RESEARCH ARTICLE

Validity and reliability of concept inventory test in human physiology

Gio Jay B. Aligway[a,1,*], Jo C. Delos Angeles [b,2], Angeli V. Collano[c,3], Eljoy P. Barroca[d,4], Anna Clarissa D. Aves[e,5], Juneflor F. Catubay [f,6], Jennifer T. Edjec [e,7], Ma. Diana A. Butaya [g,8], Sylvester T. Cortes[h,9,*],

[a] City of Naga Integrated Center for Science Technology Culture and Arts High School, City of Naga, Cebu, Philippines

[b] Science in Biology Department, College of arts and sciences, Cebu Doctors' University, North Reclamation Area, Mandaue City, Philippines

[c] Talamban National High School, Department of Education Region 7, Cebu City, Philippines

[d] Western Mindanao State University, Normal Road, Baliwasan, Zamboanga City, Philippines

[e] Dolores National High School, Ormoc City, Philippines

[f] Professional Academy of the Philippines, City of Naga, Cebu, Philippines

[g] Luray II National High School, Toledo, City, Philippines

[h] Cebu Technological University, Cebu City, Philippines

[1]giojay.aligway@deped.gov.ph; [2]jodelosangeles120@gmail.com; [3]angelicollano@gmail.com; [4]eljaybarroca@gmail.com; [5]annaclarissa.aves@deped.gov.ph; [6]jfcatubay08@gmail.com; [7]jenniferedjec@gmail.com; [8]madiana.butaya071@deped.gov.ph; [9]sylvester.cortes@ctu.edu.ph

For correspondence:
sylvester.cortes@ctu.edu.ph

Article history:
Received: 07 October 2023
Revised: 01 January 2024
Accepted: 24 March 2024
Published: 30 March 2024

10.22219/jpbi.v10i1.29558

p-ISSN: 2442-3750
e-ISSN: 2537-6204

How to cite: Aligway, G. J. B., Angeles, J. C. D., Collano, A. V., Barroca, E. P., Aves, A. C. D., Catubay, J. F., Edjec, J. T., Butaya, M. D. A., & Cortes, S. T. (2024) Validity and reliability of concept inventory test in human physiology. *JPBI (Journal Pendidikan Biologi Indonesia), 10*(1), 273-282. https://doi.org/10.22219/jpbi.v10i1.29558

Abstract: Biology education plays a vital role in nurturing the understanding of learners about the intricacy of life. Various efforts have emerged to strengthen learning biological concepts but there were still studies that showed that learners have low mastery in some aspects. To determine how well students understood various biological topics, including human physiology, Concept inventory tests (CIT) were used. The concept inventory test may be able to spot students' misconceptions and ultimately lead to improved comprehension. The crafted CIT developed with the aid of a table of specifications based on Bloom's taxonomy of cognitive domain was assessed according to its validity and reliability. In validation, content validity and item analysis were considered while reliability test was employed through Cronbach's alpha. Distractor analysis was also performed to determine possible source of misconception per item. The CIT was administered to 120 senior high school STEM students (50.8% from the private schools, 37.5% in regular public schools and 11.7% from public schools with special programs in science). The results displayed high content validity with a mean of 4.83 for content validity and an average Aiken's validity coefficient of 0.98. It also highlighted that the test is moderately difficult with the test difficulty of 0.58, as well as, discriminatory with a discriminating level of 0.46. After item classification, 63 items were retained (39 accepted, 24 for revisions) and Chronbach's alpha (α=0.74) indicated good internal consistency. The concept inventory test propounds to be a good classroom test with minor items to be revised.

Keywords: biology education; concept inventory test; human physiology; test development; test validation

Introduction

Biology learning, which explores the intricate web of life, is an essential part of one's education (Allawan, 2021). The biology curriculum covers numerous topics such as cells, genetics, ecosystems and human physiology (Kaptan & Timurlenk, 2012). This provides students with knowledge and scientific skills which play crucial roles in nation building (Villarino & Villarino, 2023). Recognizing its vital role in education, numerous initiatives have emerged to strengthen the curriculum through innovation in pedagogical approach and technology integration (Safitri et al., 2017). Other effort is promoting STEM-focused project-based learning. It is intended to help students expand their content knowledge while honing their global and 21st century abilities (O'Connor & Hite, 2017).

The K-12 program in the Philippines offers General Biology 1 and 2 as a specialized subject for Science, Technology, Engineering and Mathematics (STEM) students in Senior High School to keep up with the biology education on a global platform. It gears towards deeper understanding and appreciation of life processes at the cellular level previously introduced in Grade 7-10 through spiral progression (Antipolo & Rogayan, 2021; Amin & Zubaidah, 2017). This serves as a preparatory path for students who opt to take healthcare, medical or other biology related courses. However, despite the efforts of the curriculum to equip students with holistic biological concepts, several factors had hindered its promising objective. In the study of Kaptan & Timurlenk (2012), inadequate educational materials, teachers and laboratory facilities were some of the challenges. Moreover, Großschedl et al.(2014), also highlighted that learners have low mastery in biology due to factors which include knowledge, intelligence and motivation.

Quality assessment is also needed to improve the quality of education including biology. According to UNESCO IIEP, assessment provides a variety of purposes ranging from, describing students learning, providing career guidance, motivating students and certifying level of competence. Assessment could also be a springboard to help educators and policy-makers in improving the curriculum (Kellaghan & Greaney, 2001). Thus, it is important that assessment must have a robust framework that ensures validity, reliability, and comprehensive coverage to promote a fair assessment process (Rezai, 2022).

Different approaches in assessment are salient in biology education because it contributes evidence in teaching effectiveness and student learning (Tanner & Allen, 2004; Agboghoroma & Oyovwi, 2015). While the probability of quality assessment is common in all areas of biology, human physiology is not exempted. Quality assessment is essential in human physiology education to foster critical thinking, practical application, and prepare students for medical and research careers (Michael, 2007; Ghazivakili et al., 2014).

In assessing student's level of proficiency of the various concepts in biology such as human physiology, concept inventory test had been utilized (Cary et al., 2019). It is composed of multiple-choice questions with plausible distractors. This is essential because it is highly efficient in administering and scoring as well as it provides objective evaluation and standardized method of assessment. Thus, studying the reliability and validity of concept inventory test valuable tool for assessing the impact of educational interventions, curriculum revisions, or instructional strategies on students' learning outcomes in human physiology (Lieu & Shaffer, 2018). Moreover, developing multiple-choice assessments with high standards for concept inventory may be able to identify pupils' misunderstandings and ultimately result in higher comprehension (Engelhardt, 2009).

Hence, this study aims in developing a valid and reliable concept inventory test for human physiology that can be used as diagnostic, summative or formative assessment as well as a misconception or placement test. In addition, the developed concept inventory test focused on assessing STEM Senior High School students in their conceptual knowledge and misconceptions on Human Physiology.

Method

This study focused on the assessment of validity and reliability of the crafted concept inventory test (CIT) in Human Physiology. The validation includes Content Validity through the checklist adopted from (Morales, 2003) supported with Aiken's validity test. Further validation included item analysis which have two features namely item difficulty and item discrimination. On the other hand, Chronbach's Alpha was used to determine the reliability of the CIT. Moreover, this study followed three phases.

Phase 1 Preparation

The crafting of the CIT was initiated by carefully selecting and collating topics under Human Physiology. Different literature and sources (Scanlon & Sanders, 2018; Silverthorn et al., 2023) including DepEd's General Biology most essential learning competencies were taken into account.

Phase 2 Development

Based on the topics from phase 1, a table of specifications with Bloom's revised taxonomy were crafted as a guide for the development of the CIT. The original version (v.1) contained 60% of cognitive domain items that supported Lower-Order Thinking Skills (LOTS) like remembering and understanding and 40%

of cognitive domain items supporting Higher-Order Thinking Skills (HOTS) like application, analysis, evaluation, and creation. The percentages of each order thinking skills were marginally based on the studies of (Amalia & Wahyuni, 2021; Winanda & Anwar, 2022). The initial iteration produced a multiple-choice concept test with 100 items (v.1). In the Philippines, typical final and standard exams often have 80 to 100 item formats. These were the considerations in the development of version 1 of the test. Revisions were also made after validation of the experts.

Phase 3 Validation

The initial constructed test was validated by two biology educators who were masters in their field and one medical practitioner who has a strong background of biology and human physiology. Recommendations and suggestions were collated and taken into account for the construction of version 2 (v.2).

The 100-item CIT was revised pursuant to the validators' comments and remarks. The test's second version (v. 2), which still had 100 items, underwent validation from the same validators a second time. Version 2 was administered to 120 Grade 11 and 12 students that had already undertaken the specialized subject, General Biology 1 and 2. Respondents were represented by 50.8% from the private schools, 37.5% in regular public schools and 11.7% from public schools with special programs in science.

To determine the quality of the test, validation and reliability tests were done. Distractor analysis was also performed for the evaluation of the students' selection of each item's response. This is done by analyzing a fraction of the entire respondents' test scores. In order to have an adequate representation, the upper 27 % who belong to the higher score group and the lower 27% who belong to the lower score group were taken into consideration (Morales, 2012). The frequency of students who chose each response is then calculated and tabulated for each item.

Results and Discussion

Test development

In the construction of the multiple-choice CIT, 20 topics in Human Physiology were considered. Topics were distributed equally with 5 items each. These items were categorized based on Bloom's taxonomy of cognitive domain, wherein remembering and understanding items were tagged as Lower-Order Thinking Skills (LOTS) while applying, analyzing, evaluating and creating items were tagged as Higher-Order Thinking Skills (HOTS) (Morales, 2012).

As shown in table 1, about 40% of the constructed questions in version 1 belong to higher-order thinking skills (HOTS) while 60% belong to lower-order thinking skills (LOTS). The third column in table 1 indicates the percentage distribution of items for each cognitive domain. It shows that remembering items has the highest percentage with 36%. Conversely, creating items were 4% and denoted as the least among the rest.

Table 1. Percentage Distribution of Items Based on Bloom's Taxonomy of Cognitive Domain (v.1)

Bloom's Taxonomy of Cognitive Domain	No. of Items	Percentage out of 100	Percentage of LOTS and HOTS
LOTS			
Remember	37	36%	
Understand	23	24%	60%
HOTS			
Apply	12	11%	
Analyze	18	18%	
Evaluate	6	7%	
Create	4	4%	40%
Total	**100**	**100**	

Moreover, in the development of this test, four choices format was followed. The choices had one right response and three distractors. Although other standardized tests have less than or more than 4 choices, the latter format is prevalent in standardized and teacher-made tests in the Philippine educational context (Morales, 2012). It also pointed out that the 4-choices format is even used in Trends in International

Mathematics and Science Study (TIMSS) as well as in National Achievement Test (NAT) which is one of the standard assessments of Philippine Department of Education for secondary levels.

Table 2. Content Validity of Version 1 (v.1)

	Expert 1	Expert 2	Expert 3
Mean	4.25	3.9	3.0
Comments	1. "Use negatively stated items stem only when learning outcomes are required." 2. "Use special alternatives such as none of the above, or none of these sparingly" 3. "Arranging options in a vertical column makes reading easier" 4. "When writing item stems and options, try to avoid lifting verbatim from the text/ reference. I suggest paraphrasing"	1. "There are items that need to be revised by changing it into an interrogative form." 2. "Most of the choices need to be arranged appropriately."	1. "Choices in several items need to be arranged according to length or alphabetical order." 2. "It's better not to have "all of the above" in the choices. Instead, write other plausible distractors or use a 'table form' distractor."
Over-all Mean	**3.72 out of 5**		

Version 1 and 2 of the CIT were validated by three experts who are two biology educators and one medical practitioner acquainted in the academy. As shown in table 2, the overall mean of the content validity of version 1 is 3.72 out of 5. This suggests that its content validity is between Undecided to Agree (Appendix A). Thus, further improvements of the inventory test should be made and it was clearly coinciding in their descriptive validation. The recurring theme of their comments centered on the arrangement of the choices. Moreover, some also highlighted the avoidance of all of the above and none of the above choices and suggestion to paraphrase or revise stem questions or the options. These were taken into consideration for the revision of the inventory test.

Table 3. Content Validity of Version 2 (v.2)

	Expert 1	Expert 2	Expert 3
Mean	**5**	**4.75**	**4.75**
Comments	No comment	The questions are well framed according to the objectives. The choices of the questions are the areas that need to be improved by arranging it from fewer terms to longer terms or from smallest to greatest.	Minor revisions only
Over-all Mean	**4.83 out of 5**		

The redesigned test (v.2) was subjected to a second round of content and face validation (Table 3). Version 2 still consists of 100 items after being revised based on the initial validation round. The same experts rated it with an overall mean of 4.83 out of 5. This exemplifies a consequential improvement from the first validation round. The over-all mean also denotes that the content validity of version 2 is between Agree to Strongly Agree (Appendix A). This indicates a good quality test in terms of construction and content validity (Morales, 2012).

Moreover, to verify that the inventory test is rated as content valid, the content validity coefficient was determined based on the checklist. This was accomplished using Aiken's content validity coefficient (Aiken, 1985). The closer the coefficient is to one, the more content validity an item has. The average content validity coefficient of version 1 is at 0.68. In comparison to the version 2 which had an average of 0.95, version 2 is much closer to one, it can be seen at Table 4. Thus, the experts who rated the items thought the items in version 2 were legitimate in terms of content. All of the items on the checklist in version 2 were rated close to one, indicating a high content validity coefficient.

Table 4. Content Validity Coefficient

Checklist Items	Aiken's V (Content Validity Coefficient)	
	v.1	v.2
1	*0.67*	1.00
2	0.75	1.00
3	0.75	0.90
4	0.67	1.00
5	0.75	1.00
6	0.67	0.90
7	0.67	0.90
8	0.67	1.00
9	0.67	1.00
10	0.67	0.90
11	0.67	0.90
12	0.67	0.80
13	0.67	1.00
14	0.75	0.90
15	0.5	0.90
16	0.75	1.00
17	0.75	0.90
18	0.5	1.00
19	0.67	1.00
20	0.75	1.00
Average	**0.68**	**0.95**

Item analysis

Item categorization based on difficulty and discrimination indices was included in the item analysis. This delivers substantial information to teachers for supplemental item customization and future exam development, as well as educational resources to help them (Siri & Freddano, 2011). Furthermore, only version 2 was subjected to the item analysis since it was the version administered to the STEM senior high school students.

Item difficulty

Item difficulty or the difficulty index, is the percentage of takers who correctly answered the given item divided by the total number of test takers which is 120. As described in Table 5, the difficulty classification of the test items was based on Morales (2003). Difficulty index of each item were calculated and classified as very easy (VE), easy (E), moderately difficult (MD), difficult (D) and very difficult (VD).

Table 5. Item Difficulty of Version 2 (v.2)

Item Difficulty Interpretation	Item Difficulty Range	Number of Items	Percentage
Very Easy (VE)	0.81 – above	14	14%
Easy (E)	0.61 – 0.80	28	28%
Moderately Difficult (MD)	0.41 – 0.60	39	39%
Difficult (D)	0.21 – 0.40	16	16%
Very Difficult (VD)	0.00 – 0.20	3	3%
Total		**100**	**100%**
Test Difficulty		**0.58**	

As reflected in Table 5, 39% of the test items were classified as moderately difficult (MD), 28% were easy (E) and 16% were difficult (D). Subsequently, 14% of the items were tagged as very easy (VE) and 3% were very difficult. The calculated test difficulty index averaged at 0.58. It means that the version 2 of the inventory test comes into the category of moderately difficult.

Item discrimination

In order to attain item discrimination, the students were ranked according to the lowest to highest test scores. Both upper and lower 27% were determined as the basis for the analysis of the groups. Item discrimination is the difference between the upper discrimination index and the lower discrimination index. Table 6, shows the summary through percentages of all item discrimination index classified into ranges along with the interpretations.

Table 6. Item Discrimination of Version 2 (v.2)

Item Discrimination Interpretation	Range	Number of Items	Percentage
Questionable (Q)	-1.00 - -0.60	0	0%
Not Discriminating (ND)	-0.59 – 0.09	9	9%
Moderately Discriminating (MD)	0.10 - 0.20	3	3%
Discriminating (D)	0.21 – 0.60	74	74%
Very Discriminating (VD)	0.61 – 1.00	14	14%
Total		**100**	**100%**
Discriminating level		**0.46**	

According to the discriminatory indices (Table 6), the majority of the items were classified as discriminating (D) with a result of 74% of the 100 items. On the other hand, 14% of the items were found to be very discriminating (VD) while moderately discriminating (MD) and not discriminating (ND) have 3% and 9%, respectively. Furthermore, there were no questionable (Q) items identified.

Furthermore, coupling the difficulty index with the discrimination index identified the decision whether items are accepted, revised or rejected (Table 7). The decision table was adopted from Morales (2012) with slight modification.

Table 7. *Item Classification Based on Item Analysis of Version 2 (v.2)*

	Number of Items	Percentage
Accept	39	39%
Revise	24	24%
Reject	37	37%
Total	**100**	**100%**

The percentage distribution of the 100-item multiple-choice CIT (Table 7) revealed that 39% were accepted, 24% needed modification, and 37% were rejected. The final version (v.3) of the CIT had 39 accepted items and 24 suggested to revise items. It connotes a total of 63 items of the Human Physiology CIT for STEM Senior High School.

Reliability

Reliability is referred to as the assessment results' consistency. It can show how consistently test results or data sets perform across applications or over time. Several types of reliability testing had been used but internal consistency reliability is best suited for a test that is only given once (Morales, 2012; Nolan & Hecker, 2021). It measures the reliability of the individual items in a test.

Table 8. Reliability Statistics (Chronbach's Alpha) of Version 2 (v.2)

Number of Items	Chronbach's Alpha
100	0.73

The computed reliability Chronbach's Alpha was 0.73 (Table 8). It Is interpreted as good according to University of Washington. It means that the most items of the CIT are suitable for a classroom test however few items could be improved. This consistency coefficient offers a good degree to which related items in the assessment are actually measuring the same concept (Caffrey, 2011). Hence, the constructed CIT exhibits a fair reliability in assessing student's conceptual knowledge in Human

Physiology.

Distractor Analysis

The sample of the Distractor analysis is constituted in Table 9. In item # 16, students from the upper and lower group frequently answered the correct answer which is letter B. In contrast, two distractors which are A & C pose a misconception among the students. Two students from the upper group answered A while five students from the lower group also answered A. These students have a misconception about Facilitated Diffusion since they mistakenly connect its concept to Endocytosis and Exocytosis. It shows that it is a good constructed item because it has several plausible distractors (Cheung & Bucat, 2002).

Table 9. Sample Distractor Analysis

Item No.	Upper (n=32)				Lower (n=32)				Remarks
	A	B	C	D	A	B	C	D	
16	**2	*30	**0	0	**5	*17	**7	3	A & C poses a misconception
17	1	*5	0	**26	6	*4	3	**19	D poses a misconception
18	2	0	*29	**1	6	5	*13	**8	D poses a misconception
19	**4	*26	2	0	**20	*10	0	2	A poses a misconception
20	3	*2	*23	**4	5	6	*8	**13	D poses a misconception

**- correct answer*

***- probable source of misconception*

Interestingly, item # 17 (Fig.1) has a different case wherein both the upper and lower group have high frequency of opting D which is a distractor. A total of 45 students from both groups answered D while 9 students chose B which is believed to be the correct answer. Upon reviewing the item, the alternatives revealed to have an error. The right response should be "D. Facilitated Diffusion". It was a slight technical error that was overlooked by both researchers and the experts. Nonetheless, this error actually coincides with the decision of the item analysis which suggest that item # 17 should be rejected.

17. Which passive transport method needs the support of particular channel or carrier proteins?

A. Exocytosis
B. Endocytosis
C. Simple diffusion
D. Facilitated diffusion

Figure 1. Sample item (#17).

19. Which of the following processes occurs to water inside a cell when immersed in a hypertonic solution?

A. Into the cell
B. Out of the cell
C. No net movement
D. No movement at all

Figure 2. Sample Item (#19).

Contrary, item # 19 (Fig. 2) exhibits a highly plausible distractor because 20 students from the lower group answered A which is not the right response. It is stated that plausible distractors should be able to draw in more than 5% of the low-performing students who were unable to choose the right response (Shin et al., 2019).Therefore, A is a plausible distractor that caused a misconception of osmosis in a hypertonic solution among lower group students. In relation, osmosis concepts call for students to conceptualize and consider chemical processes at the molecular level, which may make them challenging for certain students to grasp its concept (Oztas, 2014). As a result, one can determine the student's alternative idea of a certain topic being measured by the item from the distractor analysis.

Conclusion

The validity and reliability of concept inventory test offers refinement of the test items that can assess student's conceptual knowledge of Human physiology. Inferring information about the students' conceptions and misconceptions from the presented process will enable educators and teachers to create a similar kind of assessment. Also, this study showed that distractor analysis does not only provide the identification of misconceptions of students but it could also serve as a tool in feedbacking of the test that had been administered. Therefore, peculiar results in distractor analysis make provision for test revisits and improvements. Overall, the concept inventory test is recommended to be used as a classroom test. The items in the concept inventory test could also serve as test banks for teachers especially those accepted items. However, continual development and refinement of this test can make this viable for standardized exam. It is suggested to increase the number of respondents to observe if the internal consistency of the items will be similar. Additionally, a comparable effort may be made to develop concept inventory tests for other themes in biology, such Cell Biology, Botany, Zoology, Genetics and Ecology.

Acknowledgment

Sincere appreciation to Cebu Technological University for their outstanding assistance and cooperation with research projects. Their dedication to creating a supportive environment for study has been crucial to the accomplishment of our research undertaking. This collaboration has greatly benefited the research experience as well as the larger academic and scientific community. We look forward to continuing our partnership with CTU as we pursue knowledge and research excellence.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contributions

G. J. B. Aligway: Performed data analysis and interpretation as well as writing and editing manuscripts; **J. C. D. Angeles:** Methodology and Implemented test validation; **A. V. Collano:** Supervised the crafting of concept inventory test and contributed to written manuscript; **E. P. Barroca:** Wrote original draft preparation; **A. C. D. Aves:** Organized the collected data and contributed to written manuscript; **J. F. Catubay:** Administered concept inventory test and collected initial data; **J. T. Edjec:** Contributed to research components and concept inventory test; **M. D. A. Butaya:** Administered concept inventory test; and **S. T. Cortes:** Performed data analysis and finalized the manuscript.

References

Agboghoroma, T. E., & Oyovwi, E. O. (2015). Evaluating effect of students' academic achievement on identified difficult concepts in senior secondary school biology in Delta State. *Journal of Education and Practice*, *6*(30), 117–125. https://eric.ed.gov/?id=EJ1081378

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Allawan, J. G. (2021). Exploring the Factors Influencing the Performance in Biology of Senior High School Students at University of Mindanao, Philippines. *International Journal of Research and Innovation in Social Science*, *05*(06), 352–359. https://doi.org/10.47772/IJRISS.2021.5618

Amalia, R. F., & Wahyuni, S. (2021). Analysis of Higher Order Thinking Skills (HOTS) Content of SBMPTN Physics Problems. *Journal of Physics: Conference Series*, *1918*(5), 052055. https://doi.org/10.1088/1742-6596/1918/5/052055

Amin, A. M., Corebima, A. D., & Zubaidah, S. (2017). The critical thinking skills profile of preservice biology teachers in animal physiology. *Proceedings of the 3rd International Conference on Education and Training*, *128*(179–183). https://doi.org/10.2991/icet-17.2017.30

Antipolo, A. M. R., & Rogayan, D. V. Jr. (2021). Filipino prospective teachers' experiences in teaching in K12 science curriculum: A cross-sectional research. *Jurnal Pendidikan Biologi Indonesia*, *7*(1), 1–10. https://doi.org/10.22219/jpbi.v7i1.15468
Caffrey, E. D. (2011). *Assessment in elementary and secondary education: A primer.* DIANE Publishing. https://books.google.com/books
Cary, T. L., Wienhold, C. J., & Branchaw, J. (2019). A Biology Core Concept Instrument (BCCI) to Teach and Assess Student Conceptual Understanding. *CBE—Life Sciences Education*, *18*(3), ar46. https://doi.org/10.1187/cbe.18-09-0192
Cheung, D., & Bucat, R. (2002). *How can we construct good multiple-choice items?.* https://www.academia.edu/download/40386455/constructMC.pdf
Engelhardt, V. (2009). An introduction to classical test theory as applied to conceptual multiple-choice tests: Getting started in physics education research. *American Journal of Physics*, *2*(1), 1–40. https://www.per-central.org/items/detail.cfm?ID=8807
Ghazivakili, Z., Norouzi Nia, R., Panahi, F., Karimi, M., Gholsorkhi, H., & Ahmadi, Z. (2014). The role of critical thinking skills and learning styles of university students in their academic performance. *Journal of Advances in Medical Education & Professionalism*, *2*(3), 95–102. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4235550/
Großschedl, J., Mahler, D., Kleickmann, T., & Harms, U. (2014). Content-Related Knowledge of Biology Teachers from Secondary Schools: Structure and learning opportunities. *International Journal of Science Education*, *36*(14), 2335–2366. https://doi.org/10.1080/09500693.2014.923949
Kaptan, K., & Timurlenk, O. (2012). Challenges for Science Education. *Procedia - Social and Behavioral Sciences*, *51*, 763–771. https://doi.org/10.1016/j.sbspro.2012.08.237
Kellaghan, T., & Greaney, V. (2001). *Using assessment to improve the quality of education.* Unesco, International Institute for Educational Planning. https://unesdoc.unesco.org/ark:/48223/pf0000126231
Lieu, R. M., Gutierrez, A., & Shaffer, J. F. (2018). Student perceived difficulties in learning organ systems in an undergraduate Human Anatomy Course. *Journal of the Human Anatomy and Physiology Society*, *22*(1), 84–92. https://doi.org/10.21692/haps.2018.011
Michael, J. (2007). What makes physiology hard for students to learn? Results of a faculty survey. *Advances in Physiology Education*, *31*(1), 34–40. https://doi.org/10.1152/advan.00057.2006
Morales, M. P. E. (2003). *Development and validation of a two-tier test in Natsci 13 (Ecology)*. De La Salle-College of Saint Benilde. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3884103
Morales, M. P. E. (2012). Development and Validation of a concept test in introductory physics for biology students. *The Manila Journal of Science*, *7*(2), 26–44. https://ejournals.ph/article.php?id=2441
Nolan, M. M., Beran, T., & Hecker, K. G. (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal*, *11*(2), 103–123. https://doi.org/10.52041/serj.v11i2.333
O'Connor, B., & Hite, R. (2017). Global Learning Using Biology PBL: A Texas-China Collaboration in Middle Grade Genetics. *Journal of Interdisciplinary Teacher Leadership*, *2*(2), 17–26. https://doi.org/10.46767/kfp.2016-0016
Oztas, F. (2014). How do High School Students Know Diffusion and Osmosis? High School Students' Difficulties in Understanding Diffusion & Osmosis. *Procedia - Social and Behavioral Sciences*, *116*, 3679–3682. https://doi.org/10.1016/j.sbspro.2014.01.822
Rezai, A. (2022). Fairness in classroom assessment: development and validation of a questionnaire. *Language Testing in Asia*, *12*(1), 17. https://doi.org/10.1186/s40468-022-00162-9
Safitri, M., Riandi, R., Widodo, A., & Nasution, W. R. (2017). Integration of Various Technologies in Biology Learning. *Journal of Physics: Conference Series*, *895*, 012145. https://doi.org/10.1088/1742-6596/895/1/012145
Scanlon, V. C., & Sanders, T. (2018). *Essentials of anatomy and physiology.* FA Davis. https://books.google.co.id/books/about/Essentials_of_Anatomy_and_Physiology.html?id=oXAotAEACAAJ&redir_esc=y
Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-Choice Item Distractor Development Using Topic Modeling Approaches. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00825
Silverthorn, D., Cafferty, P., Casagrand, J., Co, E., Flemming, M., McFarland, J., O'Loughlin, V., Scott, D., & Tomicek, N. (2023). Introducing the HAPS Physiology Learning Outcomes. *HAPS Educator*, *27*(1), 79–86. https://doi.org/10.21692/haps.2023.010
Siri, A., & Freddano, M. (2011). The Use of Item Analysis for The Improvement Of Objective Examinations. *Procedia - Social and Behavioral Sciences*, *29*, 188–197. https://doi.org/10.1016/j.sbspro.2011.11.224

Tanner, K., & Allen, D. (2004). Approaches to Biology Teaching and Learning: From Assays to Assessments—On Collecting Evidence in Science Teaching. *Cell Biology Education*, *3*(2), 69–74. https://doi.org/10.1187/cbe.04-03-0037

Villarino, R. T., & Villarino, M. (2023). *Academic Performance of Rural Junior High School Students in Biology: Basis for Learning Activities Development*. *4*, 1–10. https://dergipark.org.tr/en/download/article-file/2928056

Winanda, G. Z., & Anwar, D. (2022). Analysis of the implementation higher order thinking skills in tasks and test of english subject at SMK Negeri 2 Padang. *Advances in Social Science, Education and Humanities Research,* https://doi.org/10.2991/assehr.k.220201.019