while analyzing (C4) and evaluating (C5) showed sensitivity to item format and stimulus heterogeneity (case specificity). Internal consistency reliability was acceptable for formative diagnostic purposes (Cronbach's Alpha = 0.623; McDonald's Omega = 0.642). The HOTS profile showed a declining gradient across cognitive levels: analyzing (C4) was high (72.00%), evaluating (C5) was moderate (46.67%), and creating (C6) was low (39.83%), indicating a gap in knowledge transfer toward argumentative evaluation and evidence-based creative problem formulation. The instrument is recommended as a formative diagnostic tool for teachers designing differentiated instruction, with the caveat that additional items per cognitive dimension and stimulus homogeneity are needed in further development to strengthen construct validity evidence.

Keywords: ADDIE; diagnostic test instrument; HOTS; respiratory system; sustainable health issues

ABSTRAK

Rendahnya pencapaian Keterampilan Berpikir Tingkat Tinggi (HOTS) di kalangan siswa Indonesia, sebagaimana tercermin dalam hasil PISA dan TIMSS, sebagian disebabkan oleh tidak adanya instrumen tes yang dirancang khusus untuk mendiagnosis kelemahan dalam menganalisis (C4), mengevaluasi (C5), dan menciptakan (C6), bukan hanya sekadar mengidentifikasi kesalahpahaman faktual. Studi ini bertujuan untuk mengembangkan instrumen tes diagnostik HOTS pada topik sistem pernapasan yang dikontekstualisasikan dalam isu kesehatan berkelanjutan (polusi udara PM2.5, patofisiologi terkait merokok, dan penyakit pernapasan menular), serta untuk menguji validitas isi, validitas konstruk, reliabilitas, dan profil HOTS siswa kelas XI di Nation Star Academy Surabaya. Studi Penelitian dan Pengembangan (R&D) ini menggunakan model ADDIE dan menghasilkan tiga set instrumen paralel (Set A, B, dan C), masing-masing terdiri dari enam item (dua item pilihan ganda untuk C4, dua item pilihan ganda dengan banyak jawaban untuk C5, dan dua item esai terstruktur untuk C6) yang mensintesis Hasil Pembelajaran Kurikulum Merdeka Fase F dengan Hasil Pembelajaran Biologi Cambridge AS & A Level. Uji coba lapangan terbatas dilakukan dengan 50 siswa Sains kelas XI. Hasil menunjukkan validitas isi yang sangat kuat (rata-rata Aiken's V = 0,907). Validitas konstruk, yang diperiksa melalui korelasi item-total yang dikoreksi, konsistensi internal sub-skala, dan analisis faktor eksploratori, memberikan bukti awal tetapi belum sepenuhnya terkonfirmasi: dimensi penciptaan (C6) terbukti paling stabil secara empiris (r item-total hingga 0,781), sedangkan analisis (C4) dan evaluasi (C5) menunjukkan sensitivitas terhadap format item dan heterogenitas stimulus (spesifisitas kasus). Reliabilitas konsistensi internal dapat diterima untuk tujuan diagnostik formatif (Alpha Cronbach = 0,623; Omega McDonald = 0,642). Profil HOTS menunjukkan gradien menurun di seluruh tingkat kognitif: menganalisis (C4) tinggi (72,00%), mengevaluasi (C5) sedang (46,67%), dan menciptakan (C6) rendah (39,83%), menunjukkan kesenjangan dalam transfer pengetahuan menuju evaluasi argumentatif dan formulasi masalah kreatif berbasis bukti. Instrumen ini direkomendasikan sebagai alat diagnostik formatif untuk guru yang merancang pembelajaran berdiferensiasi, dengan catatan bahwa item tambahan per dimensi kognitif dan homogenitas stimulus diperlukan dalam pengembangan lebih lanjut untuk memperkuat bukti validitas konstruk.

Kata Kunci: ADDIE; instrumen tes diagnostik; HOTS; sistem pernapasan; isu kesehatan berkelanjutan

A. Introduction

Twenty-first century education demands a fundamental transformation in learning practice, in line with national and international curriculum policies that foreground the holistic development of essential student competencies. This paradigm shift drives a move from teacher-centered toward student-centered learning, with a focus on developing critical, creative, collaborative, and communicative thinking (the 4Cs) as prerequisites for success in an era of information and technological disruption (Hujjatusnaini et al., 2022). Indonesia's Merdeka Curriculum explicitly emphasizes meaningful and adaptive learning, which is best achieved when content is connected to real problems such as sustainable health issues. Within this context, Higher Order Thinking Skills (HOTS) covering analyzing (C4), evaluating (C5), and creating (C6) in the revised Bloom's Taxonomy become an inseparable element of learning (Anderson & Krathwohl, 2001).

Nevertheless, Indonesian students' HOTS achievement generally remains at a concerning level. PISA 2022 placed Indonesia 70th of 81 countries in science

literacy, with a mean score of 383, far below the OECD average of 485 (OECD, 2023). TIMSS 2019 results likewise indicate that most Indonesian students can only answer low-cognitive-level (knowing) items, with only a small proportion reaching the reasoning level that requires analysis and evaluation (Mullis et al., 2020). This low achievement is closely tied to a fundamental problem in the assessment instruments in use: most available instruments are still designed to measure factual and rote knowledge (C1–C2) and therefore cannot accurately capture students' analysis, evaluation, or creation abilities (Alkhatib, 2022). The absence of adequate diagnostic instruments leaves teachers without sufficient data to identify students' specific cognitive weaknesses at each HOTS level, so that instructional interventions become poorly targeted and less effective (Wiliam, 2018).

The urgency of developing a HOTS-based diagnostic test instrument is reinforced by its connection to sustainable health issues. The respiratory health domain offers complex, multidimensional real-world problems that make it an ideal stimulus for HOTS testing. Globally,

approximately 4.2 million premature deaths per year are associated with ambient air pollutant exposure (World Health Organization, 2022), while Indonesia records a very high burden of acute respiratory infection, particularly among school-age children. Relevance of health issues as a diagnostic stimulus lies in their capacity to force students to transfer knowledge applying respiratory system concepts to unfamiliar real situations (Brookhart, 2010). When a student is asked to analyze acute respiratory infection prevalence data in an industrial city, evaluate the effectiveness of air quality policy based on scientific evidence, or design an intervention grounded in biological understanding, that student demonstrates all three HOTS levels simultaneously, generating rich diagnostic data on cognitive strengths and weaknesses that rote items cannot reveal.

A review of the literature on diagnostic instrument development reveals a clear research gap. Existing diagnostic instruments, such as three-tier multiple-choice tests, are able to identify factual-conceptual misconceptions but are not designed to diagnose weaknesses in higher-

order thinking processes such as data analysis, argument evaluation, and solution creation. Other studies have produced valid and reliable HOTS instruments that are summative rather than diagnostic in function, or that touch only a single HOTS dimension without an authentic health-data stimulus. Consequently, no instrument yet exists that is simultaneously diagnostic, grounded in real health-issue context, and capable of comprehensively measuring all HOTS levels (C4–C6) among high school students.

Based on preliminary observation at Nation Star Academy Surabaya a school that implements Cambridge International AS & A Level Biology alongside the Merdeka Curriculum the test instruments in current use still focus predominantly on factual knowledge (C1–C2), which is not fully aligned with the Cambridge assessment's emphasis on higher cognitive skills (C4–C6). The potential of the respiratory system topic to be linked with sustainable health issues such as air pollution impact, respiratory infection epidemiology, and smoking-related chronic disease risk has therefore not been optimized to map and develop students' HOTS.

This study was conducted to fill that gap through the development of a diagnostic test instrument to measure HOTS on the respiratory system topic based on sustainable health issues, designed to have strong content and construct validity, adequate reliability, and the capacity to generate an informative, actionable HOTS diagnostic profile for teachers' instructional planning.

Based on this background, the study addressed the following research questions: (1) how was the HOTS-based diagnostic test instrument on the respiratory system topic, contextualized within sustainable health issues, developed; (2) what is the content validity of the instrument based on expert judgment; (3) what is the construct validity of the instrument based on field-trial data; (4) what is the reliability of the instrument; and (5) what is the HOTS profile of Grade XI students at Nation Star Academy Surabaya based on the instrument's implementation.

B. Methods

This study is a Research and Development (R&D) study employing the ADDIE model (Analysis, Design, Development, Implementation,

Evaluation), selected for its systematic, iterative approach and its suitability for developing a diagnostic test instrument that requires thorough needs analysis, structured blueprint-based design, staged validation, and formative evaluation at every stage.

The study was conducted over five months (February–July 2026). The Analysis and Design stages, covering curriculum analysis, needs analysis, formulation of the diagnostic construct, blueprint development, and drafting of the instrument and scoring rubric, were carried out at the Biology Education Study Program, Universitas Negeri Surabaya (UNESA). The Development stage, focused on item construction, expert review, and instrument refinement, was likewise conducted at UNESA. The Implementation stage consisted of a limited field trial at Nation Star Academy Surabaya, involving Grade XI Science students as the research subjects, selected purposively to evaluate the instrument's validity and reliability under authentic learning conditions. The Evaluation stage, again conducted at UNESA, analyzed the instrument's psychometric quality and finalized the product.

In the Analysis stage, curriculum analysis examined the coherence between the Merdeka Curriculum Fase F Learning Outcomes, the Cambridge AS & A Level Biology Learning Outcomes (AO2 and AO3 for Gas Exchange), and relevant health-literacy competencies. Student analysis found that although students already possessed adequate factual knowledge of the respiratory system, their ability to analyze health data, evaluate scientific arguments, and formulate evidence-based solutions had not been systematically measured due to the absence of a suitable diagnostic instrument. Concept analysis identified four priority subtopics: the macroscopic and microscopic structure of the respiratory tract, gas-exchange mechanisms at the alveolus, mucociliary escalator function and its disruption by smoking, and ventilation regulation in the context of chronic pollutant exposure.

construct was operationally defined as three distinguishable latent cognitive factors analyzing (C4), evaluating (C5), and creating (C6) when confronted with complex, non-routine, health-issue-based stimuli. A test blueprint was developed based on

the two-dimensional revised Bloom's Taxonomy (Anderson & Krathwohl, 2001), crossing cognitive process (C4, C5, C6) with knowledge dimension (conceptual and procedural). Each parallel set comprises six items distributed proportionally: two items targeting C4, two targeting C5, and two targeting C6, with one conceptual- and one procedural-knowledge item per cognitive-level pair. Health-issue stimuli particulate air pollution (PM2.5), smoking-related pathophysiology, and communicable respiratory disease (tuberculosis and acute respiratory infection) were selected based on conceptual relevance and data authenticity, sourced from WHO (2022) and Ministry of Health Indonesia reports. Three item formats were purposively designed to target distinct HOTS levels: multiple-choice items for C4, multiple choice multiple answer (MCMA) items for C4–C5, and structured essay items with analytic rubrics for C6.

In the Development stage, item construction (Draft I) referenced the blueprint directly, attending to content, construction, and language quality (Arikunto, 2019). Internal review by the researcher and supervisor

produced Draft II, which was then evaluated by an expert panel comprising a biology education specialist and an educational evaluation expert using a four-point Likert validation sheet (1 = not valid to 4 = very valid) covering content, construction, language, and stimulus-relevance aspects. Expert feedback led to Draft III, deemed ready for field testing.

In the Implementation stage, Draft III was trialed on 50 Grade XI Science students at Nation Star Academy Surabaya under controlled, standardized conditions, distributed across the three parallel sets (Set A: $n = 17$; Set B: $n = 17$; Set C: $n = 16$). Each student completed only one set, following a parallel-forms design intended to minimize practice effects and item exposure. Data collected included per-item student responses, used for construct-validity and reliability analysis, and independent double-rater scores on the essay items, used for inter-rater reliability analysis.

In the Evaluation stage, formative evaluation was embedded throughout the Development and Implementation stages, while summative evaluation comprised a

comprehensive psychometric analysis: content validity via Aiken's V , construct validity via corrected item-total correlation and Exploratory Factor Analysis (EFA; Principal Axis Factoring extraction, Oblimin rotation, eigenvalue > 1 retention, loading threshold ≥ 0.40), internal-consistency reliability via Cronbach's Alpha and McDonald's Omega, and inter-rater reliability via the Intraclass Correlation Coefficient (ICC, Two-Way Mixed Effects, Absolute Agreement) and weighted Kappa. Given the sample-to-item ratio ($50:6 = 8.3:1$, below the recommended minimum of 5:1 for stable EFA per set and 20:1 for confirmatory analysis), EFA results were treated as indicative and exploratory rather than confirmatory (Hair et al., 2019). The HOTS profile was analyzed descriptively, with achievement categorized as high ($\geq 70\%$), moderate (40–69%), or low ($< 40\%$) following Arikunto (2013).

C. Result and Discussion

Instrument Development Process

The ADDIE-based development process produced three equivalent parallel instrument sets (Set A, B, C), each with six mixed-format items (two multiple-choice/C4, two MCMA/C5,

two structured essay/C6), integrating Merdeka Curriculum Fase F Learning Outcomes with Cambridge AS & A Level Biology Learning Outcomes and using sustainable health issues (PM2.5 pollution, smoking pathophysiology, tuberculosis) as authentic stimuli for both conceptual and procedural knowledge.

Content Validity

Content validity was evaluated by two expert validators from the Biology Education field at Universitas Negeri Surabaya, who reviewed all 18 items across content, construction, and language aspects (five indicators each) using a four-point relevance scale. The aggregated Aiken's V coefficients across the three sets ranged from 0.833 to 1.000, yielding an overall mean of 0.907 well above the 0.80 acceptance threshold (Azwar, 2018; Retnawati, 2016) indicating very strong content validity evidence for all indicators assessed.

Qualitative expert feedback led to three key restructuring actions before field testing. First, a monodomain principle was applied to operational verbs: items previously requiring double-barreled cognitive demands (e.g., "analyze and conclude") were revised so each item

isolated a single, unambiguous cognitive level, preventing diagnostic imprecision about which process a student failed at. Second, visual stimuli (graphs and clinical data displays) were redesigned at higher resolution with clearer axis labeling to reduce extraneous cognitive load (Sweller et al., 2019), consistent with expert comments that stimulus graphics needed sharper presentation. Third, item stems were syntactically pruned to remove wordy, non-essential background narrative that risked shifting the construct measured toward reading comprehension rather than biological reasoning (Messick, 1995).

Construct Validity

Construct validity was evaluated across three complementary layers: per-set corrected item-total correlation, per-dimension internal consistency on pooled data ($n = 50$), and Exploratory Factor Analysis (EFA) on pooled data. Using a threshold of $r \geq 0.30$ (Hair et al., 2019), the essay items (C6) consistently showed the highest item-total correlations across all three sets ($r = 0.419\text{--}0.781$), while MCMA items (C5) showed moderate, variable correlations ($r = 0.133\text{--}0.459$), and multiple-choice items (C4)

generally showed the lowest correlations ($r = 0.178\text{--}0.517$). This gradient reflects the mathematical consequence of variance restriction in dichotomous items rather than a genuine failure to measure the intended construct (Reckase, 2009); two C4 items in Set B additionally showed a ceiling effect ($P = 0.824$ and 0.882), which further attenuates item-total correlation without indicating construct weakness.

Per-dimension internal consistency (pooled $n = 50$) likewise showed a level-dependent gradient: C4 ($\alpha = 0.480$, weak), C5 ($\alpha = 0.341$, very weak), and C6 ($\alpha = 0.660$, approaching adequate). The exceptionally low C5 consistency was attributable in part to stimulus heterogeneity between the two MCMA items, which addressed different health contexts (PM2.5 exposure versus smoking-related mucociliary damage), introducing case-specificity variance beyond the intended construct.

EFA using Principal Axis Factoring with Oblimin rotation on pooled data ($KMO = 0.573$, "mediocre" but adequate for exploratory purposes; Bartlett's $\chi^2(15) = 53.813$, $p < 0.001$) retained three factors with

eigenvalues of 2.147, 1.354, and 1.030, jointly explaining 75.51% of variance. However, the extraction did not reach full convergence (> 25 iterations, final convergence = 0.002), and the resulting pattern matrix showed only partial alignment with the hypothesized three-dimension structure: Factor 1 was dominated by both essay items and one MCMA item, Factor 2 by the second MCMA item, and Factor 3 by both multiple-choice items, with essay item 2 showing substantial cross-loading. Inter-factor correlations were low to moderate ($r = -0.070$ to 0.264), supporting the oblique rotation choice while indicating the three factors are empirically distinguishable but not fully independent.

Taken together, the evidence converges on the creating dimension (C6) as the most empirically well-defined construct in the instrument supported by the highest item-total correlations, the highest sub-scale consistency, dominant EFA loadings, and excellent-to-good inter-rater agreement ($ICC = 0.980$ and 0.885). Differentiation between C4 and C5, in contrast, is not yet strongly evidenced; low C4 correlations largely reflect format and ceiling-effect limitations

rather than construct failure, while the low C5 consistency reflects stimulus heterogeneity. The construct validity of this instrument should therefore be characterized as initial, indicative evidence rather than fully confirmed, warranting caution in interpreting individual-level C4 and C5 scores while supporting confident use of the C6 sub-scale.

Reliability

Internal-consistency reliability on pooled data ($n = 50$, six items) yielded Cronbach's Alpha = 0.623 and McDonald's Omega = 0.642, both categorized as acceptable for a formative diagnostic instrument (Taber, 2018). These coefficients are defensible given the short test length, mixed dichotomous/polytomous item format, and inherent multidimensionality of the construct, all of which mathematically attenuate internal-consistency estimates without indicating poor instrument quality (Cortina, 1993; Reckase, 2009; Hayes & Coutts, 2020). Inter-rater reliability on the two essay items was excellent to good, with ICC (3,1) = 0.980 (essay 1) and 0.885 (essay 2), weighted Kappa in close agreement (difference < 0.002), and exact agreement rates of 96.0% and 90.0% respectively,

confirming that the strong C6 psychometric evidence reflects genuine student ability rather than single-rater bias.

Students' HOTS Profile

Analysis of the 50 field-trial students' performance revealed a declining achievement gradient across the three HOTS dimensions: analyzing (C4) was high at 72.00% of maximum score, evaluating (C5) dropped to moderate at 46.67%, and creating (C6) fell to low at 39.83%, with an overall HOTS achievement of 45.10% (moderate), as summarized in Table 1. This gradient is consistent with cognitive load theory and the hierarchical structure of the revised Bloom's Taxonomy: students were highly capable of deconstructing factual and graphical information (C4) but struggled increasingly as cognitive demand shifted toward convergent argument evaluation (C5) and, most sharply, toward generating original, evidence-based interventions (C6). This pattern indicates that while the Cambridge-aligned curriculum has successfully built students' analytical logic, it has not yet adequately cultivated divergent, creative thinking, an essential component of functional health literacy (Nutbeam, 2008).

Table 1. HOTS Achievement Profile by Cognitive Dimension

HOTS Dimen sion	Ma x Sco re	Me an Sco re	Achieve ment (%)	Categ ory
Analyzi ng (C4)	2	1.4 4	72.00	Hight
Evaluat ing (C5)	6	2.8 0	46.67	Moder ate
Creatin g (C6)	12	4.7 8	39.83	Low
Total (Overall HOTS)	20	9.0 2	45.10	Moder ate

D. Conclusion

This study developed a HOTS diagnostic test instrument for the respiratory system topic, framed around sustainable health issues, using the ADDIE model. The instrument comprising three parallel, mixed-format sets integrating Merdeka Curriculum and Cambridge AS & A Level Biology Learning Outcomes demonstrated very strong content validity (Aiken's $V = 0.907$) following expert review and structural refinement. Construct validity evidence was strongest for the creating (C6) dimension, which was consistently well-defined across item-total correlation, internal consistency, and factor analysis, while the analyzing (C4) and evaluating (C5) dimensions showed sensitivity to item format and stimulus heterogeneity,

indicating initial rather than fully confirmed construct validity. Internal-consistency reliability (Cronbach's $\alpha = 0.623$; McDonald's $\Omega = 0.642$) and inter-rater reliability for essay items (ICC up to 0.980) were acceptable to excellent for formative diagnostic use. The resulting HOTS profile of Grade XI students showed a declining gradient high in analyzing (72.00%), moderate in evaluating (46.67%), and low in creating (39.83%) pointing to a specific instructional need for scaffolding the transfer of analytical knowledge into evaluative and creative health-related problem solving. Teachers are encouraged to use this profile to design differentiated instruction emphasizing Socio-Scientific Issues and Problem-Based Learning approaches, while future instrument development should expand the number of items per cognitive dimension, homogenize stimuli within item pairs, and pursue confirmatory testing on a larger sample (minimum $n = 200$) to strengthen the construct validity evidence base.

REFERENCE

Book:

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning,*

- teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Arikunto, S. (2019). *Dasar-dasar evaluasi pendidikan* (Edisi revisi). Jakarta: Bumi Aksara.
- Azwar, S. (2018). *Reliabilitas dan validitas* (Edisi 4). Yogyakarta: Pustaka Pelajar.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Andover: Cengage Learning.
- Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Boston: Pearson.
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. Paris: OECD Publishing.
<https://doi.org/10.1787/53f23881-en>
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.
- Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian (Panduan peneliti, mahasiswa, dan psikometrikawan)*. Yogyakarta: Parama Publishing.
- William, D. (2018). *Embedded formative assessment* (2nd ed.). Bloomington, IN: Solution Tree Press.
- World Health Organization. (2013). *Health literacy: The solid facts*. Copenhagen: WHO Regional Office for Europe.
- World Health Organization. (2022). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. 2021 update*. Geneva: World Health Organization.
- Journal Article:**
- Alkhatib, O. J. (2022). An effective assessment method of higher-order thinking skills (problem-solving, critical thinking, creative thinking, and decision-making) in engineering and humanities. In *2022 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1–6). IEEE.
- Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139–148.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104.
- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *Communication Methods and Measures*, 14(1), 1–24.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting

- intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4), 212–218.
- 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.
- Nutbeam, D. (2008). The evolving concept of health literacy. *Social Science & Medicine*, 67(12), 2072–2078.
<https://doi.org/10.1016/j.socscimed.2008.09.050>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292.
- 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.