Rasch analysis in physics education research: Why measurement matters

Maja Planinic,[1,*] William J. Boone,[2] Ana Susac,[3] and Lana Ivanjek[4]

[1]*Department of Physics, Faculty of Science, University of Zagreb, Bijenicka c. 32, 10000 Zagreb, Croatia*
[2]*Department of Educational Psychology, McGuffey Hall, Miami University, Oxford, Ohio 45056, USA*
[3]*Department of Applied Physics, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia*
[4]*University of Vienna, Faculty of Physics, Porzellangasse 4, 1090 Vienna, Austria*

(Received 31 August 2018; published 3 July 2019)

[This paper is part of the Focused Collection on Quantitative Methods in PER: A Critical Examination.] The Rasch model is a probabilistic model which describes the interaction of persons (test takers or survey respondents) with test or survey items and is governed by two parameters: item difficulty and person ability. Rasch measurement parallels physical measurement processes by constructing and using linear person and item measures that are independent of the particular characteristics of the sample and the test items along a unidimensional construct. The model's properties make it especially suitable for test construction and evaluation as well as the development and use of surveys. The evaluation of item fit with the model can pinpoint problematic items and flag idiosyncratic respondents. The possibility of determining sample—independent item difficulties makes it possible to use the Rasch model for linking tests and tracking students' progression. The use of the Rasch model in PER is continuously increasing. We provide an overview and examples of its use and benefits, and also outline common mistakes or misconceptions made by researchers when considering the use of the Rasch model. We focus in particular on the question of how Rasch modeling can improve some common practices in PER, such as test construction, test evaluation, and calculation of student gain on PER diagnostic instruments.

DOI: 10.1103/PhysRevPhysEducRes.15.020111

I. INTRODUCTION

The pursuit of objective measurement lies at the heart of science, and physics education research (PER) as a science should also try to bring its measurements closer to the standards of objective measurement. The Rasch model was developed by Danish mathematician Georg Rasch with that particular purpose. The use of the Rasch model for data analysis is not new to PER practitioners. The Rasch model has been used in a number of PER studies to date (e.g., [1–25]). However, it is likely that the Rasch approach is not generally well understood and that many researchers do not understand the rationale and benefits for its use in physics education research. The intention of this article is to present the basic ideas of the Rasch model, and the motivation for using it in PER, but also to present some common misunderstandings of the model. We will also try to suggest some ways in which the use of the Rasch model could improve some common PER practices. These include construction and evaluation of diagnostic instruments (tests and surveys), linking of tests, monitoring learning progression, and measuring learning gain. Since it is not possible to cover all aspects of the Rasch model and its use in a single paper, we have chosen to refrain from technical aspects as much as it is possible and reasonable, to make the article easier to read and help readers primarily focus on the main conceptual issues. Readers who are interested in a more detailed and in-depth presentation of many technical aspects of Rasch analysis are referred to other sources (e.g., [26–30]). An important aspect of the Rasch model is that it is not just another statistical technique to apply to data, but it is a perspective as to what is measurement, why measurement matters, and how to achieve better quality measurement in an educational setting. After an introduction to these ideas and the Rasch model itself, we will outline the process of test construction and test evaluation with Rasch analysis and also provide some essential practical advice for novice analysts as how to avoid some common misunderstandings and pitfalls. In the end, we will discuss the common practice of using pretesting, posttesting, and normalized gain in PER and discuss them from a Rasch perspective.

[*]Corresponding author.
maja@phy.hr

II. BASIC PRINCIPLES AND PROBLEMS OF MEASUREMENT

A. Objective measurement

Natural science was built upon striving to achieve as objective measurements as possible, which means that objective methods were determined and utilized to transform observation into measurement. In physics, methods for measuring have been developed which are specific for the intended measurement and independent of the variation in the other characteristics of the measured object or the measuring instrument utilized. To achieve objective measurement, the calibration of measuring instruments must be independent of the objects used for calibration and the measurement of an object must be independent of the instrument that was used [28]. These conditions are satisfied to a large degree in most physics measurements. For example, if we want to measure the height of an object, we can use different measuring instruments, and obtain the same result with each instrument measuring height, within the limits of the associated measurement uncertainties of the different measurement instruments measuring height. The obtained result will not depend on other objects which we may have measured with the same instruments. Therefore, the obtained measure is, in principle, independent of the instrument used (one can refer to this as instrument-free measurement), as well as of other objects on which measurement was performed (one can name this sample-free measurement). However, this is in striking contrast with measurements routinely performed in education, which are strongly test and sample dependent. For example, one student may achieve 90% on one test, and 50% on another (more difficult) test, covering the same content, or be placed in the 80th or 60th percentile of their class. Students' ability estimates, obtained in this way, are clearly very dependent on both the characteristics of the test and the performance of other people in the reference group.

B. Unidimensionality

Objective measurement requires creation of unidimensional measurement scales [31]. Unidimensionality means that we are trying to describe and measure only one attribute of the phenomenon under observation at a time. In physics, scientists have succeeded in creating a great number of such unidimensional scales for the measurement of many different physics quantities (e.g., length, mass, temperature, etc.). Unidimensionality is important, since we can only understand the meaning of the obtained measure if we have clearly isolated one trait (the dimension) which is being measured. This may seem almost impossible to achieve in educational measurement, where there are so many factors and traits which seem to complicate each measurement. However, the Rasch perspective is that we still must try to measure a single trait. We know that there will always be noise in educational measurements, but steps can be taken to limit such noise. It is important to try to work toward unidimensionality, bearing in mind that unidimensionality is necessarily always an approximation, but one that can be empirically tested. Maximizing the quality of the unidimensional measure improves the quality and confidence we can have in the analysis of the collected test data.

When contemplating unidimensionality, it is important to distinguish between a psychometric dimension, expressed through persons' responses, and psychological dimensions, which may or may not be the same as the psychometric dimension [32]. For example, physics test problems presented to students will usually involve several psychological dimensions, e.g., physics knowledge, math knowledge, reading, etc. Although psychologically multidimensional, such problems can in many practical cases define a single empirical psychometric dimension. The main requirement is that the items work sufficiently together to define a trait [33]. There are many ways in Rasch analysis to test the unidimensionality of the dataset (the methods will be discussed later), to explore the possible presence of other dimensions, as well as their size and impact on measurement, and to enable the analyst to make the decision of whether the test is sufficiently unidimensional for their purpose, or if it needs further refinements.

C. Abstract measurement scale

Objective measurement requires that we move from simple counting to the construction of abstract continuous measurement scales. Counts are typically not measures. For example, we can count objects, but the same number of objects will not always imply the same underlying quantity, since objects can vary in size. To solve the problem, we have to resort to an abstract quantity (e.g., mass), and express its value on an abstract continuous scale with a measurement unit that has the same meaning on any part of the scale (e.g., kilogram). Measurement requires the construction of an abstract quantity, expressed in linear abstract units, whose meaning does not change along the scale [28,31]. It also requires a well-constructed and calibrated measurement instrument. In the example of measuring the mass of objects, such an instrument could be a spring scale. But, at the core of each instrument, there is a measurement model, which describes the interaction of the object of measurement with the instrument.

In educational measurement counts of correct answers are often used as measures, although they do not possess the necessary characteristics of measures. Counts or percentages of correct answers are not linear in the variable that they represent [27,28]. There are other related problems with counts and raw scores. Raw scores are limited to be between 0% and 100%, whereas linear measures do not have such bounds, and a score increase of 1% will not represent the same increase in ability along the entire percentage scale.

It is important to stress that any mathematical operations and statistical analyses performed with nonlinear measures may produce distorted results. It is important therefore to switch from nonlinear counts to abstract linear measures if one wishes to perform such mathematical operations and statistical analyses (e.g., calculating mean values, conducting t-tests, conducting analysis of variance, etc.).

III. THEORETICAL FOUNDATIONS OF RASCH MODELING

The construction of measures requires a model which can describe what happens when a test taker interacts with a test item (e.g., when a student attempts to answer an item), and which can produce a method for converting the counts of correct answers (often called the raw scores) to person ability measures. It is important to note that the Rasch meaning of the term "person ability" does not refer to the general intellectual ability of a person, but instead refers only to the degree of the latent trait under investigation (e.g., knowledge of mechanics, understanding of electromagnetism, attitude toward physics, etc.) possessed by the person. The first such model is the Rasch dichotomous model [34], which was later followed by other extensions of the Rasch model, such as the rating scale model [35], the partial credit model [36], and the many faceted model [37].

The requirement of unidimensionality assumes the existence of an underlying variable (sometimes named a "latent trait"), which is operationalized with a certain number of test items, each of which can be characterized by its difficulty (D_i). These items are used to determine where a person is located on the variable (e.g., are they at a high level on the variable or a low level on the variable; see Fig. 1).

Persons are described by a parameter called person ability (B_n). The measurement model should be probabilistic, and not deterministic, since persons of the same ability may respond differently to the same item, so we can only predict the outcome in terms of probabilities.

If a test taker of ability B_n answers a test item of difficulty D_i, in the simplest case of a dichotomous item, they can succeed or fail on the item. The probability of success will be some function of the difference $B_n - D_i$: the larger the difference, the higher the probability of success. Also, the function should be normalized to provide probability values between 0 and 1. Many functions that satisfy these conditions can be constructed, but Rasch showed that the logistic function

$$P(X = 1|B_n, D_i) = e^{(Bn-Di)}/(1 + e^{(Bn-Di)}) \tag{1}$$

is the only one that allows the separation of parameters, and their independent determination [28]. Equation (1) is known as the dichotomous Rasch model [34]. Of all proposed latent trait models, the Rasch model has fewest parameters: one ability parameter for each person, and one difficulty parameter for each item [38]. The Rasch model is most often expressed in terms of the log odds (L):

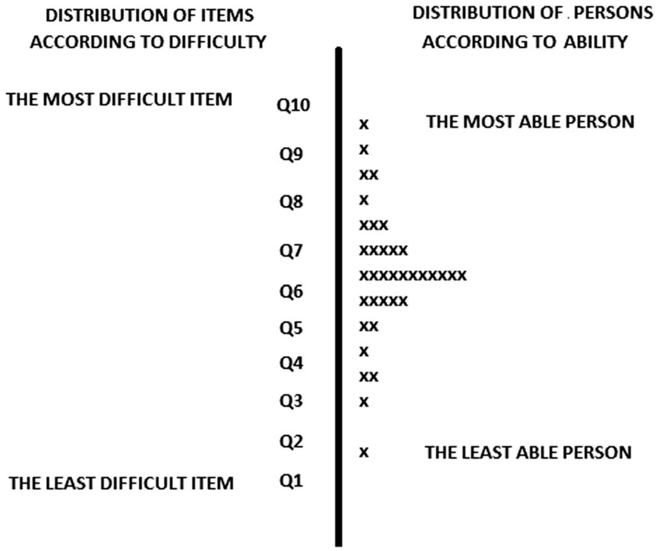

FIG. 1. An example of the Wright map (item-person map), representing a hypothetical distribution of items (Q1–Q10) and persons (each person is represented by "x") along the same variable. The variable is represented by a straight line in the middle and increases in upward direction.

$$L = \ln[P_{ni}/(1\text{-}P_{ni})] = B_n - D_i. \tag{2}$$

From Eq. (2) we can see that both B_n and D_i range from negative infinite to positive infinite. They also have some additional important properties:

$$L_{n1} - L_{n2} = B_{n1} - B_{n2} \quad (\text{for the same item of difficulty } D_i), \tag{3}$$

$$L_{i1} - L_{i2} = D_{i2} - D_{i1} \quad (\text{for the same person of ability } B_n). \tag{4}$$

Equations (3) and (4) demonstrate the mutual independence, as well as linearity, of person abilities and item difficulties expressed as log odds. An important aspect of Eq. (3) to note is that the difference of log odds for the two persons attempting the same item is determined only by the difference of the persons' abilities, and is not influenced by the item's difficulty. The same holds for the difference of item difficulties—item difficulties are independent of the ability of the person who answered the items. The Rasch model has therefore item and person invariance properties—unlike the commonly used "percentage correct" statistics which are strongly sample dependent and test dependent [30].

Through its mathematical form the Rasch model defines a general mathematical unit called the logit (log odds unit).

A person's ability in logits is their log odds of succeeding on items chosen to define the scale origin; for $D_i = 0$, ln $[P_{ni}/(1 - P_{ni})] = B_n$. An item's difficulty in logits is the log odds of failure on that item of persons with abilities at the scale origin; for $B_n = 0$, ln $[(1 - P_{ni})/P_{ni}] = D_i$. Since it is the difference $B_n - D_i$ that governs the probability of the correct answer, it is possible to add the same arbitrary constant to both B_n and D_i without changing the probability. The origin of the logit scale of the latent variable is therefore arbitrary. Commonly in analyses, the origin (zero logits) is set at the average difficulty of the test items. The size of the logit depends on the way the variable is operationalized by the items, and is not the same in each analysis. This is often described as analogous to the difference of temperature scales, e.g., the Fahrenheit and Celsius scale. The comparisons of results obtained by different Rasch analyses will therefore first require the equating of their logit scales, just as in the case in which one wishes to utilize both Celsius and Fahrenheit temperature data for a study.

It is important to mention that the measures B_n and D_i are not counts, and cannot be observed directly, but only inferred and estimated from the response pattern of examinees. If the difference between B_n and D_i is zero, the probability of success is exactly 50%. Different values of probabilities of success for different values of $B_n - D_i$ are displayed in Table I. Figure 2 shows a typical curve that illustrates the nonlinearity of raw scores which are plotted against the corresponding Rasch person ability measures (the raw scores to measures curve, also called the logistic ogive). The nonlinearity is most strongly expressed at the high and low ends of the raw score scale [1]. This means that the low and high performers on tests are particularly impacted when raw score totals are used as proxies for "measures."

The Rasch model enables us to construct measures of students' abilities and item difficulties from their response pattern [in the case of a dichotomous test the patterns of 0 (incorrect) and 1 (correct) answers to each item by a respondent], with the use of different software packages for Rasch analysis (commercial software such as e.g., WINSTEPS, QUEST/CONQUEST, or RUMM, or free software such as BIGSTEPS, MINISTEPS, or R). It is important to note that Rasch software is typically very user friendly, simple to use, and requires neither the learning of a new programming language nor extensive coding.

TABLE I. Probability of a correct answer to dichotomous questions for different values of $B_n - D_i$.

Difference of person ability and item difficulty, $B_n - D_i$ (in logit)	Probability of a correct answer of person n to item i
5	0.99
4	0.98
3	0.95
2	0.88
1	0.73
0	0.50
−1	0.27
−2	0.12
−3	0.05
−4	0.02
−5	0.01

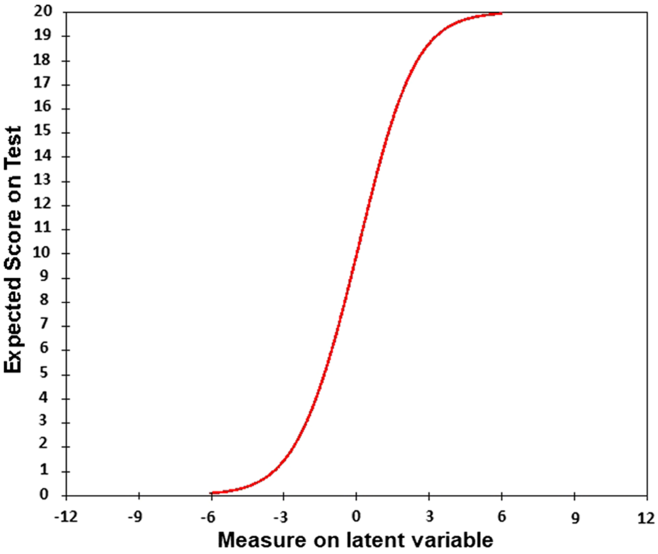

FIG. 2. A typical Rasch raw scores to measures curve for an example of a test of 20 dichotomous item, each scored 1 point, plotted from Winsteps [26]. Rasch measures are expressed in logit.

According to Wright [31], the measures must be inferences by stochastic approximation, expressed in abstract units, linear, of unidimensional quantities, and impervious to extraneous factors. The measures obtained through Rasch modeling fulfill all those criteria.

Rating scales (e.g., surveys using a Likert scale) are often used in educational research. Andrich perceived that categories of a rating scale could be thought of as a series of dichotomies [32]. The dichotomous Rasch model can be expressed as

$$\ln(P_{ni1}/P_{ni0}) = B_n - D_i, \tag{5}$$

where P_{ni1} and P_{ni0} stand, respectively, for probabilities of success and failure of person n on a dichotomous item i. In a rating scale model each item will have several rating scale categories. The probability of a person n endorsing category j over previous category $(j - 1)$, or being observed in category j of item i, can be expressed in a Rasch-Andrich rating scale model [35] as

$$\ln(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_j, \tag{6}$$

where F_j is the Rasch-Andrich threshold (step calibration), or the point on the latent variable where the probability of person n being observed in category j of item i equals the probability of the same person being observed in category $(j-1)$. F_j is estimated from the category frequency, and the difficulty of the item is now located at the point where the highest and the lowest categories are equally probable.

In a rating scale model each item has the same thresholds for the rating scale categories. This means that one can compute the rating scale steps from category to category, for example from SA to A, from A to D, from D to SD, with no assumption made as to the distance of a step. But, importantly, with this model, the rating scale structure which is determined is treated as if the structure is the same for each item.

A version of the rating scale model, where the rating scale is specific to each item (thresholds are different for different items), is called the Rasch-Masters partial credit model [36]. This model is described mathematically as

$$\ln(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_{ij}. \tag{7}$$

In this model F_{ij} stands for the threshold between categories j and $j-1$ on item i. In contrast to the rating scale model, in the partial credit model the items of the same raw score can have different values of Rasch difficulties if the pattern of category usage on those items is different. The partial credit model is suitable for analysis of items which do not share the same category structure. It is possible to evaluate, for example, a set of rating scale data (e.g., using a Likert scale SA, A, D, SD), but not assert that rating scale steps have the same structure for each item.

A further extension of the Rasch model is the Rasch-Linacre many faceted model [37], which was developed for situations where the performance of a person on a specific task is judged by several raters:

$$\ln(P_{nijk}/P_{ni(j-1)k}) = B_n - D_{gi} - C_k - F_{gj}. \tag{8}$$

C_k represents the severity (or leniency) of rater (judge) k, who awards the ratings j to person n on item i in group g. Unlike the previous three models, this model has been rarely used in physics education research, but we feel that in cases when a variety of judges are used to evaluate student performance, using items, this model should be considered for use.

IV. TEST CONSTRUCTION AND EVALUATION USING THE RASCH MODEL

One of the most important uses of the Rasch model in PER is to help guide test and survey construction and evaluate their functioning. In this section we outline the general process and the most important aspects of test construction and evaluation with the Rasch model in order to facilitate this process for PER researchers. To provide readers with guidance we find the suggestions of Liu [39] to be helpful. Construction of a measurement instrument with the Rasch model is according to Liu [39] "a systematic process in which items are purposefully constructed according to a theory and empirically tested through Rasch models in order to produce a set of items that define a linear measurement scale." Liu includes the following steps (also summarized in Table II):

1. Define the construct that can be characterized by a linear attribute.
2. Identify the behaviors corresponding to different levels of the defined construct.
3. Define the outcome space of behaviors (item pool).
4. Field test with a representative sample of the target population.
5. Conduct Rasch modeling.
6. Review item fit statistics and revise items if necessary.
7. Review the Wright map and add or delete items if necessary.
8. Repeat (4) to (7) until a set of items fit the Rasch model and define a scale.
9. Establish validity and reliability claims for the measurement instrument.
10. Develop documentation for the measurement instrument.

The starting point, from the Rasch measurement perspective, for the development of an instrument is the use of theory. The focus on theory ensures that the constructed instrument will have high construct validity [39]. The construct which is to be measured must be defined in step 1 in terms of a unidimensional progression in student knowledge or attitudes, from a lower to a higher level. This progression needs to be based on the respective theory of the investigated topic. In step 2, the types of items which can be used to obtain adequate information from examinees about the topic are specified. Of importance is that the items are constructed in a way to ensure that different levels of the construct require different levels of cognitive reasoning. This means that to define items along a construct, for example for a test, one needs to include items of varied difficulty. In step 3, an initial item pool and item scoring keys or rubrics will be developed. In this step some qualitative testing of items can be of great help (e.g., conducting interviews with selected representative participants, using "think aloud" protocols). In step 4, a draft of the instrument is tested with a representative sample of the target population with an adequate spread of abilities (for a test there should be, when possible, lower ability, middle ability, and high ability respondents). The obtained data should then be analyzed with Rasch software in step 5, and the fit of data with the model is examined in step 6. Fit can be understood as the calculation and comparison of the differences between the theoretical and experimental values (residuals) for both persons (respondents) and items. There

TABLE II. Short description of the steps in the construction of a measurement instrument.

Step in the construction of a measurement instrument	Description of the step
1. Definition of the construct	The construct which is to be measured has to be characterized by a linear attribute and based on the respective theory of the investigated topic. The defined construct should be unidimensional and show progression of student knowledge or attitudes.
2. Identification of different levels of the defined construct	Different levels of the construct should correspond to different levels of cognitive reasoning or attitude. The items corresponding to these levels are developed and tested in the subsequent steps.
3. Delineation of the construct into items	The initial set of items is developed based on the previous research or a qualitative study (e.g., interviews).
4. Field testing with a representative sample of the target population	The initial version of the instrument is tested on a representative sample with respondents of different abilities or attitudes.
5. Rasch modeling	The Rasch analysis is conducted by using Rasch software.
6. Evaluation of the item fit statistics and revision of items	The fit of data to the model is evaluated by comparison of the theoretical probabilities for the success of each person on each item with the observed values. The fit statistics and the point-measure correlations are inspected. Poorly fitted items should be reviewed or removed from the instrument.
7. Evaluation of the Wright map and adding/deleting of items	The Wright map helps with evaluating the targeting of the test to the sample and the structure of the test. If large gaps between items are detected, new items of appropriate difficulty should be added. If too many test items of the same difficulty are found, some of them can be removed.
8. Iteration of steps 4–7 until a set of items fit the Rasch model and define a scale	Steps 4–7 are repeated until a set of items obtains the characteristics of a measurement instrument. The invariance properties of the constructed instrument should also be evaluated. If an item behaves very differently for the two subgroups of examinees (e.g., female and male examinees), it should be revised or removed from the test.
9. Validation of the measurement instrument	The theoretical construct validity should be present in the initial steps of defining the construct, whereas empirical construct validity is evaluated through Rasch analysis (fit statistics, item correlations, and instrument unidimensionality). The reproducibility of measures is estimated by calculating person and item reliability.
10. Developing documentation for the measurement instrument	The instrument documentation includes the description of the instrument development and pilot testing and guidelines for users of the instrument. A conversion scale of Rasch measures to raw scores can also be included.

are a number of fit statistics commonly considered. Outfit MNSQ (mean squares) statistics is the arithmetic mean of simple squared residuals, whereas infit MNSQ is a weighted mean of squared residuals [27]. Outfit is more sensitive to outliers, and infit to respondents' responses to items whose difficulties are close to respondents' abilities. Outfit and infit can be based alternatively on ZSTD values, where ZSTD is a normalized Z score of the residual. It is generally accepted that the items with infit and outfit MNSQ values between 0.7 and 1.3 and infit and outfit ZSTD values between -2 and 2 have good model fit [39], and all items with infit and outfit MNSQ values between 0.5 and 1.5 can be regarded as productive for measurement [26]. Item functioning can be evaluated with the help of fit statistics and point-measure correlations which indicate how a specific item contributes to the whole person or item measure. A bubble chart [27] is one way of visualizing the overall functioning of test and the fit of its items. It is a graph of item difficulty vs item outfit MNSQ-infit MNSQ-outfit ZSTD-infit ZSTD. Each item is represented by a circle, whose size is proportional to its standard error of the calibration. No matter the means of plotting the interplay of fit and items as well as fit and persons, ideally items should be as close as possible to a modeled value of 1 for outfit MNSQ and infit MNSQ, or 0 for outfit ZSTD and infit ZSTD (Fig. 3).

Poor fit of some items may indicate problems with those items' structure (e.g., partial credit, multiple choice, etc.), wording, scoring, or content. The fit of persons should also be examined, because sometimes person misfit can be the cause of the item misfit. Analysis of the answer patterns of misfiting persons can reveal problems with person behavior, such as guessing, and these persons should be eliminated from an analysis. It is important to emphasize that

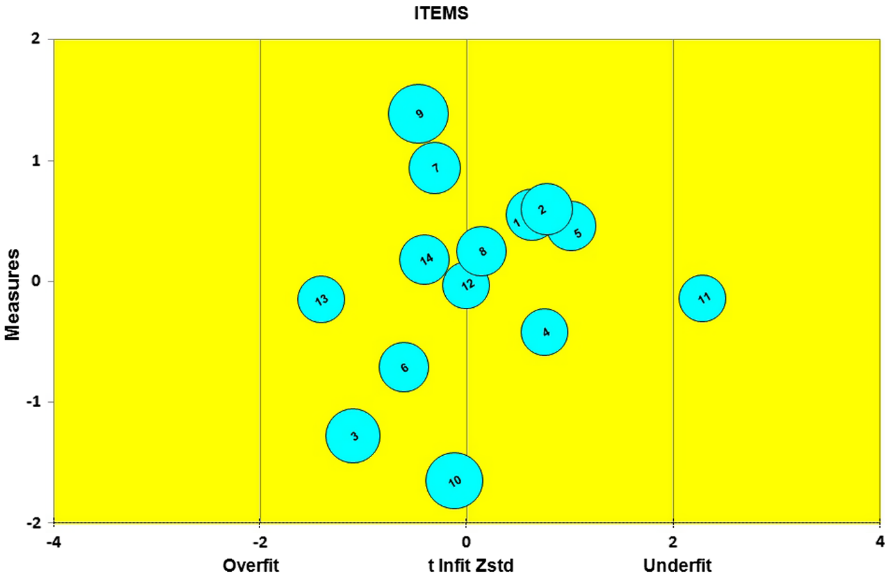

FIG. 3. An example of a bubble chart for 14 test items, plotted from Winsteps [26]. Each bubble represents an item, whose size is proportional to the standard error of item difficulty calibration. Well-fitting items are close to the central vertical line. Item 11 is outside the acceptable range of misfit (infit ZSTD greater than 2).

the specific logic of Rasch modeling is not to fit the model to the data, but to construct the instrument which is in good agreement with the theory and basic requirements of objective measurement. That leads to the necessity of discarding some of the data when there is significant misfit present, for those cases when, for example, items do not seem to define the trait, and persons seem not to be concentrating as they respond to items.

In step 7 the structure of the test is examined with the use of a Wright map. A Wright map (also known as item-person map) presents both item difficulties and person abilities arranged along the same logit scale (Fig. 1). Through the use of a Wright map one can visualize the targeting of the test to the sample, as well as the targeting of individual items to persons. A well-constructed instrument should match the width of the target population ability distribution with the width of the distribution of test items. The presence of large gaps between the item difficulties in the test means that persons within those gaps cannot be measured precisely enough because of the lack of close items near their ability level. This is akin to a meter stick that is missing some marks and attempting to measure an object whose length falls within the gaps between marks. A person of ability B_n is best measured with the items of difficulties within ± 1 logit from B_n. To fix such gaps in a test, new items of appropriate difficulty should be added to the test. It is also important to note that too many test items of the same difficulty are not necessary (especially at the high and low ends of the test), and if such situation is detected, some of them can be removed to shorten the test. The unidimensionality, an essential requirement of the Rasch model, should also be inspected when the functioning of the test is evaluated. One way to evaluate dimensionality is through the analysis of point-measure correlations and item fit (misfit of items can sometimes be a sign of the presence of another dimension in the test, different from the one that was intended for measurement). Another way is to examine the dimensionality of Rasch residuals through the principal component analysis of residuals. If strong correlations are found among residuals, which should in principle be uncorrelated if the test is unidimensional, it is possible that there are one or more additional dimensions in the test. Rasch software (e.g., Winsteps [26]) performs this analysis and identifies items which might belong to a different dimension.

In step 8, when problematic items are identified, they are reviewed, and consequently corrected or removed from the test. Some new items may also be added, and a new cycle of steps 4–7 begins, until a satisfactory set of items is established, and the instrument obtains the properties which suggest a test functioning as is required for measurement instruments. At this point it is time to examine the invariance properties of both the person and item measures. This is usually done through scatter plots of item measures of different subgroups of examinees (e.g., plots in which item measures obtained from two subgroups are plotted one against the other to see how they compare), or of person measures obtained from two tests which measure the same construct (e.g., plots in which measures of ability of the same persons tested by two tests are plotted against each other). If there is good fit of the data collected with the instrument and the model, it should be expected to find essentially the same measures in those comparisons (e.g., in a plot of item measures obtained from analysis of female examinees vs those obtained from male examinees for the same items, one should find that all obtained points are close to the identity line), within the limits of their standard errors. A point in such a plot that departs very much from the identity line would suggest that the item represented by that point behaves differently for the two groups of examinees [this is called differential item functioning (DIF)], and the item in question should be further examined, revised, or possibly removed from the test.

In step 9, the validity and reliability of the constructed instrument are evaluated. Often validity has been considered to be of various types (e.g., criterion related, content, and construct validity). The current view is that validity is unitary [39], and might simply be called construct validity. It relies on two foundations: theoretical and empirical. Since the construction of a test using Rasch theory is based on what it means to measure, and begins from the theoretical considerations of the investigated latent trait, and the hierarchical organization of the trait (e.g., what skills should be exhibited at the lower end of the trait and what skills exhibited at the higher, more advanced, end of the trait), the theoretical validity should already be present through the construction process requiring a definition of the variable. The theoretical construct is operationalized through the choice of items, which define the construct. The face validity of items needs to be investigated by experts in the field. How well the chosen items have empirically succeeded in defining a sufficiently unidimensional and consistent construct (construct validity) can be investigated through Rasch analysis, namely the earlier described analysis of item fit, item correlations, and test unidimensionality. The predicted order of item difficulty can be compared to the pattern observed on the Wright map.

Test reliability can be viewed as the reproducibility of measures. A current view on the issue of reliability concerns a systematic analysis of various sources of errors associated with items, but also with other facets of measurement, such as raters or testing setting [39]. Rasch analysis reports person reliability, which is analogous to the test reliability of classical test theory (Cronbach alpha), and item reliability, which has no analog in classical test theory. A reliability of 0.5 is considered to be the minimum meaningful reliability, whereas 0.8 is the lowest person reliability for any decision making involving students' abilities [26]. It is important to emphasize that high reliability does not necessarily imply good quality of the test. High reliability simply means that the test scores are reproducible, but the meaning of the scores is a different matter. The quality of the test depends also on the quality of its items and the degree to which they define a meaningful construct.

One technique utilized in Rasch measurement concerns the computation of the person separation index. The person separation index G refers to the precision of measurement and indicates how well one can differentiate examinees' abilities with a test. The number of ability strata which can be resolved is provided by the formula $(4G + 1)/3$ with the assumption that different ability levels are 3 standard errors apart [27]. For example, a person separation index of 2 implies three distinct ability levels which the test can differentiate whereas a person separation index of 3 implies four distinct ability levels.

In step 10 the documentation which will make the use of the instrument easier is developed. It often includes the description of the process of instrument construction and pilot testing and guidelines for users of the instrument. A conversion scale of Rasch measures to raw scores is often included.

Until recently, Rasch analysis was mostly used to evaluate the functioning of the previously developed PER diagnostic tests [4,5,16,21,25]. Planinic [4] evaluated the functioning of the Conceptual survey of electricity and magnetism and Ding [16] reevaluated the widely used Brief Electricity and Magnetism Assessment with Rasch analysis. Similarly, the most famous and widely used PER instrument, the Force Concept Inventory (FCI), was analyzed using the Rasch model to examine its structure and functioning on two different samples of students (non-Newtonian and predominantly Newtonian), detect possible problems, and suggest further improvements [5]. Taasoobshirazi *et al.* [21] performed the Rasch analysis of Physics Metacognition Inventory to assess its construct validity. The Test of Understanding of Vectors was also analyzed using the Rasch model [25].

More recently, PER researchers started to follow the procedure described above for test construction and use the Rasch model in the process of test development. Based on the initial learning progression, Neumann *et al.* [11] constructed a measurement instrument, the Energy Concept Assessment. Testa *et al.* [19] developed and validated a two-tier multiple-choice questionnaire about the change of seasons, solar and lunar eclipses, and moon phases. Hofer *et al.* [24] constructed and evaluated the test of basic Mechanics Conceptual Understanding that is adapted for secondary school students. Planinic *et al.* [14] developed and evaluated a test on graphs in different contexts. Ene and Ackerson [18] developed the Physics of Semiconductors Concept Inventory using Rasch analysis on a relatively small sample of students enrolled in the course Introduction to Physics of Semiconductors. Aslanides and Savage [12] developed and calibrated similarly their Relativity Concept Inventory.

The Rasch model is especially suitable for linking tests and tracking students' progression, which are both very important for PER. "Learning progressions" are most often built using Rasch model calculations; see e.g., [11,13,17,19,20,40,41]. Fulmer *et al.* [17] used the Rasch model and the Force Concept Inventory to explore a proposed force and motion learning progression on a sample of high school and university students. In a recent large-scale study on learning progression for energy ideas from upper elementary through high school, the authors also used Rasch analysis [40]. Paik *et al.* [41] developed a four-level learning progression and assessment for the concept of buoyancy.

Because Rasch item difficulties are sample independent (within their statistical limits), the Rasch model can be used to build item banks, which can make test construction much easier [42].

V. COMMONLY ENCOUNTERED MISUNDERSTANDINGS WHEN CONDUCTING A RASCH ANALYSIS

When Rasch analysis is attempted in PER, there are often misunderstandings encountered, some of which are quite common. It is our goal in this section of the paper to briefly address some of these misunderstandings and help analysts (especially novices) to avoid such pitfalls.

A. Rasch analysis is not only number crunching

It is important to understand that Rasch analysis is not just number crunching, but that it is about the conceptualization of a variable. When utilizing Rasch analysis, we think about what we want to measure, and we make predictions of what it means to measure. By doing so we are guided as to what questions we want to ask, to help us locate a respondent on a trait. Thinking about what we want to measure also helps us think about what it means to go up and down the measurement scale (what skills, for example, does one student have with a higher measure as opposed to a student with a lower measure).

In Rasch analysis we evaluate data quality, and if the data is of low quality, we might remove the data. For example, through the analysis of each person's response pattern we can identify respondents who are behaving in a manner that is quite unexpected, often in a manner that suggests they might be wildly guessing on an exam. On the other hand, there may also be items which perform poorly and do not contribute to the intended test construct. With Rasch we do not view data as sacred; rather we view data which does not match the requirements of fundamental measurement as data we might not use for item calibration (determining where items fall on the trait) and the computation of person measures (determining where persons fall on the trait). Rasch analysis provides the means for a strong quality control of the instrument used and the data obtained. No statistical model can save bad data, and bad data is not even worth saving. Therefore, instead of looking for how to change the model to fit the data, with Rasch we try to construct quality instruments and perform with these instruments high quality measurements. In that process, we sometimes have to discard some data which do not conform to the requirements of the measurement model.

The Rasch model is viewed as a definition of measurement and what it means to measure, so the model is not altered to fit the data. Rasch measurement theorists prefer to view the assumptions of the model as requirements for objective measurement [33], which must not be altered if one wishes to perform such measurements. This is the philosophical point that distinguishes Rasch fundamentally from other modeling approaches, e.g., Item Response Theory (IRT), although the math of some of the IRT models may look similar to that of the Rasch model.

B. Rasch analysis does not necessarily require large data sets

A possible misunderstanding of those first using Rasch is that one needs large datasets to conduct a Rasch analysis. This is most likely the result of many large-scale assessments using Rasch. However, Rasch analysis can be conducted with small datasets; see e.g., Refs. [12,18]. An example is the Relativity Concept Inventory (RCI) development: the pretest sample was 70 students, and the posttest sample was 63 students, of which 53 matched both measures [12]. Linacre suggests [43] that for obtaining stability of item calibrations within one logit, for dichotomous items, already 50 well-targeted examinees can be enough. He also provides guidelines for optimal sample sizes for different testing purposes and different intended precisions of item calibrations [43]. One should be aware that for rating scale analysis of a survey with N items, in which each item stems from e.g., five categories, the Likert scale is treated as a separate item ($5N$ items), proportionally larger samples will be needed than for the analysis of tests with N dichotomous items, to obtain the same density of data in both cases [27].

Another misunderstanding may be that there is a set number of test items which are needed for an instrument. In fact, the number of items needed depends upon how well items are distributed along the trait, what one wishes to measure, with what precision, where respondents are along the trait, and what sort of decision one will be making with the results.

C. Logit may have different sizes

The measurement scale of Rasch is in logit. These are the units used to express both person ability (in the case of a test) and item difficulty. A common error is assuming that a logit person measure on one scale means the same as a logit person measure on another scale. For example, a 25-item mechanics physics test is developed, and person and item measures are computed. Now let us imagine a different 30-item mechanics test which was developed. Also, item and person measures are computed. The person measures are all expressed in logits, but the meaning of the logit numbers is based upon the scale defined by the items and the location of the origin (the value of "zero logits" is commonly set at the average item difficulty value). To compare the measures from two different logit scales, the scales first need to be equated, either through anchoring of common items, anchoring of common persons or by other methods (for more detail see, for example, Ref. [26]). It is even possible, in certain situations, to equate the scales of two tests which do not share either items or persons, by the procedure of virtual test equating [10,26].

D. Measurement errors should not be ignored

Rasch analysis, unlike other methods, provides, along with measures, estimates of person measure error and estimates of item measure error. Sometimes researchers compute person and item raw scores in the form of percentages or total number of points, and treat those scores as being infinitely precise measures of student ability or item difficulty. No measurement error is reported or considered in the analysis. On the other hand, some researchers may even compute Rasch person and item measures, and still ignore the standard errors of those measures in further analysis. Taking into account the measure together with its error is an important part of Rasch measurement, just as it is important in physics measurements. It is possible to compute person measures for a very short test, for example, a 4-item test. The values for each person might look precise (e.g., 2.03 logit), but one must look at the measurement error term, which will be quite large in that case, at least of the order of 1 logit. That will mean that the mentioned person calibration is not 2.03 logit, but somewhere between 1 and 3 logits approximately. Standard errors determine the precision of the estimated measures of item difficulty and person ability, and they have to be taken into account when making any further conclusions or when conducting statistical analyses with those measures. The size of the standard error of item difficulties is mostly influenced by the sample size N_s ($\mathrm{SE} \sim 1/\sqrt{N_\mathrm{s}}$), whereas the size of the standard error of person calibrations is mostly influenced by the number of the items in the test N_i ($\mathrm{SE} \sim 1/\sqrt{N_\mathrm{i}}$). However, poor targeting can also increase standard errors. That is the reason that items and persons at the high and low end of the scale will tend to have higher standard errors than items and persons in the middle of the scale.

E. Rasch is not just for instrument development but also for computing person measures

A common error is to not remember that Rasch is used not only to develop and improve instruments but also for the computation of person measures (these are the measures of respondents on a linear scale). Sometimes beginners use Rasch to develop their instruments, but they forget that they must also compute and evaluate person measures using the logit scale. They may also not know how to evaluate those person measures, and the measures should be, at least initially, evaluated with parametric statistics.

The Rasch person measures may be used for statistical tests. For example, if one wishes to examine the performance of boys and girls with a physics test one must compute the girl and boy person measures on the logit scale, and then make a statistical comparison, for example a t-test, of the boys and the girls, using those logit person measures. Many other statistical tests and analyses can also be conducted using Rasch measures.

F. Construct validity should be evaluated before reporting measures

Novice analysts are sometimes focused too much on obtaining person and item measures through Rasch analysis of test data, and forget to first evaluate the validity of the construct underlying the test. The Wright map (e.g., for a multiple-choice test) which provides a plot with persons on one side of the map and items on the other side of the map (Fig. 1), and the bubble chart of item difficulties vs their outfit or infit values (Fig. 3) can help in construct evaluation. An error of some novice researchers may be to overlook the need to evaluate the construct validity of their instrument. The Wright map and the bubble chart can facilitate a review of the ordering and spacing of test items as a function of difficulty. The bubble chart can allow one to evaluate the level of item misfit with the model. If the ordering of items matches theory, that is evidence supporting the construct validity of an instrument. If the ordering of items does not match theory, or there is too much misfit in the data, that is evidence of the instrument having suspect construct validity. Unless there is reasonable construct validity of the test and reasonable fit of data with the Rasch model, there is not much sense in reporting the measures obtained through Rasch analysis.

However, some researchers may make the opposite error and place too much confidence in the good numbers and charts obtained by Rasch analysis, forgetting that these are not enough by themselves to conclude that the instrument is valid. The face validity and overall quality of test items should also be closely inspected (in addition to performing numerical analysis) before such a conclusion can be reached.

The Rasch model allows us to see if items conform to our idea of what it means to measure one trait. If an item does not conform, it might show misfit, and if we run the model separately for two subgroups, e.g., men and women in the sample (and then cross plot the obtained separate item difficulties) we might see possible differential item functioning, in that for some items the item difficulty for men is very different than for women (if a consistent difference is found on all items, e.g., all items are more difficult for men than for women, that is not a sign of DIF, but of different abilities of the two groups). When we do see DIF, we need to analyze the items showing DIF and then decide about the steps we need to take, taking into account the size of DIF. Small size DIF can sometimes be ignored, but if an item is strongly biased against one or the other group, then that item should be excluded from the test. It is important to stress that we can come close to objective measurement only with the use of quality and unbiased instruments. It is therefore always the first step in Rasch analysis to inspect the functioning and structure of the instrument. Only if the instrument is found to be well constructed, the measures it produces can be taken as meaningful and reliable.

G. Problems related to rating scale analysis

Sometimes researchers do not understand how an attitude can be marked on a trait. Researchers, generally, do not have a difficult time understanding how dichotomous items can be used to mark a trait. But it may be difficult for some researchers to understand how an attitude trait can be marked by survey items. For example, a survey may provide a number of attitudinal items, and some of those items will be harder to agree with than the other items. In the context of the rating scale analysis, items that are harder to agree with are the more "difficult" items, and this difference in the agreeability of items can define a scale for the investigated trait. Bond and Fox [27] give an example of two statements in a survey on computer anxiety, which are clearly of different difficulty (agreeability): 1. I am so afraid of computers I avoid using them; 2. I am afraid that I will make mistakes when I use my computer. It takes a higher level of computer anxiety to strongly agree with the first statement than with the second, but in a classical survey analysis both answers would carry the same number of points. Similarly as for dichotomous items, the Rasch model will determine the difficulty of each item stem, based on the answer pattern of the examinees. From this pattern of responses, a rating scale structure for the survey will be determined. The distances between different Likert scale categories most often do not turn out to be the same, whereas in a classical survey analysis these distances will be automatically assumed to be the same. Two models can be used for analyzing rating scales, the rating scale model or the partial credit model. Depending on the details of the study or the type of the data, one or the other may show to be a better choice [8]. The rating scale model produces the same rating scale structure for all items, whereas the partial credit model allows different structure for different items.

Another important feature of rating scale analysis is the analysis of the functioning of Likert scale categories in a survey, which novice analysts sometimes forget to investigate. The Rasch analysis may show that some categories of the rating scale do not function optimally and that it is better to use a smaller number of categories (collapsing some categories) [32]. Analyzing category probability curves that can be produced by some Rasch software can help determine when collapsing of categories is advisable [32].

It is important to mention that the Rasch model can also be used to evaluate partial credit tests, as well as tests which have several types of items (e.g., 20 dichotomous items, 2 items worth between 0 and 3 points and one item worth between 0 and 5 points).

VI. PROBLEMS WITH PRETESTING, POSTTESTING, AND NORMALIZED GAIN

It is a common practice in PER to evaluate students' "learning gain," or to evaluate the effectiveness of instruction, through pretesting and posttesting students with the same diagnostic instrument. We have already touched on this issue in one of our previous studies related to the Force Concept Inventory [5] and suggested that several problems may exist with pretesting and posttesting with the same instrument. In that study we have shown that the FCI had poor targeting and width on a typical pretest (pre-Newtonian) and posttest (predominantly Newtonian) student populations, suggesting that the FCI is not very well suited for measuring either one of these populations [5]. This is not surprising, since it is difficult, if not impossible, to construct a single test of reasonable length that will precisely measure both populations' ability spans. We have also shown in previous work that the construct defined by the FCI may have changed from pretest to posttest (manifested by a different order of item difficulties), making it problematic to compare the obtained scores [5]. A significant change in item ordering means, in effect, that the measurement instrument has changed. Some other studies have suggested that some of the FCI items may be gender biased; see, e.g., Ref. [44]. This may also be inspected with Rasch analysis by analyzing items for possible differential item functioning, as described in the previous section.

The typical way of evaluating student progress with diagnostic instruments usually includes computing of the Hake's normalized gain g [45] from posttest and pretest individual or class means in percentages as

$$g = (\text{posttest} - \text{pretest})/(100\text{-pretest}). \qquad (9)$$

Hake introduced g as a measure of student gain on conceptual inventories because it seemed not to be correlated with the pretest class means and therefore seemed to be a suitable measure for comparison of diverse populations of students [45]. Hake's normalized gain soon became a widely used measure of student success and/or instruction effectiveness. However, g was also criticized by some researchers. Coletta and Philipps found that g was correlated with class pretest means [46]. Bao [47], as well as Marx and Cummings [48], pointed to the pretest bias of Hake's g and Bao showed that different ways of calculating g (with class pretest and posttest means or as a mean of individual student gains) led to different results [47]. Willoughby and Metz [49] found in their study that different ways of defining gain influenced whether performance of male and female students on conceptual tests in astronomy and biology appeared different or not, with Hake's g being the only measure that suggested gender inequities in contrast to several other measures which indicated statistically equal performance of both genders on the conceptual tests [49].

Although the normalized gain has played a major role in the development of PER, and is widely used in the PER community, we have to consider that it is computed from raw scores, which are nonlinear. The normalized gain can therefore also suffer from nonlinearity issues. This was suggested by Planinic, Ivanjek, and Susac [5], and further analyzed by Wallace and Bailey [6] and Ene and Ackerson [18]. Wallace and Bailey state that Hake's normalized gain, being constructed from ordinal data, may be at most an ordinal measure of learning gain [6]. When using Rasch analysis or IRT, it was suggested by Embretson and Reise [50] to compute learning gain as the difference in students' abilities obtained from the analyses of their pretest and posttest data with equated logit scales. Wallace and Bailey have analyzed the Star Properties Concept Inventory results with the Rasch model, and computed Rasch gains in the way suggested by Embretson and Reise. When they compared the obtained Rasch student gains with the normalized gains computed from raw scores, they did not find a one-to-one correspondence between them. Instead, for any value of Hake's normalized gain multiple Rasch gains were found [6]. Since normalized gain favors students with high pretest results (for the same absolute gain, a student with a higher pretest result will have a larger normalized gain), it was shown to be even possible that a student with a smaller increase in Rasch ability achieves a larger normalized gain than another student with a much larger increase in ability [6]. Ene and Ackerson also suggest the use of Rasch gain instead of the Hake's gain [18]. They warn that the same Rasch gain could respond to different Hake's gains. They give a simulated example in which they show that the same Rasch class gain of +0.8 logit may correspond to a g of 7.5% for a class with the pretest value of 20%, and a g of 35% for a class with the pretest value of 43% [18]. Lasry, Guillemette, and Mazur have also found (when analyzing more than 13 000 FCI student answers) that Hake's gain favors students with higher pretest results [51]. Pentecost and Barbera problematized the issue of normalized gain and its nonlinearity and suggested the use of Rasch gain instead [15]. It is obvious from these analyses that the normalized gain should be used with extreme caution. In addition to Rasch gain, other alternative measures have also been proposed. In a recent study Nissen *et al.* suggested the use of Cohen's d instead of Hake's g [52]. Marx and Cummings suggested a new measure c, called normalized change, to replace g [48]. However, since both Cohen's d and normalized change c are computed from raw scores, they may also suffer from the issues of nonlinearity, leaving Rasch gain as the only solution for this problem for now.

To improve measurement in PER, we also suggest that instead of using the same diagnostic instrument for pretest and posttest, two instruments are constructed for the same topic sharing several common items, which would serve to link the two tests and enable comparisons of the obtained student abilities on the same logit scale. It is an important advantage of Rasch modeling over standard approaches that we do not have to use exactly the same instrument to be able to compare student abilities or monitor learning progress. With two tests instead of one, better targeting for both pretest and posttest populations could be obtained, the width of the tests could be better adjusted, and there would be less risk of students remembering items from the first administration or being bored by taking the same test twice. Better targeting would ensure smaller calibration errors and therefore more precise measurement, without floor and ceiling effects. Normalized gain could simply be replaced by the difference of students' abilities on posttest and pretest (Rasch gain), and class means, as well as other statistical measures, could be computed with linear measures, which are on an interval scale, avoiding completely the possible nonlinearity issues.

VII. CONCLUSIONS

The development of many quantitative assessment instruments in PER starting from the 1990s has greatly influenced the way that physicists and physics teachers viewed teaching and learning of physics at both the high school and university levels. This happened because the results obtained with those instruments were often repeatable, and showed that they had some general meaning, that was not only limited to a specific group of students that was tested. The introduction of measurement in physics education research was of great importance and greatly impacted the whole paradigm of educational research in physics. However, the standards of measurement in PER are still not unified, and measurements are of very differing quality. One of the main concerns for PER researchers should be how to bring the standards of educational measurement closer to the standards of objective measurement, like those found in physics. We believe that Rasch modeling is an important step in that direction. It allows us to depart from counts and observations and construct abstract linear measures whose meaning can transcend the testing occasion. It also enables the careful construction of diagnostic assessment instruments and their quality control. It can lead to construction of item banks and enable comparisons and longitudinal studies of development. Common criticism toward the Rasch model is that its assumptions are too rigid, especially the assumption of unidimensionality. However, we hope that we have managed to show that this assumption is necessary for any measurement and that in reality unidimensionality is always an approximation, which should be checked empirically. The use of the Rasch model in PER is constantly increasing, but its values and possibilities are still not widely known and appreciated. In this article we have tried to show some of the benefits of using the Rasch model in research and to point to some common difficulties and misunderstanding of it. We hope that it may contribute to improving measurement quality in PER.

[1] W. J. Boone and K. Scantlebury, The role of Rasch analysis when conducting science education research utilizing multiple-choice tests, Sci. Educ. **90**, 253 (2006).

[2] M. Planinic, W. J. Boone, R. Krsnik, and M. Beilfuss, Exploring Alternative Conceptions from Newtonian Dynamics and Simple DC Circuits: Links between Item Difficulty and Item Confidence, J. Res. Sci. Teach. **43**, 150 (2006).

[3] A. Kauertz and H. E. Fischer, Assesing students' level of knowledge and analysing the reasons for learning difficulties in physics by rasch analysis, in *Applications of Rasch Measurement in Science Education*, edited by X. Liu and W. J. Boone (JAM Press, Maple Grove, MN, 2006), pp. 212–246.

[4] M. Planinic, The Rasch model-based analysis of the Conceptual Survey of Electricity and Magnetism, in *Proceedings of GIREP Conference 2006: Modeling in Physics, and Physics Education*, edited by E. van den Berg, A. L. Ellermeijer, and O. Slooten (University of Amsterdam, Amsterdam, 2006), pp. 923–927.

[5] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

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

[7] J. D. Plummer and J. Krajcik, Building a learning progression for celestial motion: Elementary levels from an Earth-based perspective, J. Res. Sci. Teach. **47**, 768 (2010).

[8] I. Neumann, K. Neumann, and R. Nehm, Evaluating instrument quality in science education: Rasch-based analysis of a nature of science test, Int. J. Sci. Educ. **33**, 1373 (2011).

[9] W. J. Boone, J. S. Townsend, and J. Staver, Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data, Sci. Educ. **95**, 258 (2011).

[10] V. Mesic and H. Muratovic, Identifying predictors of physics item difficulty: A linear regression approach, Phys. Rev. ST Phys. Educ. Res. **7**, 010110 (2011).

[11] K. Neumann, T. Viering, W. J. Boone, and H. E. Fischer, Towards a learning progression of energy, J. Res. Sci. Teach. **50**, 162 (2013).

[12] J. S. Aslanides and C. M. Savage, Relativity concept inventory: Development, analysis, and results, Phys. Rev. ST Phys. Educ. Res. **9**, 010118 (2013).

[13] J. C. Hadenfeldt, S. Bernholt, X. F. Liu, K. Neumann, and I. Parchmann, Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter, J. Chem. Educ. **90**, 1602 (2013).

[14] M. Planinic, L. Ivanjek, A. Susac, and Z. Milin–Sipus, Comparison of university students' understanding of graphs in different contexts, Phys. Rev. ST Phys. Educ. Res. **9**, 020103 (2013).

[15] T. C. Pentecost and J. Barbera, Measuring learning gains in chemical education: A comparison of two methods, J. Chem. Educ. **90**, 839 (2013).

[16] 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).

[17] G. W. Fulmer, L. L. Liang, and X. F. Liu, Applying a force and motion learning progression over an extended time span using the Force Concept Inventory, Int. J. Sci. Educ. **36**, 2918 (2014).

[18] E. Ene and B. J. Ackerson, Assessing learning in small sized physics courses, Phys. Rev. Phys. Educ. Res. **14**, 010102 (2018).

[19] I. Testa, S. Galano, S Leccia, and E. Puddu, Development and validation of a learning progression for change of seasons, solar and lunar eclipses, and moon phases, Phys. Rev. ST Phys. Educ. Res. **11**, 020102 (2015).

[20] G. W. Fulmer, Validating proposed learning progressions on force and motion using the Force Concept Inventory: Findings from Singapore secondary schools, Int. J. Sci. Math. Educ. **13**, 1235 (2015).

[21] G. Taasoobshirazi, M. Bailey, and J. Farley, Physics metacognition inventory part II: Confirmatory factor analysis and Rasch analysis, Int. J. Sci. Educ. **37**, 2769 (2015).

[22] G. W. Fulmer, H.-E. Chu, D. F. Treagust, and K. Neumann, Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model, Asia-Pac. Sci. Educ. **1**, 1 (2015).

[23] L. Ivanjek, M. Planinic, M. Hopf, and A. Susac, Student difficulties with graphs in different contexts, in *Cognitive and Affective Aspects in Science Education Research*, edited by K. Hahl, K. Juuti, J. Lampiselkä, A. Uitto, and J. Lavonen (Springer International Publishing, New York, 2017), pp. 167–178.

[24] S. I. Hofer, R. Schumacher, and H. Rubin, The test of basic Mechanics Conceptual Understanding (bMCU): Using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton's mechanics, Int. J. STEM Educ. **4**, 18 (2017).

[25] A. Susac, M. Planinic, D. Klemencic, and Z. Milin Sipus, Using the Rasch model to analyze the test of understanding of vectors, Phys. Rev. Phys. Educ. Res. **14**, 023101 (2018).

[26] J. M. Linacre, A user's guide to Winsteps Ministep Rasch-model computer programs, 2018, http://www.winsteps.com/winman/copyright.htm.

[27] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Mahwah, NJ, 2007).

[28] B. D. Wright and M. H. Stone, *Best Test Design* (MESA Press, Chicago, 1979).

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

[30] *Applications of Rasch Measurement in Science Education*, edited by X. Liu and W. J. Boone (JAM Press, Maple Grove, MN, 2006).

[31] B. D. Wright, A history of social science measurement, 1997, https://www.rasch.org/memo62.htm.

[32] J. M. Linacre, Winsteps Rasch tutorials 1–4, https://www.winsteps.com/tutorials.htm.

[33] J. R. Sick, Rasch measurement in language education, part 5: Assumptions and requirements of Rasch measurement, Shiken **14**, 23 (2010); http://hosted.jalt.org/test/sic_5.htm.

[34] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danmarks Paedagogiske Institut, Copenhagen, 1960).

[35] D. Andrich, Rating formulation for ordered response categories, Psychometrika **43**, 561 (1978).

[36] G. N. Masters, A Rasch model for partial credit scoring, Psychometrika **47**, 149 (1982).

[37] J. M. Linacre, *Many-Facet Rasch Measurement* (MESA Press, Chicago, 1989).

[38] B. D. Wright, Solving measurement problem with the Rasch model, J. Educ. Meas. **14**, 97 (1977).

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

[40] C. F. Herrmann-Abell and G. E. DeBoer, Investigating a learning progression for energy ideas from upper elementary through high school, J. Res. Sci. Teach. **55**, 68 (2018).

[41] S.-H. Paik, G. Song, S. Kim, and M. Ha, Developing a four-level learning progression and assessment for the concept of buoyancy, Eurasia J. Math. Sci. Technol. Educ. **13**, 4965 (2017).

[42] V. Mesic, K. Neumann, I. Aviani, E. Hasovic, W. J. Boone, N. Erceg, V. Grubelnik, A. Susac, Dz. Salibasic Glamocic, M. Karuza, A. Vidak, A. Alihodzic, and R. Repnik, Measuring students' conceptual understanding of wave optics: A Rasch modeling approach, Phys. Rev. Phys. Educ. Res. **15**, 010115 (2019).

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

[44] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[45] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[46] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[47] L. Bao, Theoretical comparisons of average normalized gain calculations, Am. J. Phys. **74**, 917 (2006).

[48] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).

[49] S. D. Willoughby and A. Metz, Exploring gender differences with different gain calculations in astronomy and biology, Am. J. Phys. **77**, 651 (2009).

[50] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists* (Lawrence Erlbaum Associates, Mahwah, NJ, 2000).

[51] N. Lasry, J. Guillemette, and E. Mazur, Two steps forward, one step back, Nat. Phys. **10**, 402 (2014).

[52] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's *d* for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14**, 010115 (2018).