Work in Progress: Evaluating the Effect of Symbolic Problem Solving on Testing Validity and Reliability

Dr. Yan Tang, Embry-Riddle Aeronautical University, Daytona Beach

Dr. Yan Tang is an associate professor of mechanical engineering at Embry-Riddle Aeronautical University in Daytona Beach, Fla. Her current research in engineering education focuses on cognitive load theory, deliberate practice, and effective pedagogical strategies.

Lin Ding, The Ohio State University

Lin Ding, Ph.D., is an associate professor in the Department of Teaching and Learning at The Ohio State University. Dr. Dingâ€™s scholarly interests lie in discipline-based STEM education research. His work includes theoretical and empirical investigation

Dr. Haiyan Bai, University of Central Florida

Haiyan Bai, PhD., is Professor of Quantitative Research Methodology in the College of Community Innovation and Education at the University of Central Florida. Her interests include resampling method, propensity score analysis, research design, measurement

Dr. Richard Catrambone, Max Planck Institute for Intelligent Systems

Richard Catrambone is a Professor in the School of Psychology at the Georgia Institute of Technology. He received his B.A. from Grinnell College and his Ph.D. in Experimental Psychology from the University of Michigan.

The question Catrambone likes to ask–and the thread that runs through the projects he does alone and in collaboration with others–is: What does someone need to know in order to solve novel problems or carry out tasks within a particular domain?

Catrambone's research interests include problem solving, educational technology, and human-computer interaction. He is particularly interested in how people learn from examples in order to solve problems in domains such as algebra, probability, and physics. He explores how to create instructional materials that help learners understand how to approach problems in a meaningful way rather than simply memorizing a set of steps that cannot easily be transferred to novel problems. He researches the design of teaching and training materials–including software and multimedia environments–based on cognitive principles that help students learn basic tasks quickly and promote transfer to novel problems. He uses task analysis to identify what someone needs to know in order to solve problems or carry out tasks in a domain and then to use the results of the task analysis to guide the construction of teaching and training materials/environments.

Catrambone has served on the Cognitive Science Society governing board from 2011-2016 and was chair of the Society in 2015. He was co-chair of the Cognitive Science Conference in 2010. He has served as a consulting editor for the Journal of Educational Psychology (1/2008 - 12/2011), the Journal of Experimental Psychology: Learning, Memory, and Cognition (6/2000 - 12/2001 and 1/2009 - 12/2009), the Journal of Experimental Psychology: Applied (1/2001 - 12/2007), and the Journal of Experimental Psychology: General (6/2000 - 12/2001). He has published his research in journals such as the Journal of Experimental Psychology: General; Journal of Experimental Psychology: Learning, Memory, and Cognition; Journal of Experimental Psychology: Applied; Memory & Cognition; Journal of Educational Psychology; Human-Computer Interaction; Human Factors; and other basic and applied journals. He has also served on grant review panels for a variety of funding agencies including the National Science Foundation and the Institute of Education Sciences (U.S. Department of Education).

Evaluating the Effect of Symbolic Problem Solving on Assessment Validity and Reliability

Abstract

Problem-solving is a typical type of assessment in engineering dynamics tests. To solve a problem, students need to set up equations and find a numerical answer. Depending on its difficulty and complexity, it can take anywhere from ten to thirty minutes to solve a quantitative problem. Due to the time constraint of in-class testing, a typical test may only contain a limited number of problems, covering an insufficient range of problem types. This can potentially reduce validity and reliability, two crucial factors which contribute to assessment results.

A test with high validity should cover proper content. It should be able to distinguish high-performing students from low-performing students and every student in between. A reliable test should have a sufficient number of items to provide consistent information about students' mastery of the materials.

In this work-in-progress study, we will investigate to what extent a newly developed assessment is valid and reliable. Symbolic problem solving in this study refers to solving problems by setting up a system of equations without finding numeric solutions. Such problems usually take much less time. As a result, we can include more problems of a variety of types in a test.

We evaluate the new assessment's validity and reliability. The efficient approach focused in symbolic problem-solving allows for a diverse range of problems in a single test. We will follow Standards for Educational and Psychological Testing, referred to as the *Standards*, for our study. The Standards were developed jointly by three professional organizations including the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). We will use the standards to evaluate the content validity and internal consistency of a collection of symbolic problems. Examples on rectilinear kinematics and angular motion will be provided to illustrate how symbolic problem solving is used in both homework and assessments.

Numerous studies in the literature have shown that symbolic questions impose greater challenges because of students' algebraic difficulties. Thus, we will share strategies on how to prepare students to approach such problems.

Introduction

In engineering dynamics tests, students' problem-solving skills are commonly assessed through open-ended items. To solve a problem, students need to set up equations and find a numerical answer. Depending on its difficulty and complexity, it can take anywhere from ten to thirty minutes to solve a quantitative problem. Due to time constraints, a conventional in-class test can include only a limited number of problems with an insufficient number of problem types. This can potentially reduce the validity and reliability of the test, two crucial factors which contribute to assessment results.

As stated in the *Standards*, validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." [1] When evaluating validity, we need to find the quantity and quality of evidence that supports interpreting test scores for an intended purpose [2]. As emphasized in the *Standards*, validity is considered "a unitary concept" rather than being composed of distinct categories (e.g., content validity or construct validity) [1, 2]. The *Standards* provides five types of evidence that support the proper interpretations of test scores [1]. For this study, the most pertinent source of evidence is validity evidence based on test content. Such evidence can be acquired by analyzing the relationships between the content of a test and knowledge and skills the test is intended to assess.

Another important indicator of test quality is reliability [3-6]. Reliability represents a test's consistency. Taking a test of high reliability, students should get the consistent score even when they take it on different days if students' knowledge and mental states remain the same. The general guideline for increasing reliability is to add more items because we could gather more information of students' learning to ensure consistency in scoring [1, 2]. But the tradeoffs is time.

Table 1 An Example of Conventional Problem Solving and Symbolic Problem Solving

<table>
<tr><td colspan="2">Example: The acceleration of a regional airliner during its takeoff run is $a = 14 - 0.0003v^2$ ft/s^2, where v is its velocity in ft/s. How long does it take the airliner to travel 3000 ft? How fast is the airliner moving when $s = 3000$ ft? [4]</td></tr>
<tr><td>Conventional Solution</td><td>Proposed Symbolic Solution (see Appendix A for the full expected solution)</td></tr>
<tr><td>$$\int_0^{3000} ds = \int_0^{v} \frac{v}{14 - 0.0003v^2}\, dv \Rightarrow$$
$$\boxed{v = \sqrt{\frac{14(1 - e^{-2\cdot 0.0003\cdot 3000})}{0.0003}} = 197.4\ \text{m/s}}$$
$$\int_0^{t} dt = \int_0^{v} \frac{1}{14 - 0.0003v^2}\, dv \Rightarrow$$
$$\boxed{t = \frac{1}{2\sqrt{14\cdot 0.0003}} \ln \frac{14 + v\sqrt{14\cdot 0.0003}}{14 - v\sqrt{14\cdot 0.0003}} = 23.9\ \text{s}}$$</td><td>$$\int_{t_0}^{t_1} dt = \int_{v_0}^{v_1} \frac{1}{a(v)}\, dv$$
$$\int_{s_0}^{s_1} ds = \int_{v_0}^{v_1} \frac{v}{a(v)}\, dv$$
Two unknowns: (t_1 and s_1)
Two equations</td></tr>
</table>

An example of rectilinear motion can help illustrate how conventional problem-solving might compromise validity and reliability. Table 1 shows the conventional solution in the left column and the proposed symbolic solution in the right column. The conventional solution presents equations with their numerical answers, while the symbolic approach provides solely the equations with unknowns denoted. Nevertheless, for most students, integrating equations as shown in the example can be a significant challenge, possibly taking up to 30 minutes to find the numerical answer. This implies that in a 50-minute test, only two or three problems can be incorporated, potentially leaving some aspects of the content unaddressed. A lack of content coverage can result in insufficient evidence to support valid interpretations of test scores. Furthermore, construct irrelevance can also impact test scores, particularly since students often allocate significant time to calculations, and capabilities such as taking integrals may be unrelated to the test's intended purpose. In addition to the aforementioned validity concerns, the reliability could be diminished due to the limited number of problems included.

In contrast, the symbolic solution entails setting up equations without calculations, as exemplified in the right column of Table 1. This approach allows for a more diverse range of content, mitigating the risk of content underrepresentation. Furthermore, since students need only to set up equations, construct irrelevance is minimized. Additionally, incorporating more items in the test can enhance its reliability.

As illustrated in the above example, the proposed symbolic approach has the potential to improve validity and reliability. By skipping time-consuming calculations, we could include more problem types to gather more information on how students solve problems. We intend to use this article to share our practice on improving validity and reliability in a practical manner such that instructors who are interested but with little knowledge of psychometrics could easily implement the process in their test design. The article is organized as follows. We will first introduce the test design and development process to delineate the considerations of evidence for validity. Then we show how we develop items followed by the evaluation of the validity and reliability of the test followed by the results.

Test Design and Development

Test design is a process of developing questions or tasks to measure students' knowledge and skill [1]. A test plan delineates the steps and considerations along with specifications for test administration and scoring procedures for this process. In this section, we will demonstrate how to develop a test plan by taking validity and reliability into account. To develop a test plan, one must first consider the intended use of the test scores and the expected interpretations that will arise from them. Subsequently, the test's content and format are carefully determined to ensure that the resulting evidence supports the intended interpretations for their respective purposes. Test items are then created based on the test specifications and are evaluated against the criteria to ensure proper use of the test. Additionally, procedures for scoring individual items and the entire test are established, reviewed, and modified as necessary. Typically, test design is an iterative process (Figure 1), and adjustments are made based on data from tryouts and operational use [1]. Upon completing the iterative process, a question bank is created for use in rotations to prevent question leak among students. While the test plan presented in this article focuses on rectilinear motion and angular motion, it can be easily extended to other learning outcomes and other subjects.

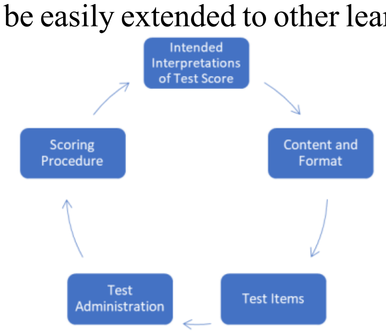


Figure 1 A schematic of a test development process

Intended Interpretations of Test Scores

In our study, we aimed to assess students' knowledge and skills in solving rectilinear and angular motion problems. Specifically, we sought to understand how students apply integrals and time derivatives correctly. This focus was based on students' knowledge deficiencies revealed in a prior knowledge assessment. The prior knowledge assessment was administered in two sections of Engineering Dynamics at Embry-Riddle Aeronautical University (ERAU), a southeastern US private institution. Over 60% of students did not represent linear functions $v(t)$ and $v(s)$ correctly from graphs, and over 70% of students failed to recognize that the kinematics equation $v = v_0 + a \cdot t$ is applicable only to motion with constant acceleration. No students could determine the first-order time derivative of the radial position function $r(\theta) = 2\cos\theta$ m. The results were consistent with findings on students' limited understanding of functions reported in [7, 8]. Such knowledge deficiencies present significant challenges to solving rectilinear motion problems where the acceleration is dependent on t, v, or s. Additionally, many students face difficulties in learning curvilinear motion with normal and tangential components due to unfamiliarity with kinematics relationships between θ, ω, and α. Owing to the similar problem-solving strategies between angular and rectilinear motion, incorporating angular motion problems can help students gain fluency with the kinematics relationships in angular motion while improving their understanding and skills in integrating functions with different functional dependencies.

Content, Format, and Test Items

Before selecting problems, we employed a test blueprint to outline the test at a high level to define what we intend to measure [2, 3]. This practice not only assisted us in creating the test, but also provided supporting evidence for validity.

Table 2 The Test Blueprint for Rectilinear and Angular Motion

Topics	Percentage	Total # of Items	Cognitive Level		
			Low	Mid	High
$\boldsymbol{a(t)}$	18%	2	0	2	0
$\boldsymbol{a(v)}$	9%	1	0	0	1
$\boldsymbol{\alpha(\omega)}$	18%	2	1	1	0
$\boldsymbol{a(s)}$	27%	3	1	0	2
$\boldsymbol{\alpha(\theta)}$	27%	3	1	1	1
Total	100%	11	3	4	4

As shown in the test blueprint (Table 2), the test comprises 11 problems across five types. All these problems aim to assess the use of definite integrals to solve rectilinear and angular motion problems. Each topic in Column 1 represents the given (angular) acceleration along with functional dependence. For example, $\alpha(\omega)$ in Row 3 represents one problem type in which the angular acceleration is given as a function of angular velocity. To solve this type of problem, students need to integrate the angular acceleration to find the angular velocity at a given instant or the time to attain a specific angular velocity. The cognitive levels are related to the difficulty level for students. Generally, problems which require more steps and higher skill levels are at a higher cognitive level. For example, setting up the equation $\alpha(t)dt = d\omega$ has a lower cognitive level than setting up the equation $dt = \frac{1}{\alpha(\omega)} d\omega$ as the latter takes one more step to apply separation of variables in $\alpha = \frac{d\omega}{dt}$; setting up $dt = \frac{1}{\alpha(\omega)} d\omega$ is easier than $d\theta = \frac{\omega}{\alpha(\omega)} d\omega$ as the latter requires the application of the chain rule, which many students find challenging to master. A task analysis was conducted to identify the cognitive levels [9]. See the sample problems in Appendix B for different problem types and cognitive levels.

Test Administration
Numerous studies have indicated that symbolic problem-solving presents significant challenges for students [7, 10-12]. To prevent cognitive overload and ensure learning success, it is crucial to offer support and prompt feedback. Consequently, these problems were implemented as formative assessments to monitor student progress and provide timely feedback. The eleven problems were distributed across five weekly assessments, spanning from the fourth week to week 10.

Scoring Procedures
All submissions were graded in Gradescope, an online grading and assessment platform. Gradescope allows instructors to create customized rubrics for grading assignments. Furthermore, it provides various statistics and analytics, including overall scores, scores by question, and average grades for the class, which significantly aids in evaluating reliability in terms of internal consistency. For each problem, 80% of the score was based on the accuracy of the equations used, while 15% was assigned for correctly representing the given information using symbols in the equations, and 5% was allocated for identifying the unknown variables. See Appendix C for an example of grading in Gradescope. All assessments were graded by the course instructor to avoid incorrect grading due to graders' lack of experience. If test correction is offered, the assessments can be graded in two rounds: the first round by the grader and the second round by the course instructor for award partial credit in addition to checking grading accuracy.

Results and Discussions

Table 3 Assessment Schedule

Time	Problems and Cognitive Level
Week 4	P1(L), P2(L)
Week 6	P3(M), P4(H), P5(M), P6(L)
Week 8	P7(M), P8(M), P9(H)
Week 9	P10(H)
Week 10	P11(H)
Cognitive Level: L: Low; M: Medium; H: High	

The five assessments were administered in two sections of Engineering Dynamics at ERAU in the Fall of 2022 (see Table 3 for the assessment schedule). The letter inside the parentheses in Table 3 represents the presumed cognitive level of each problem. Figure 2 presents the scores for the 11 problems. A notable increase in learning performance was observed when comparing the scores of P3-6 in week 6 to those of P7-9 in week 8. Similarly, solving problems with the high cognitive level exhibited significant improvement from week 6 to week 8. However, no further progress was seen in solving difficult problems beyond week 8.

When evaluating the internal consistency of the formative assessments for all eleven items using Cronbach's alpha with 61 samples, the result was α = .88. The score falls in the range for Cronbach's alpha scores from .84 to .90, indicating the test is reliable using the criteria from [13].

The assessment scores from the five assessments reflect students' learning progress. As indicated in Figure 2, students' performance increased steadily from week 6 to week 8 followed by a plateau after week 8. This can be partly explained by the students' practice methods. Throughout weeks 3 and 7, students were assigned daily homework to help them develop skills for applying the symbolic problem-solving approach. The homework problems were designed following the instructional design model Four Component Instructional Design (4C/ID), which provides systematic strategies for developing learning tasks to help novices learn effectively and efficiently [9, 14]. Since the problem-solving process can be divided into two stages -- setting up equations and solving equations -- each stage requires different skills, so students need to practice with problems targeting the desired skills. When solving problems aimed at setting up equations, students can focus on applying separation of variables and the chain rule, if necessary, to integrate the equations without considering how to solve the equation. When solving equations to find numeric solution, they only need to consider how to take integrals. Such practice is designed to help a learner develop a specific skill beyond the learner's current ability [15, 16]. The benefit of this practice strategy also includes increasing the variability of problems, which facilitates learning transfer [14]. During these five weeks, students were assigned 29 problems, including 18 problems on rectilinear motion and 11 problems on angular motion. After week 8, students were not assigned any problem, which might explain why a plateau was reached. Further studies need to be conducted to explore the factors correlated with progress.

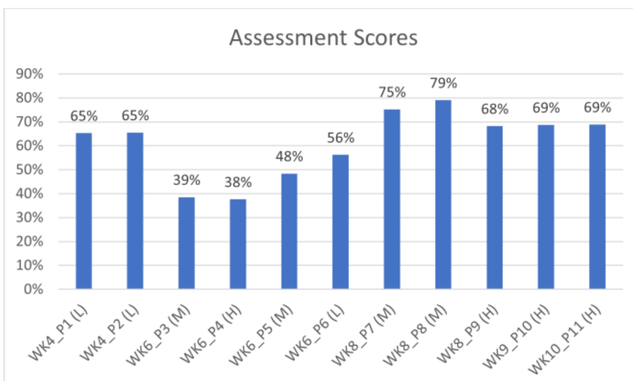


Figure 2 Assessment Results

Although the symbolic problem-solving approach can enhance validity and reliability, it presents challenges in teaching, as the literature has shown that students often struggle with symbolic questions due to algebraic difficulties [7, 10-12, 17]. To implement this approach, scaffolding is necessary to provide support and guidance, helping students develop the required skills and knowledge [14]. For instance, before students were assigned to solve problems in week 3, they were asked to only represent givens and finds and identify unknowns. This helps them become familiar with the symbolic representation of given information. Since this is not the focus of the study, we will elaborate on the design in a separate paper.

Conclusions

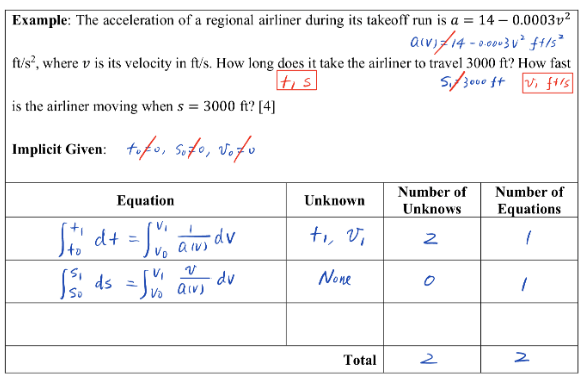
In this article, we have shared our experience in following the test development process recommended by the *Standards* for designing assessments on solving rectilinear and angular motion problems. By adopting the symbolic problem-solving approach, we were able to incorporate a more diverse range of problems, while maintaining a focus on the pertinent construct. Consequently, this approach offers the advantage of improving both validity and reliability in comparison to traditional methods. Moreover, we have highlighted the importance of providing students with effective scaffolding to improve their skills in symbolic problem-solving. These strategies not only foster a deeper understanding of the subject matter but also enable students to overcome the challenges often associated with algebraic difficulties. By combining a well-designed assessment with targeted instructional support, educators can promote deep learning effectively.

Acknowledgement

This study is supported by the National Science Foundation under Grant Nos. 1927284 and 1927290.

Appendix A: Complete Solution Format

Figure 3 illustrate the complete solution format students are expected to employ in both homework and assessments. Each problem is presented with double spacing, providing ample room for representing each piece of given information's variable beneath its corresponding statement. For a given function, students must include the function's argument to show the functional dependency, facilitating separation of variables when setting up equations. For given kinematics information at different instants, students must include a subscript to denote the corresponding instants (e.g., add the subscript "0" to s and v to represent the position and velocity at the initial instant). When examining unknowns in each equation, students are encouraged to cross out used givens to ensure that all information is presented and all unknowns are identified in each equation. The ultimate goal is to set up equal number of independent equations for the unknowns.



Example: The acceleration of a regional airliner during its takeoff run is $a = 14 - 0.0003v^2$ ft/s², where v is its velocity in ft/s. How long does it take the airliner to travel 3000 ft? How fast is the airliner moving when $s = 3000$ ft? [4]

$a(v) = 14 - 0.0003v^2$ ft/s²; t_1 s; $s_1 = 3000$ ft; v_1 ft/s

Implicit Given: $t_0 = 0$, $s_0 = 0$, $v_0 = 0$

Equation	Unknown	Number of Unknows	Number of Equations
$\int_{t_0}^{t_1} dt = \int_{v_0}^{v_1} \frac{1}{a(v)} dv$	t_1, v_1	2	1
$\int_{s_0}^{s_1} ds = \int_{v_0}^{v_1} \frac{v}{a(v)} dv$	None	0	1
	Total	2	2

Figure 3 The Full Solution Format

Appendix B: Sample Problems

Example 1 (Type: $\alpha(\omega)$; Cognitive level: mid) : The rotor of a jet engine is rotating at 10,000 rpm (revolutions per minute) when the fuel is shut off. The ensuing angular acceleration (in rad/s^2) is α=−0.02ω, where ω is the rotor's angular velocity in rad/s. How long does it take the rotor to slow to 1000 rpm?

Example 2 (Type: $a(v)$; Cognitive level: high) : A rock that's dropped from the top of a cliff will experience an acceleration due to gravity, along with a deceleration due to drag. The total downward acceleration is $g - c_d\dot{s}^2$, where g = 32.2 ft/s^2, c_d = 0.01 m^{-1}, and the downward speed $\dot{s}$ is in ft/s. If the rock is dropped from 100 ft up, calculate the rock's impact velocity.

Example 3 (Type: $\alpha(\theta)$; Cognitive level: high) : A motorcycle moves along a circular track 300-m in radius and its angular acceleration is given by $\alpha = -0.001\theta$ rad/s^2. If its speed at the initial position is 30 m/s, determine its speed and the time when it travels another quarter revolution.

Appendix C: Grading Rubric

Figure 4 shows the grading interface in Gradescope. Each marked rubric indicates an error. Since both equations were not set up correctly and three givens (i.e., t_0, s_0 v_0) were not provided, the first three rubric items were marked.

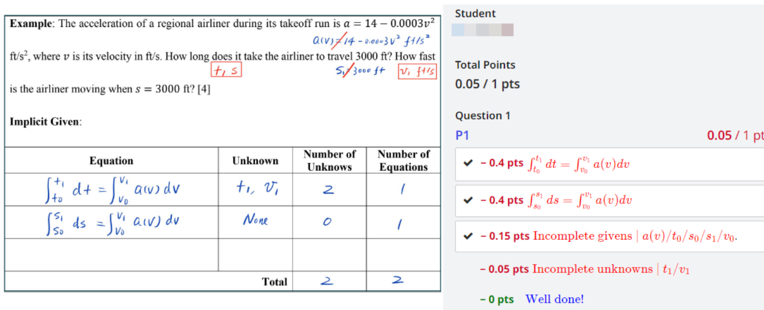


Figure 4 Grading Interface in Gradescope

References

[1] American Educational Research Association, American Psychological Association and National Council on Measurement in Education, *Standards for Educational and Psychological Testing.* American Educational Research Association, 2018.

[2] J. E. Balogh, *A Practical Guide to Creating Quality Exams.* Intelliphonics, 2016.

[3] T. Kubiszyn and G. D. Borich, *Educational Testing and Measurement.* John Wiley & Sons, 2016.

[4] P. S. Steif and J. A. Dantzler, "A statics concept inventory: Development and psychometric analysis," *J Eng Educ,* vol. 94, *(4),* pp. 363-371, 2005.

[5] D. Hestenes and I. Halloun, "Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller," *The Physics Teacher,* vol. 33, *(8),* pp. 502, 1995.

[6] L. Ding, R. Chabay, B. Sherwood and R. Beichner, "Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment," *Physical Review Special Topics.Physics Education Research,* vol. 2, *(1),* pp. 010105, 2006. . DOI: 10.1103/PhysRevSTPER.2.010105.

[7] J. M. Clark, F. Cordero, J. Cottrill, B. Czarnocha, D. J. DeVries, D. St. John, G. Tolias and D. Vidakovic, "Constructing a schema: The case of the chain rule?" *The Journal of Mathematical Behavior,* vol. 16, *(4),* pp. 345-364, 1997. . DOI: 10.1016/S0732-3123(97)90012-2.

[8] E. Redish, "Using Math in Physics: 5. Functional dependence," *The Physics Teacher,* vol. 60, *(1),* pp. 18-21, 2022. . DOI: 10.1119/5.0040055.

[9] Y. Tang, H. Hai and R. Catrambone, "Task-analysis-guided deliberate practice for learning free-body diagrams," in *ASEE Annual Conference Proceedings,* 2021, .

[10] E. Kuo, M. M. Hull, A. Gupta and A. Elby, "How students blend conceptual and formal mathematical reasoning in solving physics problems," *Science Education (Salem, Mass.); Sci.Ed,* vol. 97, *(1),* pp. 32-57, 2013. . DOI: 10.1002/sce.21043.

[11] E. T. Torigoe and G. E. Gladding, "Connecting symbolic difficulties with failure in physics," *American Journal of Physics,* vol. 79, *(1),* pp. 133-140, 2011.

[12] E. F. Redish and E. Kuo, "Language of Physics, Language of Math: Disciplinary Culture and Dynamic Epistemology," *Sci & Educ,* vol. 24, *(5-6),* pp. 561-590, 2015. . DOI: 10.1007/s11191-015-9749-7.

[13] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Res Sci Educ,* vol. 48, *(6),* pp. 1273-1296, 2018. . DOI: 10.1007/s11165-016-9602-2.

[14] J. J. Van Merriënboer and P. A. Kirschner, *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design.* Routledge, 2017.

[15] K. A. Ericsson, "The influence of experience and deliberate practice on the development of superior expert performance," *The Cambridge Handbook of Expertise and Expert Performance,* vol. 38, *(685-705),* pp. 2-2.3, 2006.

[16] K. A. Ericsson, R. T. Krampe and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance." *Psychol. Rev.,* vol. 100, *(3),* pp. 363, 1993.

[17] E. T. Torigoe, "Unpacking Symbolic Equations in Introductory Physics," 2015.