Measuring students' conceptual understanding of wave optics: A Rasch modeling approach

Vanes Mešić,[1,*] Knut Neumann,[2] Ivica Aviani,[3] Elvedin Hasović,[1] William J. Boone,[4]
Nataša Erceg,[5] Vladimir Grubelnik,[6] Ana Sušac,[7] Džana Salibašić Glamočić,[1]
Marin Karuza,[5] Andrej Vidak,[8] Adis Alihodžić,[1] and Robert Repnik[9]

[1]*Faculty of Science, University of Sarajevo, Zmaja od Bosne 33-35,*
71000 Sarajevo, Bosnia and Herzegovina

[2]*Leibniz Institute for Science and Mathematics Education (IPN) at the University of Kiel,*
Olshausenstraße 62, 24118 Kiel, Germany

[3]*Department of Physics, Faculty of Science, University of Split, R. Boškovića 33, 21000 Split, Croatia*

[4]*Department of Educational Psychology, Miami University, 201 McGuffey Hall, 45056 Oxford, Ohio, USA*

[5]*University of Rijeka, Department of Physics, R. Matejčić 2, 51000 Rijeka, Croatia*

[6]*Faculty of Electrical Engineering and Computer Sciences, University of Maribor,*
Koroška cesta 46, 2000 Maribor, Slovenia

[7]*Faculty of Electrical Engineering and Computing, University of Zagreb,*
Unska ul. 3, 10000 Zagreb, Croatia

[8]*Faculty of Chemical Engineering and Technology, University of Zagreb,*
Marulićev trg 19, 10000 Zagreb, Croatia

[9]*Faculty of Natural Science and Mathematics, University of Maribor,*
Koroška cesta 160, 2000 Maribor, Slovenia

(Received 4 September 2018; published 25 February 2019)

Even graduate physics students have many misconceptions about basic wave optics phenomena. This suggests that there is much room for improvement of the traditional wave optics curriculum. An effective way for initiating a curriculum change is to reconsider and revise the expected learning outcomes and corresponding assessment instruments. By systematically enriching our wave optics instruction and assessment with conceptual tasks, we may increase the probability of students actively engaging in learning the conceptual aspects of wave optics. In this paper, we present the process of developing an item bank for measuring understanding of wave optics in typical introductory physics courses at universities. Thereby, the Rasch modeling approach has been used. The development of the item bank has been guided by results from multiple expert and student surveys, as well as from group interviews and think aloud interviews. Altogether 65 multiple-choice items with a single correct answer and three distractors have been prepared for field testing. Until now, 35 out of 65 items have been field tested by means of a paper and pencil survey which included 188 participants from five universities in Bosnia and Herzegovina, Croatia, and Slovenia. The field test showed that 32 out of 35 items have good psychometric characteristics and that they may be very useful for uncovering students' misconceptions in wave optics.

DOI: 10.1103/PhysRevPhysEducRes.15.010115

I. INTRODUCTION

Supporting students in developing scientific conceptions about physics in order to explain natural phenomena is one main aim of school physics. However, students learn about natural phenomena in school physics but also through everyday experience [1]. Although the conceptions originating from everyday experience help the students to cope with everyday life, they are less powerful than scientific conceptions when it comes to providing precise predictions and explanations of a wider range of phenomena [1,2]. Some authors suggest that everyday and scientific conceptions can coexist, whereas others believe that learning of scientific concepts is best promoted through the process of conceptual change [2]. Both highlight that helping students become aware of their misconceptions and the limits of their scope is crucial for developing scientific conceptions. However, in order to best support students in this process, the teacher needs to know at any time where students are in

[*] vanes.mesic@gmail.com

their process of developing scientific conceptions and the misconceptions they still hold about the area of physics in question (e.g., wave optics).

In physics education research, students' conceptions are typically explored through oral interviews and/or written surveys. Although oral interviews allow us to study student conceptions at great depth, written surveys are more efficient for assessing the conceptions of larger number of students (e.g., in a university course). Over the course of time physics education researchers have carefully examined students' conceptions in many different areas of physics. The results of that research have been used for developing multiple-choice tests (e.g., for development of attractors and distractors in these tests). Some authors suggest that an even better approach than developing a simple assessment instrument is to develop comprehensive item banks that can be used for tailoring assessment instruments in order to meet different assessment goals [3]. Maybe the most effective approach to developing such item banks is the Rasch modeling approach, which is situated within the probabilistic test theory [4]. Although there was some highly valuable research in the field of students' misconceptions in wave optics, to the authors' knowledge at this moment there is no wave optics item bank that would allow for assessing university students' understanding of wave optics (e.g., light wave concept, superposition, thin film interference, multiple slit interference, and diffraction) [5].

In this paper we therefore describe the development of a wave optics item bank using the Rasch modeling approach and demonstrate its potential for measuring university students' understanding of wave optics. Building on prior research on student conceptions of wave optics we authored, tested, and revised multiple-choice questions. We finally administered these items in a field study and used the Rasch modeling approach to combine theoretical rationales and empirical evidence for purposes of supporting the adequacy of potential inferences and actions from data obtained through our items [6].

II. THEORETICAL BACKGROUND

A. Students' ideas about wave optics phenomena

Students' ideas about light are shaped by sensory experience, social and cultural practice, as well as by informal discovery learning and different meaning of terms in everyday and scientific language [7]. In primary school, students typically learn that light is something we need to see objects around us and something that plants need to grow. They also learn to name various sources of light, and develop first conceptions about the formation of shadows. In lower-secondary school, students in most countries learn that light is an electromagnetic wave and they are taught about the spectrum of electromagnetic waves. At this level, students typically begin to develop the concept of the light ray and learn to apply it in basic contexts related to rectilinear propagation, reflection, and refraction of light. In the upper-secondary school, the approach to teaching optics becomes more quantitative and the contexts in which the ray model is applied become more complex. Typically, ray optics is thereby used for explaining the principle of how various optical instruments work. However, after being exposed to instruction about basics of wave optics (e.g., double slit interference, single slit diffraction, and optical grating) the students' conceptions of light change and hybrid (particle-wave) models of light may develop [8]. At the level of introductory physics courses at the university, students often additionally learn about thin film interference and diffraction on a circular aperture [9–14]. Learning about multiple slit interference is sometimes part of the introductory physics curricula, too [11,14].

Earlier research has shown that students at all educational levels hold many misconceptions about light phenomena [7]. First, it is important to note that students hold somewhat inconsistent views about the concept of light. On the one hand, they often fail to conceptualize light as something that exists apart from its source and effects [7,15]. For example, many students are not aware that there is laser light between the illuminated single-slit mask and the screen placed directly opposite to the mask. On the other hand, the students realize that we are able to see the objects in our living room because everything in the room is immersed in a *sea of light* [16]. As a result of learning about ray optics some students begin to believe that rays are actual constituents of light waves, and as a result of wave optics instruction many students end up with hybrid conceptions which combine various models of light [8]. When it comes to propagation of light, for most students the idea of a light ray and rectilinear propagation of light is intuitively acceptable. On the other hand, many students struggle with interpreting the sinusoidal representation of the propagation of a plane electromagnetic wave [17]. Concretely, students often misinterpret the sinusoid as something that spatially delimits the wave propagation (e.g., *higher amplitude means a broader wave*) or as a trajectory of "light particles" or photons [18]. Such a mechanicistic reasoning is also reflected in the erroneous belief that reducing wavelength results in reducing of other "dimensions of the wave" [19]. Sometimes students also struggle with distinguishing the spatial and temporal versions of sinusoidal representations of a light wave which is probably a difficulty that has its origins in the mere characteristics of traditional wave optics instruction [20]. Besides difficulties in interpreting representations of light waves, students often fail to correctly apply the Huygens principle, although they know to verbally reproduce it. For example, many students believe that only the points at the edges of a slit become sources of secondary waves when the slit is illuminated with laser light [18]. Taking into account the difficulties related to understanding basic characteristics of a light wave it is no surprise that students also struggle with conceptualizing the superposition of two or more light waves. Some students believe that superposition only occurs if light waves are mutually coherent. Others believe that the resulting wave has always a larger amplitude and/or a larger wavelength compared to the individual, interfering waves [21]. Furthermore, there are also students who believe that two identical waves that propagate towards each other cancel out. Finally, many students tend to apply the superposition principle for individual pairs of waves that arrive at a certain point, instead of applying it for all waves at the same time [22]. When it comes to students' ideas about the characteristics of the interference pattern, many of them believe that there is only maximal constructive and maximal destructive interference of waves on the screen [19]. This is probably due to the fact that exactly these conditions (i.e., *maximal* constructive and *maximal* destructive interference) are mostly discussed in introductory physics courses. An alternative explanation is that students typically do not perceive the variations in irradiance within a single interference fringe in the context of a laboratory. As a result of an intensive use of mathematical relations for maximal constructive and maximal destructive interference many students in traditional courses learn how to use proportional reasoning for predicting how the fringes separation changes as a result of changing slit separation or slit width. However, thereby many students do not have a correct visual notion of the processes modeled by the mentioned mathematical relations. For example, they are often not sure how many waves are interfering at a given point of the screen. Particularly, they are confused by the fact that in double-slit interference we assume that there is an interference of only two secondary waves at an arbitrary point of the interference pattern, whereas in single-slit diffraction we typically assume that there is an interference of an infinite number of secondary waves at an arbitrary point of the diffraction pattern. This is reflected in the erroneous belief that in single-slit diffraction we should observe a first order *maximum* when the difference in path lengths of the two waves originating at the edges of the slit amounts to exactly 1 *wavelength* [23].

A possible source of students' difficulties in learning wave optics is that it is very demanding to create internal visual representations of a light wave which is a function of two variables [24]. Thereby, for students it is particularly difficult to overcome the mechanistic notions that are often associated with the sinusoidal representation of the light wave. Additionally, in wave optics phenomena it can be also very demanding to correctly visualize the scale of the given system (e.g., approximately parallel paths of light from edges of the slit towards an arbitrary point of the screen in the Fraunhofer approximation). Finally, it seems that in certain situations conceptual understanding is hindered due to a lack of factual knowledge regarding the scope of the ray model of light [18].

B. A Rasch model approach to item bank building

Speaking from a technical point of view, conceptual tests can be designed either in the classical test theory framework or in the probabilistic test theory framework. In classical test theory, it is assumed that a person's observed test score is comprised of their "true" score and a measurement error, whereas a defining feature of *probabilistic* test theory is that its models allow us to make statements about outcome *probabilities* for certain manifest (observable) variables [25]. According to Liu [26] the large majority of conceptual tests in the field of science education are based on classical test theory. However, most researchers agree that probabilistic test theory has many advantages in comparison to classical test theory. Hambleton, Swaminathan, and Rogers [27] name the following advantages of probabilistic test models: (a) item characteristics (e.g., difficulty parameter) are not group dependent, (b) examinee ability estimates are not test dependent, (c) test model is expressed at the item level, and (d) the possibility of providing measures of precision for each ability score.

There are various probabilistic test models which differ with respect to the shape of the item characteristic function—a function that models the relationship between examinees' responses to items and corresponding measures of examinees' ability [27]. One of the simplest probabilistic test models is the dichotomous (simple) Rasch model [28,29]. It distinguishes itself from other probabilistic models by the specific objectivity feature (differences between item difficulty estimates are sample independent), as well as by the possibility to generate relatively stable item estimates already for samples as small as 100 students [4,30,31]. For the Rasch model, the item characteristic function is as follows [27]:

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}, \tag{1}$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee of ability θ correctly responds to item i and b_i is the difficulty parameter for item i.

We can see that in the Rasch model there is only one item parameter which in interaction with the ability of the examinee predicts the probability of correctly solving the item. This is the difficulty parameter of the item. It corresponds to the point on the ability scale for which the probability of a correct solution amounts to 50% (Fig. 1). Both, the person ability and the item difficulty, are measured in the same units, i.e., in logits. Thereby, *"one logit is the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718"* [32].

Only if empirical data fit the item characteristic curves, the item parameters are as independent as it is statistically possible for the particular sample of persons from a homogeneous population [27,33]. The data-model fit is

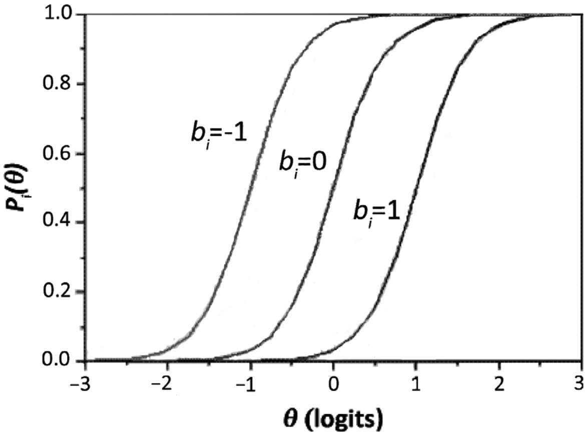


FIG. 1. Examples of item characteristic curves for the Rasch model.

of great importance for many practical applications of the Rasch model. In the Rasch context, fit statistics indicate how accurately or predictably data fit the model [34] and we can say that a single response pattern exhibits misfit if that pattern is improbable or too probable given the Rasch model or given the other patterns in a sample [35]. If that response pattern represents a single person's set of responses we speak of person misfit, and if that response pattern represents all persons' answers on a single item we speak of item misfit.

Another important measure in the Rasch modeling approach is the item information function. The information function of an item tells us how much the item contributes to ability estimation at various points along the ability scale, whereby the functions of individual test items are mutually independent, i.e., they can be estimated without knowledge about other items included in the test [27]. For example, a very difficult item usually offers much more information regarding differences among high-proficiency students than among low-proficiency students. When it comes to test construction, the Rasch modeling approach makes it possible to adjust the precision of the test at different points of the ability continuum. Thereby, the standard error of an ability parameter θ is calculated based on the corresponding value of the test information function [36]:

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}, \tag{2}$$

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta), \tag{3}$$

where $I_i(\theta)$ is the information function of item i at θ and k is the number of items.

According to Szabo [4] the advantages of the Rasch model can be best utilized by development and use of item banks. Thereby, the function of an item bank is to *"store a large number of test items with information concerning the content and psychometric characteristics of each, so that the user can select from this a set of items to construct a test which suits his/her requirements"* [37]. Typically, information about item ID, item type, difficulty parameter, point biserial, item information function, content coverage, as well as item text and responses are provided in the item bank. Development and use of item banks is most effectively performed through psychometrics software, such as TestAssembler or FastTest [38]. Within these software packages items with required properties (content and psychometric characteristics) can be easily identified and multiple test forms of similar characteristics can be generated. For example, it is possible to construct two tests whose items completely differ, yet achievement on them may be related to a common measurement scale [39]. Moreover, we can tailor the characteristics of the instruments to meet various testing aims. To that end, it is first necessary to decide on the shape of the desired test information function (often called target information function) and to select items from the item bank until the test information function matches the target information function [27]. For example, if it is important to have equally precise measures along the whole proficiency continuum then we would attempt to obtain an approximately flat test information function. On the other hand, if our aim is to decide whether examinees are below or above a certain cutoff score (e.g., indicating mastery level) our goal would be to obtain a test information function that has its maximum at the corresponding cutoff value of the ability parameter [27].

Tailoring tests to the needs of various target samples only works efficiently if we have developed large item banks. The criteria that have to be met for purposes of successful development of item banks in the probabilistic test theory framework are as follows [4]: (a) a sufficiently large sample of respondents is available for field testing of items, (b) a suitable test model is selected, (c) clear criteria are set for adding items to the item bank (e.g., in terms of item and person fit statistics), (d) an efficient linking schema is used for continuous adding of new items to the item bank, and (e) adequate software is available.

An important application of item banks is in computerized adaptive testing (CAT). In CAT the central idea is to have the computer select the items that best suit the abilities of the particular respondent [3]. Thereby, the estimation of the respondents' ability level starts already after their answer to the first or second question, whereafter the computer selects an item that is aligned with the current estimate of the proficiency and the proficiency estimation process continues until the proficiency estimate converges. Such an approach typically results in quicker and more precise testing.

Generally, the main criteria for assessing the quality of testing are objectivity, validity, and reliability [40]. Objectivity is related to the requirement that the testing, scoring, and interpretation procedures are objective, i.e., independent from the people who implement them [41]. Reliability defines the precision of measurement [40] and in the Rasch approach it can be assessed by means of the mentioned test information functions. Finally, validity can be defined as an [6] "*integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inference and actions based on test scores or other modes of assessment*."

III. METHODOLOGY

In developing our measurement instrument we mainly followed the procedure proposed by Liu [26]: (i) Defining the construct, (ii) delineating the construct into items, (iii) small scale item try out (group interviews, individual think-aloud interviews, written survey), (iv) final field testing, and (v) Rasch modeling. Through all the above-mentioned stages the instrument has been iteratively improved and validity arguments have been constructed. We investigated content validity, cognitive validity, construct validity, and correlational validity and checked for which student population our test ensures the most reliable and valid measures.

A. Defining the construct

The aim of our efforts was to develop a wave optics item pool capable of measuring university students' conceptual understanding of wave optics. The item pool was supposed to be able to identify misconceptions (i.e., have a diagnostic function) and to obtain insights into different levels of students' conceptual understanding for comparing the effectiveness of different wave optics courses (i.e., have a summative function).

Defining conceptual understanding first required us to define the content to be addressed by test items. This served as a basis for identifying relevant misconceptions and meaningful levels of understanding wave optics. According to McKagan, Perkins, and Wieman [42], if we want to use a test for purposes of evaluating the efficacy of different university courses, we should make sure that its content cover the intersection of content taught in the corresponding courses. In order to identify this intersection of wave optics content taught in introductory physics courses, we analyzed several introductory physics textbooks that are used worldwide [9–14]. In addition to the content typically covered in standard introductory textbooks, we included content that we considered relevant for a comprehensive understanding of wave optics, but that are sometimes left out from introductory physics textbooks (e.g., spatial and temporal coherence, multiple slit interference, combined interference and diffraction). Combined interference and diffraction is, for example, relevant for understanding double-slit experiments in practice, since due to the finite width of the slits we always have interference and diffraction effects which results in differences in the brightness of the fringes. Still, we decided to leave out polarization and scattering, since inclusion of this content would likely reduce homogeneity of construct and increase the risk of violating the unidimensionality assumption central to the Rasch modeling approach. Based on these considerations the final list of content areas included the following: (i) Fundamentals of wave optics (wavelength, frequency, period, amplitude, light intensity, phase, difference in optical path lengths, phase difference, coherence, superposition of waves, Huygens-Fresnel principle); (ii) interference on thin films; (iii) double slit interference (very narrow slit approximation); (iv) multiple slit interference (very narrow slit approximation [43]), (v) single slit diffraction; (vi) combined interference and diffraction; (vii) optical grating; (viii) diffraction from a rectangular, circular aperture, or obstacle.

Based on our review of the literature we identified misconceptions that students would typically hold in these content areas [17–19,21,22] (Electronic supplement A [44]). Many of these misconceptions are related to misapplication of geometrical optics in contexts that require a use of the wave model (e.g., *narrowing a slit results in a narrower central diffraction maximum*), whereas others are related to misinterpretations of some widely used representations in wave optics (e.g., *sinusoids represent photon trajectories*).

In order to measure students' levels of conceptual understanding about wave optics we developed a learning path model for wave optics based on the results of earlier studies on learning and teaching wave optics, as well as on our experience in teaching introductory physics courses. For each content area, we specified broad learning goals. These learning goals were then arranged in terms of increasing complexity. That is, in the learning path model initial learning goals related to basic features of the light wave were followed by goals related to superposition of two light waves, superposition of multiple light waves, single slit diffraction, combined interference and diffraction, concluding with diffraction in two dimensions. The learning path model served as the basis for delineating the construct into items. The complete version of the learning path model is provided as Supplemental Material B [45].

B. Delineating the construct into items

A core aspect of developing a measurement instrument is to make sure the items and subsequently students' performance on the items yield sufficient evidence to draw conclusions about the to-be-measured construct [46]. To ensure sufficiently alignment between the construct and student performance we expressed our ideas about students' understanding of wave optics in terms of observable behaviors that can reasonably be expected from students in introductory physics courses at the university level.

In assessing students' understanding of wave optics we aimed to focus on the application as the ability to apply the explanation of a phenomenon or predict what will happen to a system if the system is perturbed across a wide range of situations [47,48]. As situations, we chose those situations that had earlier been shown to trigger student misconceptions about wave optics. For these situations we delineated appropriate misconceptions from the list of misconceptions identified from the literature and the general objectives in the learning path model into observable student behaviors (i.e., performance expectations) in order to obtain a productive basis for purposes of writing the test items. That is, the level of concreteness of the specified observable behaviors was between the broad learning goals and concrete physics items. Based on these observable behaviors we authored items suitable to foster such behaviors [49].

In authoring items we first created item stems describing the situation and raising an issue or questions meant to elicit behaviors such as the ones described previously. In order to visually represent the given phenomena as well as make our items more attractive for students, we situated many items within a wave optics laboratory context. However, in order to take into account the *"faculty buy-in component"* [42] we also situated several items in contexts typically used in conventional wave optics instruction. Based on the observable behaviors specified we authored response options. For each item stem we created four response options: one correct option based on the learning goals and four incorrect ones based on misconceptions identified.

Initially, altogether 60 items were created, whereby most of these items were original. After the specification of the initial item pool, we decided to conduct the process of content validation. This process proceeded through two stages.

In the first stage of content validation, we asked five physicists whose field of expertise is atomic, molecular and optical physics to take the questionnaire from Supplemental Material C [50]. Unfortunately, one of them did not respond to our request at all, one responded only partially, and three responded in detail. Besides having subject matter expertise, all three of the experts who provided detailed opinions were engaged in teaching physics at the university level. They all had experience with teaching introductory physics and/or authoring introductory physics textbooks. Generally, all three experts agreed that the learning goals from the model of the learning path are mostly adequate for an introductory physics course, at the university level. Furthermore, they agreed that learning goals are correctly linked to specific learning objectives (i.e., operationalizations of goals) and test items. However, they also identified room for improvement when it comes to wording of certain items.

Based on feedback obtained from the first stage validation surveys, we attempted to further improve the model of the learning path and the original item pool. Some items from the initial pool were deleted, and some new were added in order to better represent the model of the learning path, which resulted in the fact that the number of items increased to 72. In the second stage of the validation process seven university professors of physics were asked three questions for each of the 72 items and were provided the opportunity to comment the items (Table I). The results of the expert survey showed that in 80% of all cases the experts agreed that the test item measures an important aspect of understanding of wave optics in the context of typical introductory physics courses, at the university level. Furthermore, the average rating of the items was 8.3 on a scale from 5 to 10. On average, the first part of the pool (i.e., fundamentals of wave optics), was rated with a 8, and the second part of the pool (i.e., applications of wave optics) received an average rating of 8.5. Based on these results, we concluded that the "faculty buy-in" [42] aspect seemed to be satisfying for our item pool. Additionally, on average the item difficulty was estimated to be 2.14 on a scale from 1 to 3. The average difficulty of the fundamentals of wave optics part of the pool was estimated to be 1.95, whereas the average difficulty estimate for remaining items was 2.3. It is interesting to note that an increase in the rating of item difficulty was positively related to the rating of item attractiveness ($r = 0.38$, $p < 0.001$). Based on the answers to the first two questions from the survey and our considerations of the integrity of the item pool as a whole, we decided to exclude 11 questions from the pool and to slightly modify the model of the learning path. For example, most experts indicated that the question which was supposed to measure students' understanding of the mathematical representation of the plane electromagnetic wave does not measure conceptual understanding, at all. So we excluded this question, as well as the corresponding learning objective from the model of the learning path.

TABLE I. Excerpt from the expert survey from the second stage of content validation. The scale from 5 to 10 corresponds to the grading scale used at most ex-Yugoslavia universities in which 5 stands for insufficient, 8 for very good, and 10 for outstanding.

Is the skill (or knowledge) measured by this item an important aspect of understanding wave optics in the context of typical introductory courses of physics at the university level? Yes No

On a scale of 5 (not at all) to 10 (very much) how much do you like the item? 5 6 7 8 9 10

On a scale of 1(appropriate for below average university students) to 3 (appropriate for above average university students) how would you rate the degree of item difficulty? 1 2 3

Here you can specify potential additional comments:

C. Small scale item try out

Before the final field testing of the item pool, we decided to conduct a small scale item try-out in order to further improve the technical and psychometric characteristics of our items. Taking into account the fact that one of the aims of our items was assessing the effectiveness of introductory wave optics curricula, we concluded that the student sample should consist of students who had recently attended and completed an introductory course of physics or optics at university. According to Ding, Chabay, Sherwood, and Beichner [51] the item try out can be conducted with students who had attended the relevant physics course 3 months or even 5 semesters earlier. In all stages of our item try-out participants were physics students who had earlier attended the introductory optics course at the Faculty of Science, University of Sarajevo. This one-semester (15 weeks) course is attended by second year students. It includes 3 h of lectures and 2 h of recitations per week. Thereby, wave optics is taught from week 10 to week 15 in the spring semester, which is 30 teaching hours, in total. The item try out mostly included third and fourth year physics students who had completed the introductory optics course between eight and twenty months earlier.

The item try out proceeded in three stages. In the first stage of item try out we asked physics teacher students (3rd and 4th year students) to solve the constructed response version of our test items and to give a written explanation for each of their answers. We decided to use the constructed response version of our items because we wanted to gain additional insight into potential student misconceptions. Because of the relatively big size of the item pool (61 item), we could not administer all of the items in only one appointment, but we had to divide the item pool into two parts and administer them in two different appointments. At the first appointment, six physics teacher students answered the first 24 items (related to basic wave optics concepts—e.g., optical path length, phase difference, coherence, superposition), and at the second appointment eight physics teacher students (mostly the same students as in the first appointment) answered the remaining 37 items (related to applications of basic wave optics concepts—e.g., thin films, slits, gratings, apertures, or obstacles). After students had finished the written survey, they were asked to further discuss their ideas about the wave optics items within a group interview.

For purposes of data analysis, for each of the 61 items, we gathered together the written information from all the students, as well as information from the video-taped group interview. Based on synthesis of the information, we tried to improve the item stems, as well as the item distractors.

In the second stage of the item try out, we conducted think-aloud interviews in order to investigate the cognitive validity of our item pool, i.e., we investigated the cognitive processes induced by our items [52]. For that purpose, we asked ten physics students from the University of Sarajevo (nine third year students, and one second year student) to think aloud about a sample of 24 out of 61 test items (multiple-choice version). Thereby, the item sample has been drawn by the procedure of proportionate stratified (random) sampling [53]. The strata consisted of item sets associated to the individual learning goals from the model of the learning path.

On average, an individual think-aloud interview lasted approximately 75 min. The think-aloud interviews were audio taped and subsequently transcribed and coded, whereby the unit of coding was a sentence. The following main coding categories were used: test item (e.g., sentences from the item stem or student's opinion about the item), strategy (e.g., sentences that provide information about item solving strategy), person and situation (e.g., sentences about the way the student feels at that moment), and other (all sentences that could not be assigned to previously mentioned categories). Furthermore, within the strategy category we tended to distinguish construct relevant versus construct irrelevant strategies, with the aim to check whether our items indeed measure wave optics understanding or something else. Sentences were coded as "construct relevant" if they provided us with relevant information about a student's level of proficiency on the measured construct, and they were coded as "irrelevant" if they failed to provide that information (e.g., when students attempted to solve the item by guessing). In certain occasions, from the analysis of a sentence it was not possible to unambiguously infer whether it reflects construct relevant processes or guessing (e.g., *"The first option is certainly correct."*) and such sentences were coded as "undecidable." Figure 2 shows the share of the individual main coding categories, whereas Fig. 3 shows how construct relevant strategies compared to construct irrelevant strategies. From Figs. 2 and 3 it followed that the cognitive validity of our item pool was satisfying. Concretely, from the 1864 sentences that described the students' thoughts related to item-solving processes, there were only 291 sentences that did not provide relevant information about

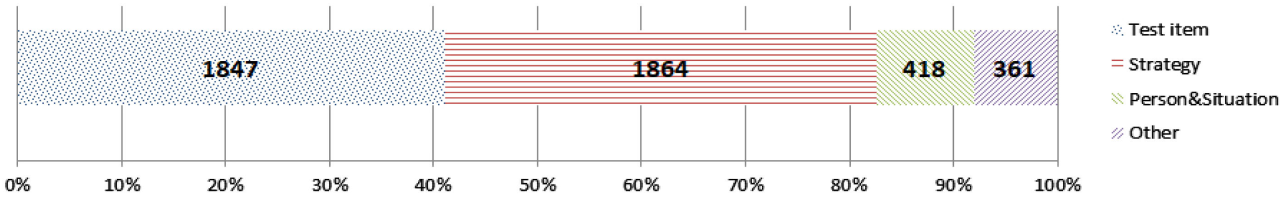


FIG. 2. Distribution of coded segments with respect to the main coding categories—a summary for the pool of 24 items.

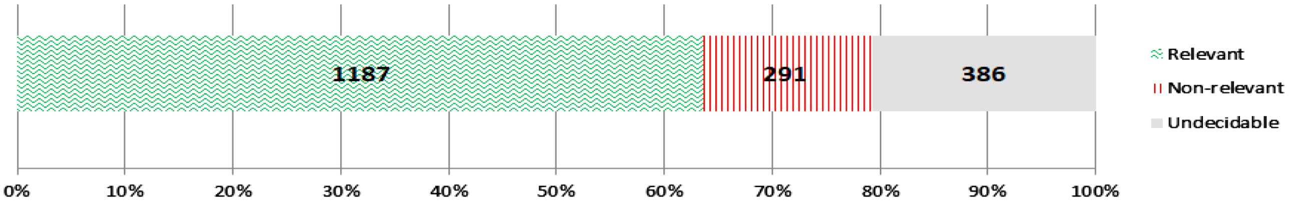


FIG. 3. Distribution of solving strategy categories—a summary for the pool of 24 items.

students' level of proficiency on the measured construct. In addition, we checked the use of construct relevant strategies for each item included in the think-aloud interviews, and drew conclusions that could be used for further improvement of potentially problematic items. In the third stage of item try out we administered the improved version of our item pool to physics students. Similarly, as in the stage 1 of the item try out, we had to schedule two appointments in order to survey our students about all the items from the item pool. At the first appointment, 15 physics students (14 third year students and 1 second year student) answered questions about the 24 basic of wave optics items, and at the second appointment 17 physics students (16 third year students and 1 second year student) answered questions about the remaining 37 items. For each of the 61 items from our pool, students were asked several questions with the purpose to gain additional insight into technical characteristics of items and their cognitive validity (Table II).

Based on the obtained data we calculated some classical psychometric measures (item difficulty index, point biserial, KR-20) for our items. The item difficulty index represents the proportion of correct answers on a given item [54], whereas the point biserial reflects the item by test-score correlation and is a measure of item discrimination [33]. KR-20 is a special case of the widely known Cronbach's alpha that is derived when items are measured exclusively on a dichotomous level [55]. Similar as Cronbach's alpha the KR-20 coefficient primarily measures the internal consistency of a test. In classical test theory the square root of the Cronbach's alpha value represents the correlation of the observed person scores and their errorless true scores [56]. Although the average difficulty index of our items was relatively low ($DI = 0.31$), most of the items proved to have satisfying discrimination properties and KR-20 amounted to 0.91 which was an indicator of very good reliability.

Based on the results of item try out, we attempted to improve wording and clarity of items that have been detected as potentially problematic. In addition, we decided to add four more items in order to ensure a better coverage of the model of the learning path. Thus we obtained an item pool consisting of 65 wave optics items that could be put to final field testing.

D. Final field testing of the item pool: characteristics of the student sample and curriculum

According to Ding *et al.* [51] and Maloney *et al.* [57] the final field testing of an assessment instrument should include a "postinstruction" population. When it comes to the sample size, Liu [26] suggests that field testing should include 5–10 times more respondents than the number of items. Linacre [30] even suggests that a "*sample of 50 well-targeted examinees is conservative for obtaining useful, stable estimates*," and 100 is an adequate size for most purposes, if our aim is to obtain item calibrations and person measures stable within ±0.5 logits. A sample size of such a magnitude has been already used by Szabo [4] for purposes of constructing a language test item bank.

Taking into account the relatively large size of our item pool ($N_i = 65$), as well as the practical difficulties that we faced in attempting to get access to a student sample of

TABLE II. Excerpt from the written student survey used for purposes of small scale item try out.

1. In the given test item underline every word (including response options) that you do not understand or do not know.
2. Do you think that some parts of the test item are confusing? If so, please explain briefly what confuses you.
Yes No
4. How confident are you in your solution of the given test item?

A: I am not confident in my solution at all. I cannot discard any of the response options and my answers are pure guessing.	B: Mostly I am not confident in my solution. I can merely discard the response option __.	C: Mostly I am not confident in my solution. I can merely discard the response option __.	D: I am completely confident in my solution.

5. Are you familiar with the content covered by the test item?
(a) Yes, I learned about them within the university course __________ (b) Yes, I learned about them in my free time (c) No
6. How much mental effort have you invested in solving the given test item?
1 (very little) 2 3 4 5 6 7 8 9 (very much)
7. How much difficult is the test item?
1 (not at all) 2 3 4 5 6 7 8 9 (very difficult)

appropriate size, we decided to conduct the item banking process through multiple stages [4]. Within the first stage of item banking a representative sample of 35 conceptual items was supposed to be evaluated. In order to ensure fairly stable parameter estimates we had to gather a sample of minimally 100 respondents who already had completed at least one university course that covered wave optics at the introductory level. In order to be in a better position to get access to a large student sample, we prepared the item pool in Bosnian, Croatian, Serbian, Slovenian, German, and English language. Eventually, the field testing of our first set of 35 items has been implemented in the period between April and July 2018 at University of Sarajevo (UNSA, Bosnia and Herzegovina), University of Zagreb (UNZA, Croatia), University of Maribor (UNMA, Slovenia), University of Rijeka (UNRI, Croatia), and University of Split (UNSP, Croatia). Altogether, the field testing included 188 physics and engineering students who agreed to anonymously volunteer in our study. The demographic section of the survey consisted of 14 items followed by 35 wave optics items subdivided into two booklets (Supplemental Material D [58]). Most students completed the survey in 45 min.

In Table III, we present the demographic features for the entire student sample and by area of study.

It is interesting to note that students consistently claimed their performance in optics was lower compared to their performance in introductory physics, in general. Only 14 out of 188 students claimed that their optics performance was better than their introductory physics performance, in general. In addition, only 17 out of 81 students who earlier attended physics competitions rated their optics performance as good (15 students) or very good (2 students).

Finally, for reasoning about the characteristics of the target population and possibilities of generalizing the results of this study, we analyzed the wave physics curricula at the sampled universities. Unsurprisingly, it could be shown that the wave physics curricula were much more extensive for physics study programs than for electrical engineering study programs (Supplemental Material E [59]).

E. Rasch modeling

Next, we had to check which of the 35 field-tested items are characterized by good psychometric features and can be combined with other items to give a scale that measures conceptual understanding of wave optics. These checks were mainly performed within the Rasch modeling framework. For purposes of Rasch modeling we used the WINSTEPS 4.0.0 software [33].

When performing Rasch modeling we attempted to follow the guidelines by Linacre [60] who recommended to firstly check for negative point-biserial correlations, followed by inspections of infit and outfit statistics, and a check for multidimensionality. The modeling procedure is iterative in its nature, and its goal is to obtain a Rasch-conform set of measures [4,26]. Concretely, our Rasch modeling approach to developing a scale for measuring understanding of wave optics was as follows:

(1) Taking into account that the wave physics curricula for physics studies were much more comprehensive than wave physics curricula in engineering studies, we first wanted to check whether our items function in the same way for the subsamples of physics and engineering students. A Rasch calibration that included all 188 students showed that in even 18 out of 35 items the differential item functioning (DIF) contrast was larger than 0.64 logits for the comparison of subsamples of physics and engineering students. According to Linacre [33] these differences are considered large. Generally, physics students largely outperformed the engineering students and classical indices of reliability such as the Cronbach's alpha were much larger for the physics subsample than for the engineering subsample. Consequently, we decided to continue the Rasch modeling process based on physics students' data only. A consequence of this approach is that the developed scale is most suitable for measuring conceptual understanding of wave optics in physics students.

(2) In the second step we reran the Rasch modeling procedure only on data obtained from the physics students ($n_1 = 119$). According to Szabo [4] a good practice in Rasch-based item banking is to first check for person misfit. We identified two persons with negative point-biserial correlations and deleted them from our database.

(3) We reran the Rasch modeling procedure and found that the item 15B had a negative point-biserial correlation. Consequently, this item has been deleted.

(4) In the next run of the Rasch modeling procedure we identified one more person with a negative point-biserial correlation and deleted the person from the database.

(5) The next Rasch calibration showed that the point-biserial statistics for items 10A and 9B was below 0.2. Consequently, these items were deleted from our database.

(6) Our final Rasch calibration (based on 116 out of 119 students and 32 and 35 items) showed that the point-biserial statistics was above 0.2 and item infit and outfit statistics were between 0.7 and 1.3 for all 32 items.

For practical applications and item banking it is of critical importance to obtain invariant, i.e., sample independent item difficulty parameters. An extreme approach to checking difficulty parameter invariance is to rank the respondents based on the Rasch ability measure, and to run separate Rasch calibrations with students from the upper half (higher proficiency subsample) and students from the bottom half (lower proficiency subsample) of the rank list.

TABLE III. Demographic features for the student sample in the field testing stage of the study.

	Total ($N = 188$) % of valid answers	Physics students ($n_1 = 119$) % of valid answers	Engineering students ($n_2 = 69$) % of valid answers
University			
Zagreb (UNZA)	47.9	60.5	26.1
Maribor (UNMA)	35.1	12.6	73.9
Sarajevo (UNSA)	8.5	13.4	···
Rijeka (UNRI)	6.4	10.1	···
Split (UNSP)	2.1	3.4	···
Age			
19–20	38.8	17.3	74.6
21–23	38.8	48.0	23.7
24–30	22.3	34.7	1.7
Gender			
Male	66.3	53.4	88.4
Female	33.7	46.6	11.6
Year of study			
1st year	28.7	2.5	73.9
2nd year	22.9	31.1	8.7
3rd year	9.6	6.7	14.5
4th year	12.2	19.3	···
5th year	26.6	40.3	2.9
Self-reported proficiency in physics			
Very good	8.6	10.9	4.4
Good	29.9	31.9	26.5
Satisfactory	49.2	48.7	50.0
Just sufficient	10.7	5.9	19.1
Poor	1.6	2.5	···
Self-reported proficiency in optics			
Very good	3.8	5.9	···
Good	19.0	23.7	10.6
Satisfactory	47.8	45.8	51.5
Just sufficient	21.8	16.9	31.3
Poor	7.6	7.6	7.6
Time passed since learning wave optics?			
Less than 3 months	51.6	41.5	69.1
3–6 months	19.4	24.6	10.3
6–12 months	8.1	10.2	4.4
12–24 months	12.4	11.9	13.2
More than 24 months	8.6	11.9	2.9
Participation in physics competitions?			
Yes	43.4	37.8	52.9
No	56.6	62.2	47.1
Participation in mathematics competitions?			
Yes	57.2	45.4	77.9
No	42.8	54.6	22.1

The estimates of item difficulties from these two separate calibrations are then cross plotted and it is checked whether some of the points in the scatterplot are outside of the 95% confidence band that is presented in the same plot [28].

According to Linacre [60], after checking for negative point biserials and item fit statistics, it is necessary to check for multidimensionality. Concretely, the Rasch model is based on the assumptions of unidimensionality and local independence of items [27]. In other words, our items are supposed to measure a single construct and there should not be a substantive relationship between residuals (differences between observed data and data expected from the Rasch model).

When it comes to checking for unidimensionality, we decided to apply the Bejar's method [61–63]. This method requires us to first run a Rasch calibration with all items (in our case 32 items) included, with the aim to estimate the difficulty parameter for each individual item. Then we are supposed to divide the sample of items into two as much as possible dissimilar subtests (subtest A and subtest B), and to run two separate Rasch calibrations on these subtests. Consequently, for each item from our item pool we get two difficulty estimates one based on the whole item pool calibration and one based on the subtest calibration. Finally, two scatterplots are drawn, showing the relationship between whole item pool estimates and subtest estimates for subtest A items and subtest B items, respectively. If the trend lines in the obtained cross plots are parallel to the identity line (line with a slope of 1), we have evidence for unidimensionality.

In our study, subtests A and B have been obtained based on our model of the learning path (Supplemental Material B [45]). Subtest A covered all the items associated with content area 1 (i.e., fundamentals of wave optics) of the learning path model, as well as the item 8B that had been initially associated with content area 3 (i.e., double-slit interference) from that model, although being primarily related to determining path length difference. For example, subtest A covered concept of phase and phase difference, optical path length and path length difference, wave front and sinusoidal representation, coherence and general concept of superposition. Subtest B covered goals associated with the remaining content areas from the model of the learning path (applications of basic concepts mainly related to analysis of patterns in single-slit, double-slit, circular-slit, rectangular-slit interference and diffraction). Eventually, subtest A (basics of wave optics) consisted of the following items: 1A, 2A, 3A, 4A, 5A, 6A, 7A, 8A, 18A, 1B, 2B, 3B, 4B, 5B, 8B, 13B. On the other hand, subtest B (applications) included the following items: 9A, 11A, 12A, 13A, 14A, 15A, 16A, 17A, 19A, 20A, 6B, 7B, 10B, 11B, 12B, 14B.

The assumption of local item independence has been checked by inspection of largest standardized residual correlations. Thereby, standardized residuals represent those parts of the data not explained by the Rasch model [33], and standardized residual correlations are correlations between these residuals. According to Linacre [33] standardized between-item correlations that are positive and larger than 0.7 may be source of concern.

F. Item banking

Items that exhibited good psychometric features (item fit statistics between 0.7 and 1.3) have been added to the item bank. In line with the system of data storage suggested by Szabo [4], we decided to include the following information in our item bank: Item ID, Rasch difficulty measure, standard error for the difficulty measure, infit statistics, outfit statistics, point-biserial measure, item stem and response options, and related goal in the model of the learning path.

IV. RESULTS

In this section summative and diagnostic use of the wave optics item pool is discussed separately. First, we report the results of Rasch modeling and describe the characteristics of the 32-items scale [wave optics test (WOT)] developed for measuring conceptual understanding of wave optics. In the second part of this section, we report our findings on the diagnostic potential of the field-tested set of conceptual questions.

A. Using the wave optics item pool for constructing a measurement scale

1. Unidimensionality and local item independence

As has been already described, unidimensionality has been checked by means of Bejar's method. Figure 4 shows the cross plots of full-test-based and subtest-based estimates of item difficulties, for basics of wave optics items and applications of wave optics items, respectively. If the points lie on a line that is parallel to the identity line this can be taken as evidence for unidimensionality of the scale [61–63]. Specifically, if the slope coefficient of the trend line is nearly 1 then the relative item difficulties are nearly the same for the full-test-based and subtest-based scale. This implies that the difference in log-odds for passing two items does not depend on which other items are included in the test. Consequently, we can say that a single trait explains performance on all items.

Next, the assumption of local item independence has been checked by inspecting the largest residual correlations. It has been found that only for two pairs of items (13A, 16A; 13B, 18A) the standardized residual correlation was higher than 0.3. Concretely, it amounted to 0.36 and 0.31, respectively. Linacre [33] points out that high local dependency of items is characterized by correlations above 0.7.

2. Reliability

The person separation coefficient amounted to 1.62, whereas the item separation coefficient amounted to 3.77. These values correspond to reliabilities of 0.72 and 0.93, respectively.

Here the separation coefficient is defined as the number of statistically different performance strata that the test can identify in the sample [64], whereas the value of person reliability coefficient can be interpreted in a similar way as Cronbach's alpha [33]. Further, the strata can be defined as statistically distinct measures [65]. Linacre [33] offers the

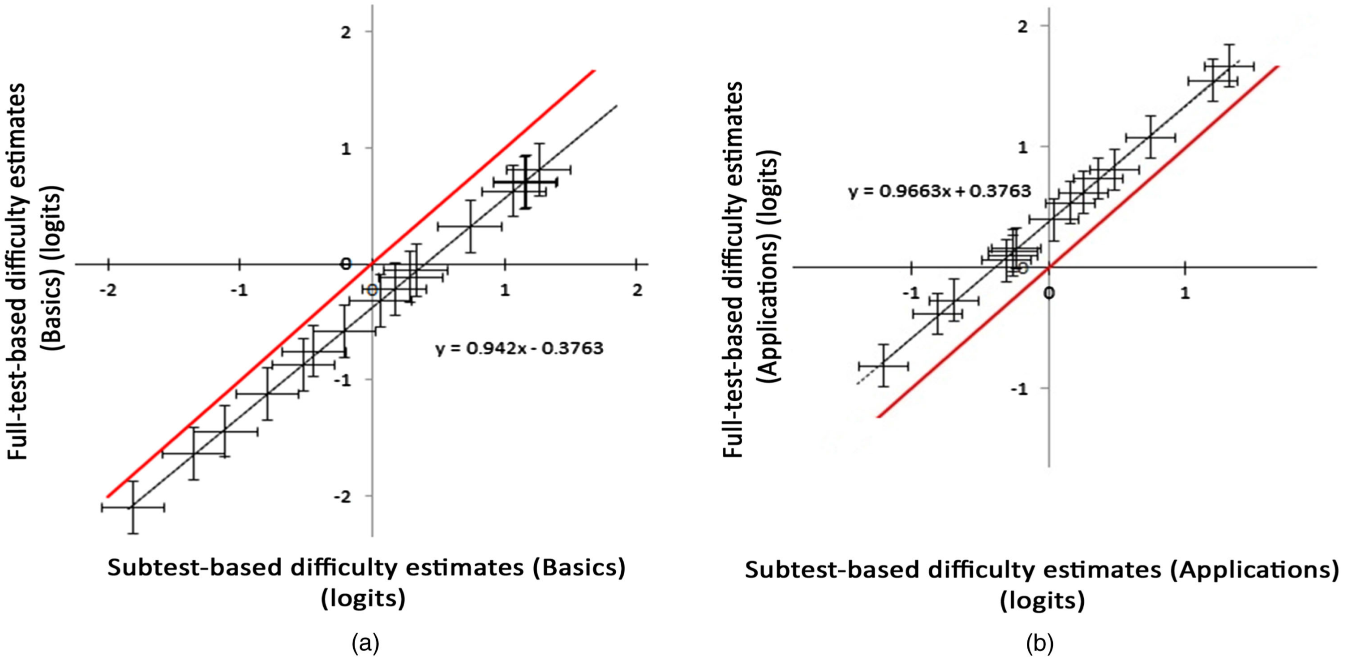


FIG. 4. Scatterplots for full-test-based versus subtest-based WOT item estimates. Identity line passes through the origin and is red in color. Cross hairs represent 68% confidence intervals. (a) Scatterplot for items related to basic wave optics concepts. (b) Scatterplot for items related to applications of wave optics concepts.

following guidelines for relating reliability coefficients to the number of performance levels (i.e., strata) the test can discriminate in the given sample: 0.9 = 3 or 4 levels; 0.8 = 2 or 3 levels; 0.5 = 1 or 2 levels. Generally, low person separation with a relevant person sample indicates that the instrument may not be sensitive enough to distinguish between low and high performers, whereas a low item separation reflects problems with confirming the item difficulty hierarchy.

3. Item fit statistics and item difficulty invariance

For our 32 items, the infit mean square (MNSQ) statistics ranged from 0.86 to 1.12, whereas the outfit mean square (MNSQ) statistics ranged from 0.71 to 1.26. No item had a standardized (ZSTD) infit or outfit statistic outside the range from -2.0 to 2.0.

For practical applications it is of high importance that the item difficulty parameters are invariant, i.e., independent of the concrete sample of respondents from a particular population. In order to check for item difficulty invariance we ranked our respondents based on ability and created two subsamples—58 students from the upper-half of the rank list were the high-proficiency subsample and 58 students from the lower-half were the low-proficiency subsample. Figure 5 shows the cross plot of item difficulties estimated in two separate calibrations, with the high-proficiency and low-proficiency group, respectively.

It is desirable that the points in the cross plot of item difficulties are not outside the 95% confidence band [28].

4. Person-item targeting

Person-item targeting can be best discussed based on the Wright's map (Fig. 6) which can be defined as a *"combined item difficulty and examinee ability diagram showing the distribution of items and examinees along a same unidimensional logit scale"* [26]. This map allows for

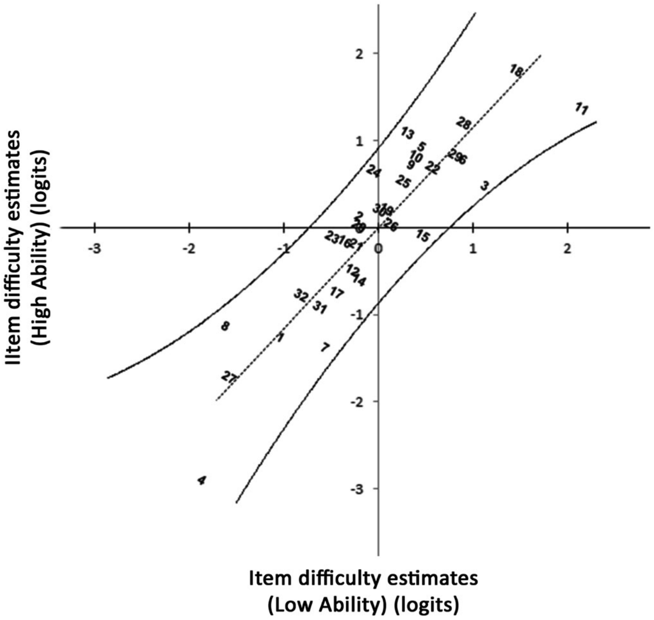
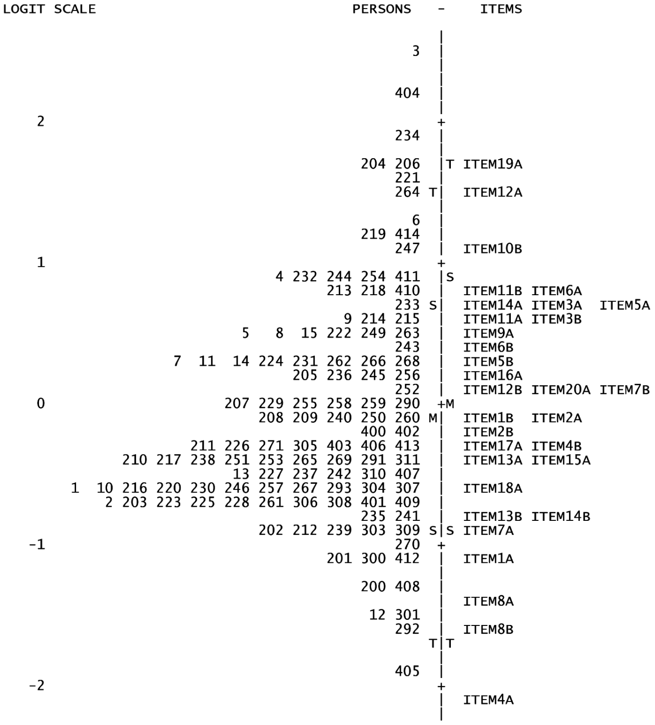


FIG. 5. Evidence for item difficulty invariance. Numbers represent items and a 95% confidence band is shown.



```
LOGIT SCALE                              PERSONS   -    ITEMS

                                                   |
                                                3  |
                                                   |
                                                   |
                                              404  |
                                                   |
    2                                              +
                                              234  |
                                                   |
                                          204 206  |T ITEM19A
                                              221  |
                                              264 T|  ITEM12A
                                                   |
                                                6  |
                                          219 414  |
                                              247  |  ITEM10B
    1                                              +
                                4 232 244 254 411  |S
                                      213 218 410  |  ITEM11B ITEM6A
                                              233 S|  ITEM14A ITEM3A  ITEM5A
                                        9 214 215  |  ITEM11A ITEM3B
                            5   8  15 222 249 263  |  ITEM9A
                                              243  |  ITEM6B
                    7  11  14 224 231 262 266 268  |  ITEM5B
                                  205 236 245 256  |  ITEM16A
                                              252  |  ITEM12B ITEM20A ITEM7B
    0                     207 229 255 258 259 290  +M
                              208 209 240 250 260 M|  ITEM1B  ITEM2A
                                          400 402  |  ITEM2B
                      211 226 271 305 403 406 413  |  ITEM17A ITEM4B
              210 217 238 251 253 265 269 291 311  |  ITEM13A ITEM15A
                           13 227 237 242 310 407  |
        1  10 216 220 230 246 257 267 293 304 307  |  ITEM18A
            2 203 223 225 228 261 306 308 401 409  |
                                          235 241  |  ITEM13B ITEM14B
                              202 212 239 303 309 S|S ITEM7A
   -1                                         270  +
                                      201 300 412  |  ITEM1A
                                                   |
                                          200 408  |
                                                   |  ITEM8A
                                           12 301  |
                                              292  |  ITEM8B
                                                  T|T
                                                   |
                                              405  |
   -2                                              +
                                                   |  ITEM4A
                                                   |
```

FIG. 6. Wright's map for inspecting person-item targeting and empirical hierarchy of item difficulties.

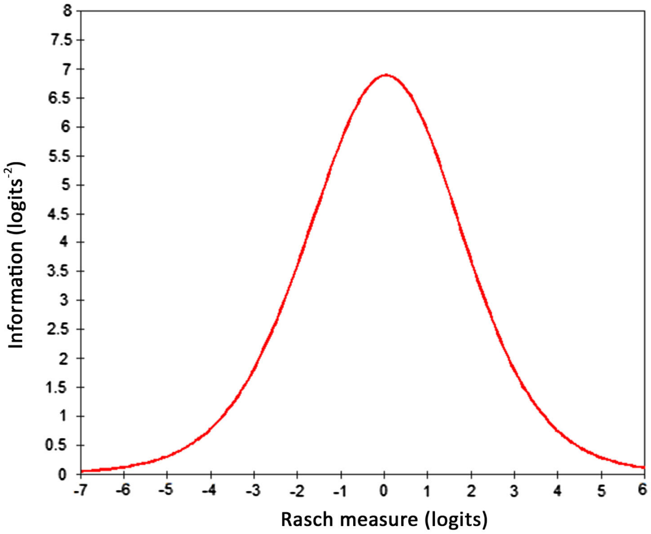


FIG. 7. The test information function for the 32-items scale.

a comparison between respondents and items. Typically, the hierarchy of respondents is shown in the left part of the map and the hierarchy of items is shown in the right part of the map. The difficulty of items and the ability of persons increase if we go from the bottom to the top of the map [29].

According to Linacre [33] good tests typically have the items targeted, i.e., lined up, with the persons. In addition, Linacre points out that it is desirable to have an even spread of items along the y axis, with no gaps. Additional insight into person-test targeting can be gained through inspection of the test information function (Fig. 7).

Note that a higher information value corresponds to more precise measurement, i.e., lower standard errors.

5. Rasch ability and demographic characteristics of respondents

Finally, we decided to check whether there is a relationship between the Rasch ability measure and following variables: gender (1-male, 2-female), self-reported performance in introductory physics (SPIP—1-very good to 5-poor), self-reported performance in optics (SPO—1-very good to 5-poor), time passed since having learned about wave optics for last time (TIMEP—1-less than 3 months to 5-more than 24 months), participation in physics competitions (PPC—1-yes, 2-no), participation in mathematics competitions (PMC—1-yes, 2-no). Results of correlational analyses are presented in Table IV.

The self-reported performance was higher for male (reverse coded M: 3.52, SD: 0.97) than for female students (reverse coded M: 3.34, SD: 0.73) and the average Rasch ability of male students (M: 0.11, SD: 0.87) was also higher than the Rasch ability of female students (M: −0.34, SD: 0.69).

B. Item-level analyses

In the previous Sec. IV-A we have shown that 32 out of 35 field-tested items can be combined into a scale that measures conceptual understanding of wave optics in physics students.

Taking into account that the items functioned very differently for physics and engineering students, we decided to construct the scale only for summative use with physics students. However, when items are applied for diagnostic uses *"the focus is on analyses of students' selections of incorrect answers and the analysis is typically done on the individual item level"* [26]. Consequently, we decided to base our analysis of students' difficulties and misconceptions on the whole sample of 35 items, and all 188 students.

TABLE IV. Relationship between Rasch ability and demographic characteristics of respondents.

	Gender	SPIP	SPO	TIMEP	PPC	PMC
Rasch ability measure	$r = -0.27$ $p = 0.004$	$r = -0.21$ $p = 0.021$	$r = -0.12$ $p = 0.221$	$r = -0.19$ $p = 0.040$	$r = -0.02$ $p = 0.793$	$r = 0.06$ $p = 0.523$

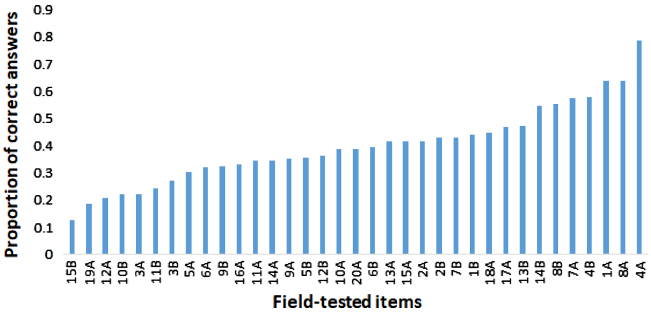


FIG. 8. Proportion of correct answers for all 35 field-tested items.

1. Difficulty of the field-tested items

An additional insight into the difficulty level of field-tested items is provided in Fig. 8.

The mean item difficulty index for the entire sample was 0.4 and only the difficulty of items 15B and 19A was outside the recommended range of 0.2 to 0.8 [66]). Items 4A, 8A, and 1A proved to be easier compared to other items from our item pool.

In general, the items were substantively easier for physics students (average item difficulty index $= 0.44$) than for engineering students (average item difficulty index $= 0.33$). The results of Fisher's exact test [67] showed that the proportion of correct answers was significantly higher ($p < 0.0001$) for physics students than for engineering students.

2. Results of distractor analysis

The purpose of the distractor analysis was to provide us additional insight into the quality of our distractors. In all of our items there was a single correct answer and three distractors. For multiple-choice items with three distractors it is desirable that each of the distractors is selected by at least 5% of the student sample [68]. In our study, 3 out of 105 distractors have not met this criterion. The response option "c" in item 1A has not been selected at all, whereas option "d" in the same item has been selected by only 2.1% of the respondents. In addition, the response option "b" in item 13A has been selected by only 4.8% of the sampled students.

Taking into account that the probability of choosing the correct answer by mere guessing was 25%, we decided to consider the distractors chosen by at least 35% of the respondents, as indicators of pronounced students' difficulties and misconceptions. We have detected 11 distractors that met this criterion (Table V).

TABLE V. A list of the most frequently chosen distractors.

Item 9A	Item 19A	Item 14A	Item 16A	Item 2A	Item 1B
a: 35.1%	c: 35.3%	a: 36.4%	d: 38.5%	c: 38.6%	b: 42.6%
Item 6A	Item 15B	Item 12A	Item 3B	Item 10B	
a: 42.6%	b: 43.5%	c: 44.6%	d: 45.4%	b: 47.5%	

V. DISCUSSION

A. Using the wave optics item pool for constructing a measurement scale

1. Unidimensionality, local item independence, and reliability

The results of Bejar's test do not indicate a violation of the unidimensionality assumption—the slope coefficients of the trend lines in Figs. 4(a) and 4(b) are very close to 1, which means that our test dominantly measures a single construct which is the conceptual understanding of wave optics.

Finally, the local item independency assumption proved to be also met for our dataset. For none of the 496 unique item pairs the standardized residual correlation was higher than 0.7. The largest standardized residual correlations were detected for the pairs "13A, 16A," and "18A, 13B." It is very interesting to note that the detected item pairs indeed thematized similar situations. In items 13A and 16A the students were expected to overcome the misconception that increasing slit separation or slit width results in wider fringes on the screen. On the other hand, in items 18A and 13B students were expected to apply the Huygens-Fresnel principle within the single-slit and sharp edge context, respectively.

The person reliability for our set of 32 items amounted to 0.72. Taking into account that this value is interpreted in the same way as the traditional Cronbach's alpha, we can conclude that the reliability of our wave optics test is at a satisfactory level [63]. Indeed, from the classical test theory perspective an obtained alpha value of 0.72 suggests that the correlation between observed person scores and their errorless true scores amounts to 0.85 which represents a strong relationship. Furthermore, the item reliability amounted to 0.93 which indicates a highly reproducible item difficulty hierarchy and speaks for the possibility to precisely locate the test items on the latent variable. Based on the tentative interpretation guidelines that relate reliability coefficients to number of performance strata [33], we can conclude that approximately two levels of student performance and four levels of item performance can be consistently identified for samples of items and respondents like the ones used in our field test.

2. Item-fit statistics and item difficulty invariance

Taking into account that the infit and outfit MNSQ values were all within the 0.7 to 1.3 range, we can conclude that our set of 32 conceptual items is characterized by very good item fit statistics [26].

Item difficulty invariance can be assessed through analysis of the item difficulty cross-plot from Fig. 5. It is evident that none of the points lies outside the 95% confidence band which suggests that the values of item difficulty parameters are not significantly different for the high-proficiency and low-proficiency student subsamples [28]. This is an important finding because it indicates that our item difficulty parameters are sample independent, which makes them suitable for item banking.

3. Person-item targeting and characteristics of the discarded items

Based on the inspection of the Wright's map it can be seen that the items are mostly lined up with persons which speaks for good person-item targeting. However, it should be noted that the item spread is not perfectly even. The most pronounced gaps are between approximately -0.5 and -0.7 logits, as well in the region above 1 logits. Concretely, in the region between -0.5 and -0.7 logits there are many persons, but only one item. Consequently, in this region of the latent trait the person ability is not measured very precisely. Based on the test information function, we can conclude that the test provided most precise measures at 0 logits which would make it a suitable cutoff score if the test were used for criterion-referenced assessment. Least precise measures are obtained in the low and high ability regions of the latent trait continuum. This issue may be resolved through the next stages of item banking in which an additional 30 items are going to be added to the wave optics item bank. We expect that at least some of these items will help to fill the identified gaps within the latent trait continuum, which will lead to more precise measurement.

Next, we will first demonstrate how to use the Wright's map for purposes of drawing conclusions about the empirical structure of students' achievement in wave optics and thereafter we will compare this empirical picture of students' achievement with our *a priori* model of the learning path in wave optics.

Based on visual inspection of gaps between larger item groupings in Wright's map, as well as based on theoretical considerations, we could identify approximately four groups of items. Empirically, this may be also related to the fact that our reliability analysis identified four different performance strata for the items. The first group of items starts with item 4A and ends with 18A. The second group starts with 15A and ends with 1B. Next, the third (largest) group starts with 7B and ends with 11B, whereas the fourth group consists of items 10B, 12A, and 19A. A common feature of items from the first (easiest) group is that they seem to require only low level transfer of knowledge. At this first level of understanding, students are able to determine the phase difference between two points on a sinusoid, and they associate the constancy in phase difference with coherence of laser light. Furthermore, they relate geometrical path length difference with occurrence of maxima or minima and they are able to identify the mathematical formula for path length difference between two waves in the context of double-slit interference. Finally, they are also able to apply the Huygens-Fresnel principle within the diffraction at single-slit and sharp edge contexts. At the next level of understanding students differentiate between geometric path length and optical path length, and they show basic understanding of the wave front representation (e.g., they are able to find the phase difference between two points). Furthermore, they also differentiate between diffraction patterns obtained from a single slit and optical grating. Finally, at this level of understanding students are also often able to correctly relate the width of the interference fringes with the separation between the two slits in double slit interference. The third group of items covered nearly all items that were related to analysis or prediction of patterns obtained in the settings of interference on thin film items, as well as most items related to single slit diffraction. Within the fourth (most difficult) group of items, students had to demonstrate understanding of patterns in combined interference and diffraction phenomena, as well as detailed understanding of the appearance of multiple slit interference patterns. Only a small number of students were found to understand the characteristics of the resulting electric field vector at the locations of minima and maxima in double-slit interference. We can conclude that the first two groups of items mostly corresponded to content area I (basics of wave optics) of our model of learning path, whereas the remaining two groups of items mostly corresponded to higher levels of the model of learning path. However, the match was not perfect—in the first two groups of items 5 out of 15 items were from higher levels of the model of the learning path, whereas in the remaining two groups of items 5 out of 17 items were from the first content area of the learning path model. Empirical results suggest that it would make more sense to put the *optical grating* content area immediately after *double-slit interference*. Items related to the superposition principle were mostly situated in a *two-point-source* context and it is eventually not surprising that these items were not easier than items related to double-slit interference. Although the content area of diffraction on two-dimensional objects has been *a priori* classified as most complex, the item 20A proved to be of average difficulty. In this item the students were required to reason about the relationship between the diameter of a circular obstacle and the diameter of the central diffraction maximum. Probably many students could solve this item through a simple analysis of the formula for diffraction on circular apertures, i.e., the item did not necessarily require the students to engage in far transfer activities. According to Kauertz [69] item difficulty depends not only on complexity of content but also on type of cognitive activity.

Although due to their poor psychometric characteristics, items 9B, 15B, and 10A were discarded from the scale development, such items can provide us often with valuable information about our curricula [42]. Indeed, items that are equally often correctly solved by high- and low-proficiency students could be an indicator of poor curriculum-test targeting. This would mean that content covered in items 9B, 15B, and 10A is probably not well covered in the sampled curricula (Supplemental Material E [59]). In both item 9B and item 15B, students were expected to judge whether or not the ray model of light is suitable for analyzing the presented phenomena. More specifically, they were expected to recognize that even at dimensions of an aperture or obstacle as large as one-tenth of a millimeter the diffraction effects are typically too large such that the ray model should not be applied [70]. On the other hand, if in double-slit interference we shorten the slits (along the y axis) from 3 to 1.5 mm the ray model of light is still reliable for predicting how light illuminates the screen along the y axis, which means that the interference fringes would become shorter (along the y axis). In item 15B the y dimension of the rectangular obstacle is 150 nm, and the x dimension is 0.15 mm. This means that diffraction happens in both dimensions, whereby it is much more pronounced along the y axis. Taking into account that the y dimension of the obstacle is even smaller than the wavelength of the used laser light, no interference minima should occur due to the diffraction along the y axis. However, interference minima occur due to diffraction along the x axis. A combination of these two effects gives rise to the diffraction pattern shown in distractor “d” of item 15B, which can be vividly simulated within the Webtop software [71]. When it comes to item 10A, it should be noted that it was the only item that was supposed to measure students’ understanding about the necessary conditions for occurrence of fringes of equal inclination.

4. Rasch ability and demographic characteristics of respondents

A statistically significant correlation of low size has been detected for the relationship between the Rasch ability of respondents and the following variables: gender, self-reported proficiency in introductory physics, and time passed since the respondent has learned about wave optics for the last time. In other words, on average, higher Rasch abilities are associated with being a male student, having reported higher proficiency in introductory physics, and having learned about wave optics more recently. The detection of the gender effect could be probably related to the fact that many of our items required performing visual tasks. In multiple earlier studies it has been found that there are significant between-gender differences when it comes to the activity of visuospatial reasoning [72].

On the other hand, for us it *was* somewhat surprising that our Rasch ability measure does not correlate significantly neither with the self-reported proficiency in optics nor with students’ participation in physics and mathematics competitions. When it comes to the relationship between the Rasch ability measure and proficiency in physics or optics, it is important to note that our correlation estimates would probably be more reliable if they had been based on more reliable and objective measures of proficiency. Indeed, much of earlier research suggests that for self-reported measures the level of reliability is often questionable [73]. When it comes to the relationship between Rasch ability and competitions, the low correlations could be probably explained by the fact that the category of “competitors” included those who competed at the school level, as well as those who competed at the level of the International Physics Olympiad.

5. Validity of score-based actions and inferences

In our study, we aimed to develop an instrument that may be used for measuring students’ understanding of wave optics. Thereby, we have intensively constructed validity arguments throughout the complete process of instrument development. Evidence about the content validity of our instrument has been obtained through two-stage expert surveys. Thereby, the sampled university professors mostly agreed that our test items cover most of the content (except the content of polarization) typically included in introductory courses of physics at the university, and they noticed that the item difficulty is appropriate for the level of university courses of physics. Information that has been gathered within this phase of validation has been used for purposes of additional improvement of the quality of our items. The improved version of the item pool has been tested for cognitive validity. Based on videotaped group interviews, audio-taped think-aloud interviews, and written surveys we found that our questions mostly function in line with our intents—students understand the questions and the questions mostly induce in students those cognitive processes that can be used as indicators of (mis)understanding of wave optics. Information gathered in this phase helped us to further improve the clarity of the item stems, as well as to significantly improve our distractors, which is important for ensuring validity of diagnostic testing [26].

In the next stage of validation, we analyzed the field test data and found that the hierarchy of items in the Wright map fairly follows our theoretical description of the learning pathway in wave optics—at least, it could be said that items related to basic wave optics concepts were more represented at the bottom of the map, while items related to more complex areas were more represented at the top of the map. This can be taken as additional evidence of construct validity. On the other hand, the correlational validity evidence proved to be relatively weak for our study and this requires further investigation.

Finally, it could be shown that our test provides more reliable measures for physics students than for engineering students. Consequently, it seems that it is more appropriate to use our test with physics students than with engineering students, at least at universities in Slovenia, Bosnia and Herzegovina, and Croatia. These results also indicate that the level of consistency of students’ knowledge structures can affect the internal consistency of a measurement instrument. Therefore it is recommended that the validity and reliability of this (and any other) test is investigated in various contexts.

B. Item-level analyses

1. Difficulty of the field-tested items

Although most item difficulty indices were inside the recommended range of 0.2 to 0.8 [66], it seems that the level of demand of our items was somewhat above the ability level of our student sample. However, it is also important to note that the targeting was better for physics students than for engineering students. Concretely, the average item difficulty index for the physics sample was 0.44 which is close to the optimal average item difficulty of 0.5 [54]. The differences between the performance of physics and engineering students were not surprising if we know that the wave optics curriculum was much more extensive for physics studies than for engineering studies (Supplemental Material E [59]).

2. Distractor analysis

Next, we will discuss the five most frequently chosen distractors from Table V.

From students' answers on item 6A, it follows that many students believe that two light waves that propagate in opposite directions along an x axis can constructively interfere in the same way as two light waves that propagate in the same direction along the same axis. Actually, only for waves propagating in the same direction the two waves' minima as well as maxima coincide all the time (e.g., this is the case all along the $\theta = 0°$ direction in double-slit interference). Generally, interference of two waves that propagate in opposite directions gives rise to standing wave patterns.

In item 15B, that has been already discussed, students were required to predict the diffraction pattern obtained on a rectangular obstacle. It seems that almost 44% of our sampled students believed that diffraction will happen only along the y axis (rectangle height of 150 nm), but not in the direction of x axis (rectangle width of 0.15 mm). This could be related to the fact that in conventional instruction students often hear that *"diffraction only occurs if the dimensions of the aperture/obstacle"* are similar to the light's wavelength. As has been earlier emphasized, diffraction effects may be pronounced even when the dimensions of the aperture or obstacle are one-tenth of a millimeter [70]. Furthermore, for large distances of the screen even larger dimensions of the obstacle or aperture can result in observable diffraction effects.

Based on students' answers to item 12A it follows that nearly 45% of students believe that the resultant electric field vector at locations of interference maxima is not changing at all over time.

In item 3B, most students believed that using a lens is the best way to increase coherence of light, even better than using a narrow aperture combined with a color filter. Earlier research has already found that students often do not understand the role of lenses in the context of wave optics phenomena [74].

In item 10B students were shown two interference patterns and they were expected to identify these patterns as three-slit and four-slit interference patterns (N-2 secondary maxima between two main maxima; N number of slits). However, nearly 50% of our students believed that they are shown two-slit and three-slit interference patterns. On the one hand, in this item it was probably relatively easy to eliminate option "a" because students are well familiar with the single slit pattern. However, instead of using the N-2 formula for determining the number of secondary maxima, it seems that many students used a N-1 formula.

C. Next steps in the item banking process

In this first stage, we have field tested 35 out of 65 wave optics items. The results of the field test showed that 32 out of 35 items exhibit good psychometric characteristics and can be combined into a scale that reliably measures understanding of wave optics in physics students. These 32 items had been added to our item bank which contains information about item ID, Rasch difficulty measure, standard error for the difficulty measure, infit statistics, outfit statistics, point-biserial measure, item stem, and response options, and related goal in the model of the learning path. In the years that follow the remaining 30 items will be field tested. Thereby they will be combined with the most precisely estimated items from the existing item bank, i.e., an item anchor-test design will be used [27]. This will allow setting all the difficulty measures on the same scale.

It is important to point out that the process of item banking can be continued even after field testing of all 65 items. By providing open access to members of the physics education community, everyone could be given the opportunity to contribute new items and to help in their field testing and adding to the item bank.

D. Limitations of the study

The final Rasch calibration was based on the answers of only 116 physics students. According to Szabo [4] such a sample size is sufficient for Rasch-based item banking, although using larger samples would certainly have helped us to obtain more stable estimates of the item difficulty parameters. It is important to note that the current scale does not provide very precise measures in the low- and high-ability regions of the latent trait continuum. The 32-item scale is mostly suitable for measuring understanding in physics students and with other student populations it may be used for purposes of uncovering misconceptions only. However, in future research we will try to identify a subset of these 32 items (or 65 items) that may be also used for summative purposes with engineering students.

Another important limitation of this study is that it failed to provide sufficient evidence on criterion validity for the developed scale. However, validation is a continuous process and in our future research we will attempt to provide more evidence on criterion (correlational) validity. In this stage, we found only a small correlation between respondents' Rasch ability and self-reported proficiency in introductory physics. This could be related to the fact that self-reported measures are often less reliable than objective measures [73], which negatively impacted our correlational analyses. Finally, it has been also found that there is room for improvement of distractors in items 1A and 13A.

VI. CONCLUSION

In this paper we described how the Rasch model can be applied for purposes of wave optics item bank building. Thereby, the content domain included all the content typically covered in introductory physics courses at the university level, apart from scattering and polarization. The distractors for each item were based on results of earlier studies, as well as on group interviews, and student think-aloud interviews. For purposes of ensuring content validity we implemented and analyzed multiple expert surveys, whereas cognitive validity has been checked through analyses of group interviews, think-aloud interviews, and written student surveys. Final field testing has been performed for 35 out of 65 items. The evaluation of the field-tested items was mostly implemented within the framework of the Rasch model. It has been shown that the Rasch model assumptions were met and that 32 out of 35 field-tested items can be combined into a scale that relatively reliably measures physics students' conceptual understanding of wave optics. The technical features of the developed scale proved to be promising—the item-fit statistics was very good and the item difficulty parameters proved to be sample independent. Furthermore, the match between the Wright's map representation of item difficulty hierarchy and our model of learning path was fair, which provides further evidence for construct validity.

The developed 32-item scale may be used for measuring physics students' understanding of wave optics. Furthermore, all of the 35 items can be used for diagnostic purposes and for sparking classroom discussions about wave optics phenomena. Indeed, the field-tested items proved to be powerful when it comes to uncovering students' misconceptions. While some of the identified misconceptions were already detected in earlier studies (e.g., *increasing width of the slit results in increasing of the fringes' width*), other misconceptions were not reported in earlier research (e.g., *at the locations of maxima the resulting electric field vector is constant over time*).

In our future research we are going to further explore the range of contexts in which our measurement instrument can allow for drawing valid inferences about students' understanding of wave optics. Also we plan to extend the existing wave optics item bank and to investigate in more detail possible differential item functioning for male and female respondents. To that end it would be desirable to get access to larger samples of well-targeted respondents.

ACKNOWLEDGMENTS

The authors would like to thank all the students and experts who took part in this study. Special thanks to assistant Damjan Osrajnik, MSc, for his help in translation of the field-testing survey into the Slovenian language. I. A. was supported by the Ministry of Science and Education of the Republic of Croatia under the bilateral Croatia-USA Project No. 1/2014. This research was also supported by the Slovenian Research Agency (Grant No. P1-0403).

[1] M. McCloskey, Intuitive physics, Sci. Am. **248**, 122 (1983).

[2] E. F. Redish, *Teaching Physics with the Physics Suite* (Wiley, Hoboken, NJ, 2003).

[3] J. B. Bjorner, C. H. Chang, D. Thissen, and B. B. Reeve, Developing tailored instruments: Item banking and computerized adaptive assessment, Qual. Life Res. **16**, 95 (2007).

[4] G. Szabo, *Applying Item Response Theory in Language Test Item Bank Building* (Peter Lang, Frankfurt am Main, 2008).

[5] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[6] S. Messick, Validity, in *Educational Measurement*, edited by R. L. Linn (Macmillan, New York, NY, 1989), pp. 13–103.

[7] C. Dedes, The mechanism of vision: Conceptual similarities between historical models and children's representations, Sci. Educ. **14**, 699 (2005).

[8] P. Hubber, Year 12 students' mental models of the nature of light, Res. Sci. Educ. **36**, 419 (2006).

[9] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics Extended*, 10th ed. (John Wiley and Sons, Hoboken, NJ, 2014).

[10] D. Giancoli, *Physics Principles with Applications* (Pearson Education, Upper Saddle River, NJ, 2005).

[11] H. D. Young and R. A. Freedman, *Sears and Zemansky's University Physics with Modern Physics: Technology Update* (Pearson Education, Upper Saddle River, NJ, 2014).

[12] R. Serway and J. Jewett, *Physics for Scientists and Engineers with Modern Physics* (Cengage Learning, Belmont, CA, 2014).

[13] R. D. Knight, *Physics for Scientists and Engineers: A Strategic Approach*, 3rd ed. (Pearson Education, Upper Saddle River, NJ, 2013).

[14] P. A. Tipler and G. Mosca, *Physics for Scientists and Engineers* (W. H. Freeman and Company, New York, 2008).

[15] M. Watts, Student conceptions of light: A case study, Phys. Educ. **20**, 183 (1985).

[16] N. J. Selley, Children's ideas on light and vision, Int. J. Sci. Educ. **18**, 713 (1996).

[17] B. S. Ambrose, P. R. L. Heron, S. Vokos, and L. C. McDermott, Student understanding of light as an electromagnetic wave: Relating the formalism to physical phenomena, Am. J. Phys. **67**, 891 (1999).

[18] B. S. Ambrose, P. S. Shaffer, R. N. Steinberg, and L. C. McDermott, An investigation of student understanding of single-slit diffraction and double-slit interference, Am. J. Phys. **67**, 146 (1999).

[19] S. K. Sengoren, How do turkish high school graduates use the wave theory of light to explain optics phenomena?, Phys. Educ. **45**, 253 (2010).

[20] V. Mesic, E. Hajder, K. Neumann, and N. Erceg, Comparing different approaches to visualizing light waves: An experimental study on teaching wave optics, Phys. Rev. Phys. Educ. Res. **12**, 010135 (2016).

[21] A. Coetzee and S. N. Imenda, Alternative conceptions held by first year physics students at a south african university of technology concerning interference and diffraction of waves, Res. Higher Educ. J. **16**, 1 (2012); http://www.aabri.com/manuscripts/121097.pdf.

[22] K. Wosilait, Research as a guide for the development of tutorials to improve student understanding of geometrical and wave optics, Ph.D. thesis, University of Washington, 1996.

[23] V. Mešić and N. Erceg, Različiti pristupi vizualiziranju superpozicije talasa u srednjoškolskoj nastavi talasne optike[different approaches to visualizing the superposition of light waves in high school instruction about wave optics], in *Proceedings of the 6th International Conference on Physics Instruction in High Schools*, edited by Lj. Nešić (Cicero, Beograd, 2018), pp. 133–141.

[24] R. D. Knight, *Five Easy Lessons: Strategies for Successful Physics Teaching* (Addison Wesley, San Francisco, CA, 2004).

[25] H. Schmid, *Psychologische Tests: Theorie und Konstruktion* (Huber, Bern, 1992).

[26] X. Liu, *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach* (IAP, Charlotte, NC, 2010).

[27] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory* (SAGE, Newbury Park, CA, 1991).

[28] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Routledge, New York, NY, 2015).

[29] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch Analysis in the Human Sciences* (Springer, Dordrecht, 2014).

[30] J. M. Linacre, Sample size and item calibration [or person measure] stability, Rasch Meas. Trans., **7**, 328 (1994); https://www.rasch.org/rmt/rmt74m.htm.

[31] T. F. McNamara, *Measuring Second Language Performance* (Addison Wesley, San Francisco, 1996).

[32] J. M. Linacre and B. D. Wright, The length of a logit, Rasch Meas. Trans. **3**, 54 (1989); https://www.rasch.org/rmt/rmt32b.htm.

[33] J. M. Linacre, *A User's Guide to WINSTEPS: Rasch-Model Computer Program* (MESA Press, Chicago, IL, 2017).

[34] J. M. Linacre, What do infit and outfit, mean-square and standardized mean, Rasch Meas. Trans. **16**, 878 (2002); https://www.rasch.org/rmt/rmt162f.htm.

[35] R. R. Meijer and K. Sitsma, Person-fit statistic—what is their purpose, Rasch Meas. Trans. **15**, 823 (2001); https://www.rasch.org/rmt/rmt152d.htm.

[36] F. M. Lord and R. M. Novick, *Statistical Theories of Mental Test Scores* (IAP, Charlotte, NC, 2008).

[37] R. Baker, Classical test theory and item response theory in test analysis, Technical Report, Special Report No 2: Language Testing Update, 1997.

[38] D. J. Weiss, Item banking, test development, and test delivery, in *The APA Handbook on Testing and Assessment*, edited by K. F. Geisinger (APA, Washington, DC, 2011), pp. 185–200.

[39] A. S. Willmott and D. E. Fowles, *The Objective Interpretation of Test Performance* (National Foundation for Educational Research, Slough, 1974).

[40] J. Rost, *Lehrbuch Testtheorie, Testkonstruktion* (Huber, Bern, 1996).

[41] H. Moosbrugger and A. Kelava, *Testtheorie und Fragebogenkonstruktion* (Springer, Heidelberg, 2008).

[42] S. B. McKagan, K. K. Perkins, and C. E. Wieman, Design and validation of the quantum mechanics conceptual survey, Phys. Rev. ST Phys. Educ. Res. **6**, 020121 (2010).

[43] K. Wosilait, P. R. Heron, P. S. Shaffer, and L. C. McDermott, Addressing student difficulties in applying a wave model to the interference and diffraction of light, Am. J. Phys. **67**, S5 (1999).

[44] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010115, for a list of common difficulties and misconceptions in wave optics.

[45] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010115, for a more detailed description of the learning path model.

[46] J. W. Pellegrino, N. Chudowsky, and R. Glaser, *Knowing what students know: The science and design of educational assessment* (National Academy Press, Washington, DC, 2001).

[47] K. Ford, Inquiry Learning: Students perception of light wave phenomena in an informal environment, Ph.D. thesis, Southern University and A&M College, Baton Rouge, LA, 2011.

[48] J. Michael and H. I. Modell, *Active Learning in Secondary and College Science Classrooms: A Working Model for Helping the Learner to Learn* (Routledge, London, 2003).

[49] M. Wilson, *Constructing Measures: An Item Response Modeling Approach* (Erlbaum, Mahwah, NJ, 2005).

[50] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010115, for more information about the questionnaire that has been used in the first stage of content validation.

[51] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[52] M. D. Smith, Cognitive validity: Can multiple-choice items tap historical thinking processes?, Am. Educ. Res. J. **54**, 1256 (2017).

[53] J. Daniel, *Sampling essentials: Practical guidelines for making sampling choices* (SAGE, Thousand Oaks, CA, 2011).

[54] R. J. Cohen and M. Swerdlik, *Psychological Assessment: An Introduction to Tests and Measurements* (McGraw-Hill Higher Education, Boston, MA, 2009).

[55] L. R. Price, *Psychometric Methods: Theory into Practice* (Guilford Publications, New York, 2016).

[56] J. C. Nunnally, *Psychometric Theory* (McGraw-Hill, New York, 1994).

[57] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[58] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010115, for the survey that has been administered for purposes of final field testing of 35 items.

[59] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010115, for obtaining information about the wave physics curricula at the sampled universities.

[60] J. M. Linacre, Detecting multidimensionality: Which residual data-type works best?, J. Outcome Meas. **2**, 266 (1998); http://europepmc.org/abstract/MED/9711024.

[61] I. I. Bejar, A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates, J. Educ. Measure. **17**, 283 (1980).

[62] C. S. Wallace and J. M. Bailey, Do concept inventories actually measure anything?, Astron. Educ. Rev. **9**, 010116 (2010).

[63] L. Ding, Seeking missing pieces in science concept assessments: Reevaluating the brief electricity and magnetism assessment through rasch analysis, Phys. Rev. ST Phys. Educ. Res. **10**, 010105 (2014).

[64] B. D. Wright, Reliability and Separation, Rasch Meas. Trans. **9**, 472 (1996); https://www.rasch.org/rmt/rmt94n.htm.

[65] B. D. Wright and G. N. Masters, Number of person or item strata: $(4 * \text{separation} + 1)/3$, Rasch Meas. Trans. **16**, 888 (2002); https://www.rasch.org/rmt/rmt163f.htm.

[66] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Routledge, London, 2015).

[67] J. H. McDonald, *Handbook of Biological Statistics* (Sparky House Publishing, Baltimore, MD, 2014).

[68] T. J. B. Kline, *Psychological Testing: A Practical Approach to Design and Evaluation* (SAGE, Thousand Oaks, 2005).

[69] A. Kauertz, Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben, Ph.D. thesis, Universitaet Duisburg-Essen, 2007.

[70] J. Orear, *Physik Band 1 und Band 2 [Physics Volume 1 and Volume 2]* (Hanser Verlag, Muenchen, 1982).

[71] Webtop, J. Foley, T. Mzoughi, and D. Banks, WebTOP—The Optics Project, 2008, https://www.compadre.org/introphys/items/detail.cfm?ID=1678.

[72] D. F. Halpern, C. P. Benbow, D. C. Geary, R. C. Gur, J. S. Hyde, and M. A. Gernsbacher, The science of sex differences in science and mathematics, Psychol. Sci. Publ. Interest **8**, 1 (2007).

[73] M. A. Zimmerman, C. H. Caldwell, and D. H. Bernat, Discrepancy between self-report and school-record grade point average: Correlates with psychosocial outcomes among African American adolescents, J. Appl. Soc. Psychol. **32**, 86 (2002).

[74] P. Colin and L. Viennot, Using two models in optics: Students' difficulties and suggestions for teaching, Am. J. Phys. **69**, S36 (2001).