Preliminary development of a simple statistical tool for estimating mental model states from a diagnostic test

Alexander Volfson,[1] Haim Eshach,[1] and Yuval Ben-Abu[2,3,*]

[1]*Department of Science Education & Technology, Ben-Gurion University of the Negev*
[2]*Department of Physics and Project Unit, Sapir Academic College, Sderot, Hof Ashkelon 79165, Israel*
[3]*Clarendon laboratory, Department of Physics, University of Oxford, United Kingdom*

(Received 25 November 2019; revised 6 October 2020; accepted 6 August 2021; published 13 September 2021)

Science knowledge is reflected in mental models that students tend to form when dealing with science phenomena. One way to identify students' mental models about scientific concepts is the use of diagnostic tests (inventories). Even though several statistical approaches and tools intended for the analysis of such inventories' results exist in the literature, there are certain inventories for which analysis might require the development of more convenient tools. Thus, there seems to be no simple way to handle the results of inventories whose items include different numbers of statements, the number of statements relating to the same mental model within an item is not fixed, and neutral distractors are possible. This exploratory study aims at meeting this challenge and suggests a relatively simple statistic F_{2m} for estimating the mental model state of a subject or a group of subjects in the case of two mental models under consideration, providing a preliminary estimation of F_{2m} consistency using the results of the sound concept inventory instrument.

DOI: 10.1103/PhysRevPhysEducRes.17.023105

I. INTRODUCTION

There is general agreement today among science educators that teachers' deep understanding of their students' knowledge is a necessary key step that might significantly help them in their efforts to design effective learning environments [1–6]. This view is expressed well by Vosniadou, Ioannides, Dimitrakopoulou, and Papademetriou [7], who write that "teachers need to be informed about how students see the physical world and learn to take their points of view into consideration when they design instruction" (p. 392). Science knowledge is reflected in mental models that students tend to form when dealing with science phenomena [8–10]. Mental models are dynamic structures created on the spot to provide explanations for these phenomena, make predictions, solve problems, and answer questions. One way to identify students' mental models about scientific concepts is by using diagnostic tools (multiple choice tests or inventories).

Several statistical methods were developed to analyze the data of diagnostic tools: classical test theory [11,12], item response theory [11,13], mental model analysis [9], etc. However, most aimed at analyzing the most modest and typical type of closed knowledge tests. That is, the examinee is faced with a certain problem for each test item, and then, of the usual three to five statements, is prompted to choose the most correct one [12]. After completing the test, a researcher can calculate the frequency of each type of answer and thus draw conclusions about the examinee's thinking. When the number of statements in each item is fixed and each item includes exactly one statement for every mental model, such an analysis is quite easy. However, the analysis becomes trickier when the number of statements per item varies, the number of statements corresponding to a given mental model varies, neutral distractors (not corresponding to any mental models under consideration) are present, and the examinees may choose more than one statement per item [5,6]. This is done in order to measure more accurately and in more details the spectrum of students' ideas. Such an analysis is not straightforward even in the dichotomous case of two possible mental models, for instance, materialistic vs process mental models of heat and sound propagation [14,15], or direct vs emergent views of these processes on some more advanced levels [6,15].

The professional literature suggests some statistical approaches to quantify examinees' responses. In the classical approach, the researcher usually calculates the average score of items within the test. The result obtained can provide us with some information about what the examinee knows, but nothing about what he does not know. Yet, when applying diagnostic tools such as the Force Concept Inventory (FCI) [16], sound concept inventory

*Corresponding author.
yuvalb@sapir.ac.il

instrument (SCII) [5], simple apparatus-based diagnostic instrument (SABDI) [6], etc., we are not usually interested in just grading the examinee but also in gaining some information about their ways of thinking and on shedding light on their mental models and misconceptions [12]. The traditional test score lacks this information as we demonstrate below. In order to probe alternative mental models of the examinees, the researcher needs to provide additional calculations, for instance, estimating the frequencies of responses relating to each model. In the case of only two possible mental models, the researcher actually can score *model A* answers with positive points, while *model B* answers with negative ones. This method, however, also demands more calculations while the output is not normalized; which makes it to be less convenient. On the other hand, Bao's model density matrixes approach [9] aims at analyzing incorrect mental models as well as correct answers. This approach applies the mathematical apparatus of quantum mechanics for data processing and produces a density matrix on the output. Thus, despite the rich informational context, this approach may seem quite complicated for use, especially in the case of only two possible mental models as discussed above.

The current exploratory study suggests a relatively simple statistical factor intended to analyze multiple choice items in the case of two target mental models, called F_{2m} (namely, factor of two models). F_{2m} is a scalar value normalized to $-1 \leq F_{2m} \leq 1$ whose calculation demands simple arithmetic calculations. Yet, it can inform the user of the mental model state of a subject or group of subjects and their consistency in thinking, as we will explain in the following sections. To the best of our knowledge, such statistics are absent in the professional literature. We further demonstrate the use of F_{2m} by applying it to analyze pre- and post-test results of SCII [5] conducted in the "Physics of Sound" course.

II. RESEARCH AIMS

The current study aims at developing a relatively simple statistic intended for estimating the mental model state of a subject or a group of subjects from the results of a multiple-choice knowledge test in the case of two mental models under consideration, while the test items include different numbers of statements, the number of statements pertaining to a certain mental model within an item is not fixed, and the examinees may choose more than one statement in each item.

III. DEVELOPING F_{2m}

Let us examine a test of N items and focus on a certain item i. The item consists of a question and a set of possible answers as follows: answers reflecting mental model 1; answers reflecting mental model 2; and neutral answers—answers that do not relate to either model 1 or 2. The subject is not limited in the number of answers chosen. Suppose that a certain examinee selected a answers of model 1, b answers of model 2, and c neutral answers. We define F^i_{2m} for this item as

$$F^i_{2m} = \frac{a_i - b_i}{a_i + b_i + c_i}. \tag{1}$$

From this definition, F^i_{2m} is normalized so that $-1 \leq F^i_{2m} \leq 1$. Furthermore, if $F^i_{2m} > 0$, the subject tends to think in the terms of mental model 1, while $F^i_{2m} < 0$ means that model 2 is dominant in their thinking. $|F^i_{2m}| \to 1$ indicates that the subject is consistent in their ideas. That is, if the examinee is fully consistent in model 1—chooses answers of type 1 only, the F^i_{2m} value will be equal to 1 no matter what a_i is (how many answers of model 1 are chosen) since, $F^i_{2m} = \frac{a_i - 0}{a_i + 0 + 0} = 1$. If on the contrary, an examinee is fully consistent in model 2, F^i_{2m} yields the output (-1) for any b_i as $F^i_{2m} = \frac{0 - b_i}{0 + b_i + 0} = -1$. This mechanism minimizes the bias which might take place in other ways of analysis due to a larger amount of answers of a certain type relative to the other one. On the other hand, $|F^i_{2m}| \to 0$ means that the subject is not consistent in their ideas. The F_{2m} of the whole inventory is the average of all F^i_{2m}. That is,

$$F_{2m} = \frac{1}{N}\sum_{i=1}^{N} F^i_{2m} = \frac{1}{N}\sum_{i=1}^{N} \frac{a_i - b_i}{a_i + b_i + c_i}. \tag{2}$$

In case all the choices of the items are contradictory (for instance, a. greater than; b. equal to; c. less than) and there is no logical possibility for more than one correct answer, F_{2m} will also work—the values of a and b in Eqs. (1) and (2) would be simply 0 or 1; however, its application might seem unnecessarily time consuming.

Eshach's sound concept inventory instrument [5] measures middle school students' understanding of sound. In particular it deals with the following two ways of thinking: (a) *process*—sound is a process of pressure and density changes propagating in the medium; and (b) *materialistic*—sound is a kind of material (e.g., special sound particles or air molecules) spreading from the source. Thus, the SCII provides a suitable case for F_{2m} analysis application. Let us focus, for example, on item 19 in the SCII (p. 010102–13):

19. ***When you stand behind the door to a room in which music is playing, you can still hear the music because***

i. The sound is made of small particles that can pass through gaps, like the one between the door and the floor.
ii. The sound is made of different sized particles. The smallest ones can get through doors and walls that are not totally sealed.
iii. The changes in air density formed in the gap between the door and the floor travel outside.
iv. The sounds in the room cause the wall to vibrate. The vibrating wall causes the air on the other side to vibrate and slightly changes the air pressure there.

TABLE I. Three hypothetical examples of responses in item 19, the appropriate F_{2m} values' calculation, and what kind of mental model state they reflect.

Student	Statements selected	F^{19}_{2m} calculation		Conclusion
1	iii	$a_{19}=1$ $b_{19}=0$ $c_{19}=0$	$F^{19}_{2m}=\frac{1-0}{1+0+0}=1$	Consistent process model
2	i, ii, iv	$a_{19}=1$ $b_{19}=2$ $c_{19}=0$	$F^{19}_{2m}=\frac{1-2}{1+2+0}=-\frac{1}{3}$	More materialistic than process thinking, but not consistent.
3	i, v—*I place my ear close to the door* (neutral statement)	$a_{19}=0$ $b_{19}=1$ $c_{19}=1$	$F^{19}_{2m}=\frac{0-1}{0+1+1}=-\frac{1}{2}$	Materialistic thinking. Neutral answer decreases the level of consistency.

v. None of the above choices fits my basic viewpoint. My basic viewpoint is (please explain your viewpoint in the space provided below).

This item includes the following four statements: items (iii) and (iv) reflect the *process model of sound—a*; items (i) and (ii) reflect the *materialistic model—b*; and the open option (v) can be materialistic, process, or a neutral model depending on the examinee's answer. Now let us examine three hypothetical examples of students whose answers are presented in Table I.

IV. ILLUSTRATING THE USE OF F_{2m} AND PRELIMINARY EXAMINATION OF ITS CONSISTENCY

F_{2m} is a new statistic. Therefore, as a first step, F_{2m} was reviewed by three experts, all of whom are professors in the Graduate Program for Science and Technology Education. All three agreed that F_{2m} is of a potential to contribute the examination of students' concepts and misconceptions by simplifying the analysis and making the output to be more vivid and informative. The second step was applying F_{2m} in analyzing the SCII results. The SCII was run twice in the acoustics course taught in the Graduate Program for Science and Technology Education at Ben-Gurion University of the Negev in 2010 before and after instruction [14]. The course emphasized the process nature of sound versus naïve materialistic models according to the recommendations of Chi and colleagues [15].

According to the literature, a minimal sample of 20 or 30 people is usually enough to obtain useful statistics [17–19]. Thus, a starting sample of 27 students (males and females) studying the course was considered as being sufficient for the preliminary analysis. Twenty-five of the participants had not studied acoustics prior to the course.

As is known from the literature, inexperienced students tend to mistakenly consider sound as a kind of material [4–6,14], or tend to be found in a mixed model state (the term is taken from Bao [9]), considering sound as a material in some contexts and as a process in others. Therefore, when running the SCII in this group prior to the course, we would expect more negative or close to zero values of F_{2m} than striving for the 1 result. Following the instruction, when the knowledge constructs are expected to be created and mental models are expected to change from materialistic to process models, we would expect to obtain mostly positive values of F_{2m} that are close to 1. That is, a statistically significant growth in F_{2m} values after the course relative to the prior ones might indicate a degree of F_{2m} consistency. Indeed, the pretest yielded an average score of 58 points (out of 100), while the score went up to 80 points in the post-test. Moreover, these numbers provide some indication of students' prior level of knowledge compared to after the course; however, they do not fully reflect their thinking. To explain this point, let us look, for instance, at the average pretest score—58. Suppose we have the following two students:

- Student 1 answered 58% of the items correctly (i.e., he chose the correct process options), while he chose materialistic distractors in the remaining 42% of the items.
- Student 2 also answered 58% of the items correctly. As to the other items, he provided 10% materialistic and 32% neutral responses.

Without a doubt, the understanding of sound by these students is quite different. These differences, however, are not reflected by personal and average scores. Indeed, students 1 and 2 both achieved the average grade of 58 points. However, student 1 provided twice as many materialistic responses compared to student 2. This information is lost when the score is calculated. In a similar manner, the score does not reflect the information regarding the incorrect responses of all participants—information that could be crucial for the research of students' conceptualization and conceptual change. F_{2m} might offer an alternative way for inventory assessment aimed at preserving this information and bringing it to the researcher in a relatively simple and convenient manner. Let us now apply F_{2m} to the SCII results.

As before, we define a_i as the number of answers reflecting a process model(s) of sound and b_i materialistic views. The SCII results yielded the following average

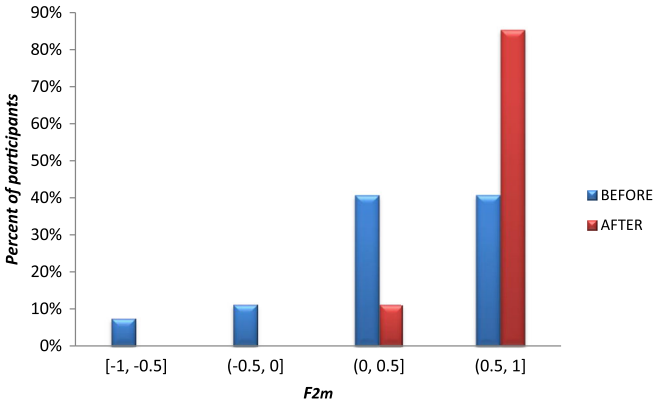

FIG. 1. Distribution of F_{2m} before and after the course.

values of F_{2m} for the group: $\langle F_{2m}\rangle_{\text{pre}} = 0.34$ before the instruction and $\langle F_{2m}\rangle_{\text{post}} = 0.8$ after, that is, a growth of $\Delta\langle F_{2m}\rangle = 0.44$. In addition to calculating $\langle F_{2m}\rangle$, we went one step further and calculated the percentage of students within the four ranges of F_{2m}, as presented in Fig. 1.

As can be seen in Fig. 1, 18.51% of the participants had a negative F_{2m} before the course, meaning more materialistic thinking about sound. The greatest number of students (40.74%) were found in the $0 < F_{2m} \leq 0.5$ range, meaning more process than materialistic thinking; however, the concepts were weak and there was no consistent theory system. Whereas 88.47% of the students reached $0.5 < F_{2m} \leq 1$ after the course, i.e., more than a twofold growth in the number of students possessing consistent process thinking about sound. As expected, no students had a negative F_{2m} after the course.

Additional evidence of F_{2m} consistency is provided by a comparison of F_{2m} values obtained in the test to students' performance in clinical or focus group interviews [20]. For instance, students who can explain sound in materialistic terms are expected to get negative F_{2m} scores; those found in a mixed model state have close to zero values; while students who succeed to explain sound as a process should acquire $F_{2m} > 0.5$ values. In the first course lesson, the lecturer inspired a whole class dialogic discussion that could be regarded as a focus group interview [20] pertaining to the nature of sound. The discussion was provided before the students underwent the SCII. A total of 77.78% of the participants failed to explain correctly what sound is. Some actually related to sound as a kind of material that can move from point A to point B, while the majority provided explanations that contained substance-based statements in addition to process descriptions. This distribution of ideas reinforces the pre-instruction average value of F_{2m} as 0.34. To illustrate this point, Table II shows some representative examples of students' answers and their analysis. The appropriate F_{2m} is noted next to each answer.

V. DISCUSSION AND CONCLUSIONS

The current preliminary study aimed at developing a simple statistic for estimating the mental model state in the dichotomous case of two possible models, e.g., materialistic vs process views of heat and sound, direct vs emergent processes [5,15,22,23], etc. The F_{2m} factor was suggested. Results obtained in this short study indicate a degree of F_{2m} consistency and should be considered as being promising enough to justify a more comprehensive study conducted on larger samples, as well as on other inventories.

As the SCII analysis example used in this study shows, F_{2m} should not be considered as an indication as to whether a subject actually *knows* the material, but only *what kind of thinking* he leans towards. Indeed, some items in the inventory have more than one process statement—answers (iii) and (iv) in item 19, for instance. If an examinee selected only statement (iii), it would result in $F_{2m} = 1$

TABLE II. Some examples of students' explanations of the sound concept, their short analysis, and appropriate F_{2m} values.

Students' explanations	Analysis	F_{2m}
Sounds are waves of particles. When we speak, particles move like a wave and when we shout, it's like a straight line.	Sound is a flux of corporeal particles propagating in wavelike and straight lines depending on the sound intensity. This explanation fits well the characteristics of a materialistic model [4,21]	−0.75
I think that sound [propagation] is not like a wave. It is more like a straight line. It is like the air pushes the sound particles.	Sound particles are pushable (by the air), which also fits the substance schema [21].	−0.19
Sound is energy. Sound is released to the air and passes to the listener's ear through sound waves. Sound wave is a disturbance of the media.	This answer combines the two models of sound. On the one hand, "*sound is released to the air and passes…*" that could be related to a materialistic view of sound passing from point A to point B according to the substance schema [21]; while, on the other hand, "*sound wave is a disturbance of the media*" that seems to be a process statement.	0.06
Sound is like waves of different frequencies and phases that cause different intensities.	This explanation is much more process. Indeed, the participant related sound in wave terms also later in the interview. However, there is still certain confusion about physical concepts which obviously reduced his F_{2m} score.	0.56

according to Eq. (1). This, however, cannot point towards a full understanding since he has not related to the correct statement (iv). Moreover, in the cases where there are several mental models of each type, F_{2m} would reflect only the type of thinking but not indicate the exact model. For instance, there are actually three possible mental models of sound in the materialistic category of thinking: (a) special sound particles [4]; (b) vibrated air molecules traveling from the source to the listener [3,6,23,24]; and (c) sound that travels as air globules of sonic data [14,25]. Similarly, there are two process models of sound: (d) transverse wave; and (e) longitudinal wave [6,24]. In this case, choosing models (a),(b) or (c) will result in a negative contribution to F_{2m} as actually reflecting materialistic thinking, whereas selecting (d) or (e) will move F_{2m} to positive values reflecting process thinking. Thus, if the obtained value of F_{2m} for a certain examinee is 1, it only indicates that they possess a process model of sound but does not specify which one. Therefore, to gain a more comprehensive picture of students' conceptions F_{2m} should be used in combination with other knowledge analysis methods.

Another point to be discussed is the possibility of guessing. In this case, F_{2m} might seem as bias towards the mental model that contains the most options. For example, if model a has more relevant options compared to b, then a randomly answering student will end up with F_{2m} pointing on model a. One way to minimize the possibility of incorrect inference based on guessing is calculating the Cronbach alpha coefficient ($0 \leq \alpha \leq 1$). Cronbach's alpha is aimed at indicating a test reliability. Tests having $\alpha \geq 0.7$ are generally considered to be reliable [11]. Incoherent guesses will make the test results unreliable, consequently yielding low Cronbach's alpha value [12].

ACKNOWLEDGMENTS

Y. B. A. and A. V. wrote and analyzed the data. H. E. and Y. B. A. supervised A. V.

[1] I. Galili and A. Hazan, Learners' knowledge in optics: interpretation, structure and analysis, Int. J. Sci. Educ. **22**, 57 (2000).
[2] D. F. Treagust, Conceptual change as a viable approach to understanding student learning in science, in *Teaching and Learning Science: A Handbook* (Rowman & Littlefield, Lanham, Maryland, 2006).
[3] Z. Hrepic, D. A. Zollman, and N. S. Rebello, Identifying students' mental models of sound propagation: the role of conceptual blending in understanding conceptual change, Phys. Rev. ST Phys. Educ. Res. **6**, 020114 (2010).
[4] H. Eshach and J. L. Schwartz, Sound stuff? Naive materialism in middle-school students' conceptions of sound, Int. J. Sci. Educ. **28**, 733 (2006).
[5] H. Eshach, Development of a student-centered instrument to assess middle school students' conceptual understanding of sound, Phys. Rev. ST Phys. Educ. Res. **10**, 010102 (2014).
[6] A. Volfson, H. Eshach, and Y. Ben-Abu, Development of a diagnostic tool aimed at pinpointing undergraduate students' knowledge about sound and its implementation in simple acoustic apparatuses' analysis, Phys. Rev. Phys. Educ. Res. **14**, 020127 (2018).
[7] S. Vosniadou, C. Ioannides, A. Dimitrakopoulou, and E. Papademetriou, Designing learning environments to promote conceptual change in science, Learn. Instr. **11**, 381 (2001).
[8] S. Vosniadou and W. F. Brewer, Mental models of the day/night cycle, Cogn. Sci. **18**, 123 (1994).
[9] L. Bao, Dynamics of student modeling: a theory, algorithms and application to quantum mechanics, Ph.D. thesis, Faculty of Graduate School of the University of Maryland at College Park, 1999.
[10] I. M. Greca and M. A. Moreira, Mental, physical, and mathematical models in the teaching and learning of physics, Sci. Educ. **86**, 106 (2002).
[11] L. H. Janda, *Psychological Testing: Theory and Applications* (The Open University, Tel Aviv 2008).
[12] W. K. Adams and C. E. Weiman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2010).
[13] B. B. Reeve and P. Fayers, Applying item response theory modeling for evaluating questionnaire item and scale properties, in *Assessing Quality of Life in Clinical Trials: Methods of Practice*, 2nd ed. (Oxford University Press New York, 2005).
[14] A. Volfson, Dialogue on sound waves: From particle to particle, Master's thesis, Faculty of Humanities and Social Sciences, Ben-Gurion University of the Negev, 2011.
[15] M. T. H. Chi, R. D. Roscoe, J. D. Slotta, M. Roy, and C. C. Chase, Misconceived causal explanations for emergent processes, Cogn. Sci. **36**, 1 (2012).
[16] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[17] L. Terry and K. Kelley, Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals, Brit. J. Math. Stat. Psychol. **65**, 371 (2011).
[18] J. C. Cappelleri, J. J. Lundy, and R. D. Hays, Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures, Clin. Ther. **36**, 648 (2014).

[19] N. A. Thompson, *Introduction to Classical Test Theory with CITAS* (Assessment Systems Corporation Cheshire Lane, Minnetonka, MN, 2016).

[20] A. Volfson, H. Eshach, and Y. Ben-Abu, Identifying physics misconceptions at the circus: The case of circular motion, Phys. Rev. Phys. Educ. Res. **16**, 010134 (2020).

[21] J. D. Slotta, M. T. H. Chi, and E. Joram, Assessing students' misclassifications of physics concepts: an ontological basis for conceptual change, Cognit. Instr. **13**, 373 (1995).

[22] M. T. H. Chi, Three types of conceptual change: Belief revision, mental model transformation, and categorical shift, in *Handbook of Research on Conceptual Change* (Erlbaum, Hillsdale, NJ, 2008).

[23] A. Volfson, H. Eshach, and Y. Ben-Abu, Introducing the idea of entropy to the ontological category shift theory for conceptual change: The case of heat and sound, Phys. Rev. Phys. Educ. Res. **15**, 010143 (2019).

[24] A. Volfson, H. Eshach, and Y. Ben-Abu, When technology meets acoustics: Students' ideas about the underlying principles explaining simple acoustic devices, Res. Sci. Educ. 1 (2020).

[25] I. Caleon and R. Subramaniam, Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves, Int. J. Sci. Educ. **32**, 939 (2010).