Medical & Biological Engineering & Computing (2021) 59:471–482
https://doi.org/10.1007/s11517-020-02301-x

ORIGINAL ARTICLE

Item response theory as a feature selection and interpretation tool in the context of machine learning

Adrienne S. Kline[1,2,3] · Theresa J. B. Kline[4] · Joon Lee[3,5,6]

Received: 26 April 2020 / Accepted: 22 December 2020 / Published online: 3 February 2021

Abstract
Optimizing the number and utility of features to use in a classification analysis has been the subject of many research studies. Most current models use end-classifications as part of the feature reduction process, leading to circularity in the methodology. The approach demonstrated in the present research uses item response theory (IRT) to select features independent of the end-classification results without the biased accuracies that this circularity engenders. Dichotomous and polytomous IRT models were used to analyze 30 histological breast cancer features from 569 patients using the Wisconsin Diagnostic Breast Cancer data set. Based on their characteristics, three features were selected for use in a machine learning classifier. For comparison purposes, two machine learning–based feature selection protocols were run—recursive feature elimination (RFE) and ridge regression—and the three features selected from these analyses were also used in the subsequent learning classifier. Classification results demonstrated that all three selection processes performed comparably. The non-biased nature of the IRT protocol and information provided about the specific characteristics of the features as to why they are of use in classification help to shed light on understanding which attributes of features make them suitable for use in a machine learning context.

Keywords Item response theory · Machine learning · Feature selection · Breast cancer

1 Introduction

The suite of machine learning approaches to data analyses is becoming a commonly used methodology in many disciplines [24]. However, there have been calls recently to improve understanding of exactly what occurs with the feature data input into a neural network as it progresses through hidden layers and reaches an output [11, 32]. The goal of opening up the "black box" and defining the specific aspects of the methodology used should enhance consumers' ability to critically evaluate the knowledge claims of any specific study. This goal dovetails with calls to increase the transparency of machine learning methods and algorithms for clinicians [11] and by doing so improve the transfer of machine learning findings' to meaningful contributions to clinical care [6]. While not all machine learning methods are inherent black boxes such as support vector machines (SVMs), decision trees, and regression models, much criticism is directed at deep learning models. However, it is worth mentioning that even image data can be visualized as it passes through each layer of a convolutional neural network (CNN).

While the calls into the black box question have focused on the treatment of feature data, none has yet to also call into question the black box nature of the feature data themselves. Instead, the focus on features has been to select those that contribute the most to the predicted outcome, always giving the highest accuracy. Irrelevant features decrease the predictive accuracy of models and are computationally costly, while relevant features provide valuable information to the machine

✉ Adrienne S. Kline
askline1@gmail.com

1 Department of Biomedical Engineering, University of Calgary, Calgary, AB, Canada

2 Undergraduate Medical Education, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

3 Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

4 Department of Psychology, University of Calgary, Calgary, AB, Canada

5 Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

6 Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

learning system [4]. This is particularly problematic when data sets suffer from high dimensionality—too many features relative to the number of observations [18], which can often be the case in clinical studies. Feature selection processes that focus on their relationships to the classification at the end of the machine learning protocol are called wrapper approaches, while filter approaches are independent of the machine learning algorithms [3, 19, 20]. These "dust-bowl" empirical approaches exploit the chance characteristics of the data set and do little to contribute to our understanding of why the particular features are relevant.

This study proposes to begin to open up this portion of the machine learning black box by focusing on the psychometric properties of the features themselves, irrespective of their relationships to the classification variable. This is a clear benefit to using IRT—it does not use case classification information in the feature selection process, then use these same features and same data to generate machine learning classification accuracies. Our approach uses models developed under the auspices of modern test theory, capitalizing on the rich history of item and test development and, in particular, in the selection of items that are most appropriate for the assessment task at hand.

Modern test theory is frequently termed item response theory (IRT) and dichotomously scored data has recently surfaced in the machine learning literature including the following: comparisons of collaborative filtering with IRT models in item responses [1]; evaluation of natural language processing systems [17]; use of IRT to identify initial computer adaptive learning items [27]; and assessment of the utility of machine learning classifiers [22]. In all these examples, the data are dichotomous in nature (composed of 0 and 1 values) and the analyses have not been directed toward feature selection to be subsequently used in machine learning systems. In this study, we demonstrate how IRT can be of use in machine learning contexts through an example that first assesses features that are dichotomously scored and then polytomously scored. Then, based on the findings of these analyses, features will be selected based on their characteristics and their classification effectiveness will be assessed. The context for this evaluation assesses 30 features and the classification problem is the diagnosis of breast tumor (benign = 0 or malignant = 1) [21, 31].

1.1 IRT background

A fundamental assumption in IRT within the context of machine learning is that there is a linkage between the values that the feature takes on (e.g., 0 or 1 in a dichotomous example, or 1, 2, 3, 4 in an integer-scaled polytomous example) and the characteristic that underlies the resulting classification (e.g., benign-malignant). This characteristic is called the latent trait and is denoted by the term and symbol, theta (Θ). This linkage takes on a logistic functional form that specifies the probability that a given value on any feature is a function of the individual case's underlying Θ-value. To generate these functions IRT models analyze the entire pattern of all feature values for all cases simultaneously using a maximum likelihood iterative approach. Characteristics of the individual features as well as of the cases themselves become part of the interpretable output of the analyses.

1.2 Dichotomous IRT models

Models based on feature data that are coded with 0 or 1 include one-, two- and three-parameter models. There have been clear explanations of these models in other sources [8, 22, 27], so only a brief description of them follows. There are three IRT models that utilize values that are dichotomously scored. The most basic model is the one-parameter logistic model (1PL), sometimes referred to as the Rasch model [28]. Only one characteristic of each feature will be generated using this approach. This is the feature "threshold" (sometimes called the "location" or "difficulty"), and labeled "*b*". The *b*-characteristic is scaled using the normal distribution with a mean of 0.0 and a standard deviation of 1.0. Thus, features with higher *b*-values have a "higher threshold" insofar as the case must have a higher level of Θ (be closer to the malignant side of the distribution) to reach a value of 1 on the feature, whereas features with lower *b*-values are "easier" insofar as the case may have a lower level of Θ (closer to the benign side of the distribution) to reach a value of 1 on the feature. An example figure of a 1-PL feature is shown in Fig. 1. Based on the features' *b*-values, combined with the individual case's scores (0 or 1) on each of the features, each case's Θ-value can be generated. This Θ-value represents that case's location on the function, and thus where each falls on the continuum of "benign" to "malignant". These Θ-values are also scaled on a normal distribution.

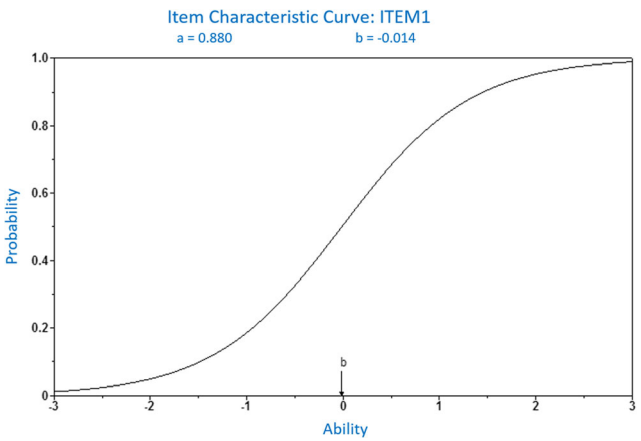

Fig. 1 Example of a feature function under a one-parameter (difficulty) logistic model; *X*-axis = Θ level and *Y*-axis = probability of belonging to the malignant group

The point at which the slope of the curve is steepest is located at the *b*-value parameter. This means that the feature is most informative for individuals who have Θ levels close to the *b*-value, in that a small change in Θ results in a large change in the probability of having a score of 1 on that feature. All of the logistic functions for features in a 1-PL model are assumed to have the same slope.

This assumption is relaxed in the two-parameter logistic model (2-PL). The allowance for features to have idiosyncratic slopes is a critically important difference between the 1-PL and 2-PL models. The slopes are denoted "a" parameters and capture the "discrimination" capability of the feature. Some features will have fairly flat slopes indicating that they are not very discriminatory (small changes in Θ result in small changes in the probability of having a score of 1 on that feature) while others will have steep slopes (very small changes in Θ result in very large changes in the probability of having a score of 1 on that feature) at the inflection point. Examples of two 2-PL features are shown in Fig. 2, one with a low slope (a) and one with a steep slope (b).

A third dichotomously scored model is the three-parameter logistic (3-PL) model, which adds a "*c*" parameter to the feature analysis. The *c*-value is a "pseudochance" parameter and is very instructive in the context of designing multiple-choice items for tests, as it assesses the degree to which a case with a very low (almost 0) level of Θ will have at least some probability of obtaining a value of 1 on the feature. This is helpful in the current context to determine if a feature is susceptible to generating "false positive" classification outcomes and thus lowering the specificity of the machine learning model. An example of a 3-PL feature is shown in Fig. 3. Because of the amount of information provided in this approach, it is the one that will be used for the dichotomous analysis in the current study.

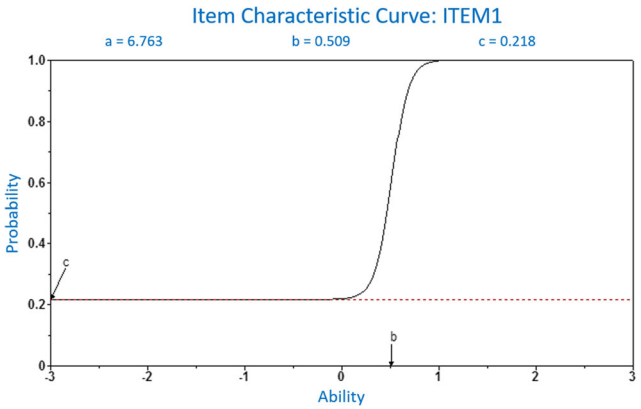

Fig. 3 Example of a feature function under a three-parameter (difficulty, slope, and pseudochance) logistic model; *X*-axis = Θ level and *Y*-axis = probability of belonging to the malignant group

1.3 Polytomous IRT model

The polytomous model that is used in this study is termed a graded response model where integer feature values are on ordered, categorical levels. In 1969, Samejima [30] pioneered this work and her model is flexible enough to allow for different numbers of categories for the various features (e.g., some features may only have two options such as gender or family history, while other features may have 4 levels or 5 levels) [30]. It is important to note, though, that in the data set-up, the options in the ordinal categories always mean "more" of the underlying construct, Θ.

A two-step process is used in the analyses of feature characteristics. First a series of *k*-1 (*k* = number of options for each feature) preliminary feature curves are generated based on "false dichotomies" that collapse feature options into two groups at various stages along the value continuum. For example, for a 4-option feature, 3 preliminary curves would be generated: (a) above 1 but less than 2, (b) above 2, but less than 3, and (c) above 3. A constraint on the model is that while different *b*-value

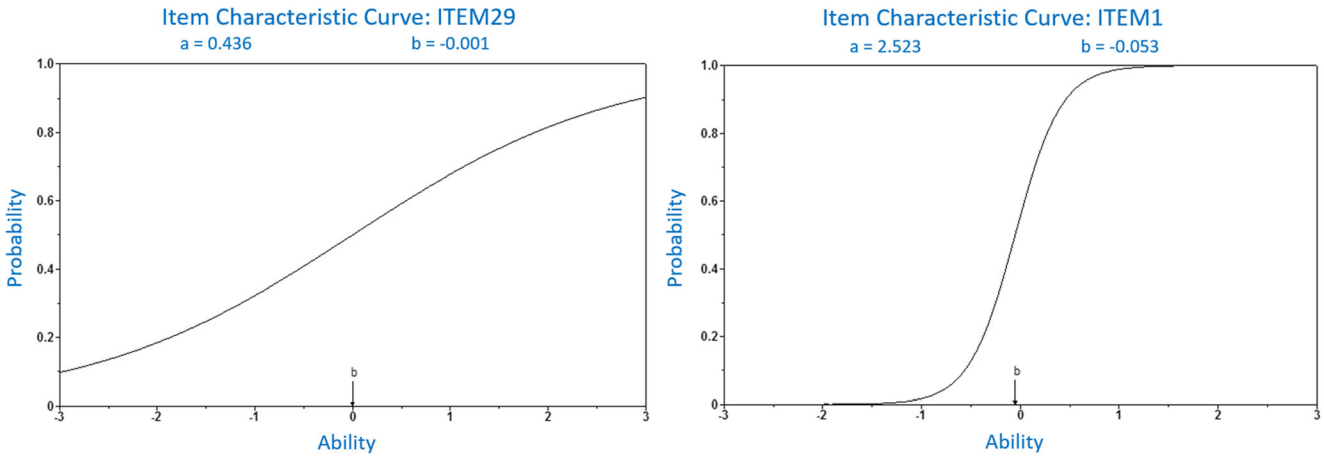

Fig. 2 Examples of two feature functions under two-parameter (difficulty and slope) logistic models, one with a low slope (a) and one with a steep slope (b); *X*-axes = Θ level and *Y*-axes = probability of belonging to the malignant group

parameters are generated for each feature's options, the a-value parameters for all feature options are assumed to be equal. The *b*-values for each option for each feature represent the Θ level needed to respond above a specific threshold with 50% probability. The *a*-values indicate slopes at the threshold.

The following threshold probabilities for each Θ level are estimated by:

(a) above 1 but lower than 2 = $P_i2(\Theta)$
(b) above 2 but lower than 3 = $P_i3(\Theta)$
(c) above 3 = $P_i4(\Theta)$

The formula for generating the separate threshold curves at various Θ levels is:

$$P_i(\Theta) = \frac{\left(\exp\left[a_i\Theta - b_{ij}\right]\right)}{\left(1 + \exp\left[a_i\left(\Theta - b_{ij}\right)\right]\right)} \quad (1)$$

Step 2 uses subtraction to estimate the probabilities for each option for each feature. The probability of responding at the lowest option or above (in this case 1, 2, 3, or 4), is 1.0; similarly, the probability of responding above the highest alternative (in this case above 4) is 0.0. The option probabilities, then are generated for each alternative in the 4-point scale example using the following equations:

Option 1: $P_i1(\Theta) = 1.0 - P_i2(\Theta)$
Option 2: $P_i2(\Theta) = P_i2(\Theta) - P_i3(\Theta)$
Option 3: $P_i3(\Theta) = P_i3(\Theta) - P_i4(\Theta)$
Option 4: $P_i4(\Theta) = P_i4(\Theta) - 0.0$

Using the above formulae, option response curves can be drawn (see Fig. 4). They show that at extremely low levels of the trait (Θ = − 3), the individual is almost certainly going to have a value of 1 on the feature. At somewhat higher levels of the trait (Θ = − 1.5), there is likely going to be a shift from a value of 1 to 2. At somewhat higher levels of the trait (Θ = − 0.75), there is likely going to be a shift from a value of 2 to 3. At even higher levels of the trait (Θ = + 0.25), there is likely going to be a shift from a value of 3 to 4. Curves such as these are extremely useful to identify at what level of Θ the particular feature differentiates. They also provide information as to the utility of the various response options.

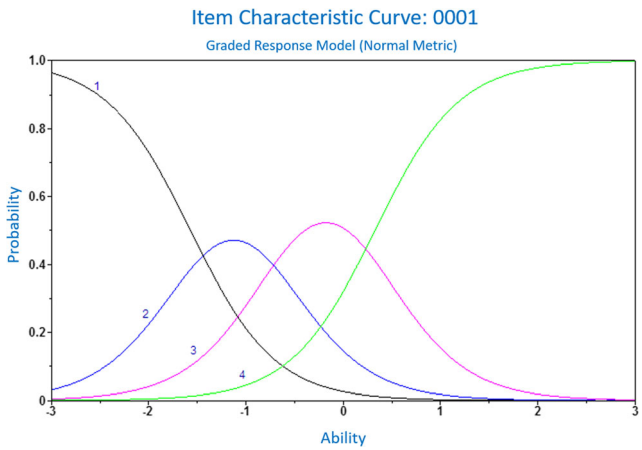

Fig. 4. Example of a feature function under a 4-level polytomous graded response model; *X*-axis = Θ level and *Y*-axis = probability of belonging to the malignant group

Both slope and location parameters are generated for each feature. Higher slopes indicate more differentiation between the category values. The overall location parameter corresponds to the level of "threshold" of the item insofar as higher values indicate more Θ is required to move into the higher categorical levels than for features with lower location parameters. In addition, category step parameters are provided to calculate the exact level of Θ at which shifts between categories occur.

1.4 Feature selection in machine learning

"Machine learning can be broadly defined as a set of computational methods using experience to improve performance or to make accurate predictions" and "consists of designing efficient and accurate prediction algorithms." [24]. These algorithms can only be as good as the "experience" (data) on which they are based, and this means the utility of the input data, or features. Thus, selecting the most appropriate features to use in a machine learning system has been of concern and the subject of much research for decades [9, 13, 15, 20]. This is particularly important when there are numerous features from which to select and a relatively small number of cases (hundreds rather than thousands) on which to run the machine learning model. How to pare down the number of features to a small, effective few has been the subject of numerous studies [26].

There are three major classes of multivariate machine learning-based feature selection methods; filter, wrapper, and embedded [12]. Filtering uses univariate approaches to feature selection; examples of this method include linear discriminate analysis (LDA) and chi-square. They relate the features with end-classification and are independent of the classification algorithm. Wrapper analyses, by contrast, are cyclical in nature iterating over the data set multiple times to identify a subset of features that provide the best classification accuracy, making it dependent on the classification algorithm and more computationally expensive than filtering. Examples include forward and backward selection and recursive feature elimination (RFE). Embedded methods, such as LASSO and ridge regression, also select features during the modeling process and are thus embedded within the algorithm. Feature weighting based on regularization penalizes features that do not contribute to the model, favoring less complex models with lower noise and avoids overfitting [5].

While there were many options to choose from, one wrapper (RFE) and one embedded (ridge regression) method were selected for several reasons to use in this study. They (1) highlight two different types of feature selection approaches; (2) are relatively easy to implement; (3) remove weakly relevant features using iterative protocols; (4) do not select algorithm-generated interaction variables that are intrinsic to the machine learning model; and (5) provide a rank-order of feature utility. These two methods in machine learning for feature reduction and selection are inherently dependent on their classification accuracies.

2 Methods

2.1 Data set

The Wisconsin Diagnostic Breast Cancer (WDBC) data set was used for this study obtained from the UCI Machine Learning Repository [7, 21, 31]. There are 569 cases with the diagnosis of either benign (n = 212) or malignant (n = 357) breast tumors. For each case, fine needle aspirate (FNA) that was taken from a breast mass was placed on a glass slide and stained to highlight cell nuclei. These images contained anywhere between 10 and 40 nuclei. Real-value features were recorded for 10 features of each nucleus for each case. The features were as follows: radius (mean of distances from the center to points on the perimeter); texture (standard deviation of gray-scale values); perimeter; area; smoothness (local variation in radius lengths); compactness (perimeter^2/area − 1.0); concavity (severity of concave portions of the contour); concave points (number of concave portions of the contour); symmetry; and fractal dimension ("coastline approximation" − 1). The final data set contains the following: mean, standard error, and worst (most negative) values for each case's set of FNA data, giving a total of 30 features.

Because all features are continuous variables, they needed to be modified to prepare them for input into the IRT programs. For the 3-PL dichotomous analysis, the data were median-split into the lower (coded 0) and upper (coded 1) 50th percentile levels for each feature. For the polytomous analysis, the feature data were quartile split 25th and below = 1, 26th–50th = 2, 51st–75th = 3, and 76th and above = 4.

2.2 Feature selection analyses

It was decided at the outset to limit the number of features selected that would be used in the classification analysis to three of the possible 30. Three were adopted to retain comparability to the number of features determined to be relevant in the original study of this data set [31].

The BILOG-MG 3 program [34] was used for the dichotomous model analysis. Marginal maximum likelihood is used to obtain parameter estimates and the program generates their estimates and standard errors. It assumes a binary logistic function of the relationship between features and the probability that a case with a specified Θ level will fall into one of two classifications on that feature [33]. Scores for each case, using the standard normal distribution, are also generated along the "degree of malignancy" continuum using all of the information in the parameter estimation process.

The PARSCALE 4.1 [25] program was used for the polytomous model analysis. It also uses a marginal maximum likelihood parameter estimation method. It assumes a log-normal prior distribution for the "a" parameters and a normal distribution for the "b" parameters [10]. The program generates the slope and location parameters for each feature as well as their standard errors. In addition, category step parameters are provided to calculate the exact level of Θ at which shifts between feature categories will occur. Similar to the dichotomous analysis, scores for each case, using the standard normal distribution, are also generated along the "degree of malignancy" continuum.

The wrapper feature selection method, RFE, iterates over all possible features (30 in the current study) and extracts the top "n", where "n" represents a user-defined number of highest-ranking features for selection. In the demonstration presented here, n = 3. It is advantageous to use fewer features in the classification analysis, as the final model is more interpretable. RFE iteratively removes features one by one, balancing maximization of classification accuracy within the prescribed number of features.

The embedded feature selection method, ridge regression, creates a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set is multi-collinear (high correlations between predictor variables). Ridge regression uses an L2 regularization (L2 penalty) (equal to the square of the magnitude of coefficients) in its feature selection protocol that decreases standard errors and reduces model complexity.

2.3 Classification analyses

A supervised machine learning protocol was adopted to assess which feature selection approach would provide the highest accuracy in case classification. The outcome for each analysis was whether the tumor was benign (0) or malignant (1). Python version 3.7.9. was used to conduct the machine learning analyses. To test the classification accuracy of using features selected via IRT, ridge regression, and RFE, a pipeline was developed in Python 3.7.9 environment using the Keras toolbox to perform a classification via a single hidden layer neural network. While it is possible to perform a Gridsearch and optimize all hyperparameters for optimization of the classification, this approach was considered out of scope and

defeated the purpose of assessing the three feature selection approaches using the same network.

Data were randomly divided into train (75%) and test (25%) data. A rectified linear unit "ReLU" activation function was selected from the input to the hidden layer as it makes backpropagation more efficient [2]. A sigmoid activation function was used from the single hidden layer to the output layer. The network was optimized using the "Adam" optimizer. "Adam" (derived from adaptive moment estimation) adjusts the learning rates for each parameter rather than using a single learning rate for all parameters [14], meaning no single learning rate was selected. Training occurred over a maximum of 150 epochs and the loss function a binary cross-entropy minimization. Size of the neural network used was 4 input dimensions, tested at 4, 8, and 12 neurons in the hidden layer. A 10-fold cross-validation was performed to eliminate bias in the model. Evaluation of feature selection approaches was based on an accuracy calculation: (TP+TN)/(TP+FP+FN+TN). "TP" denotes true positives, "TN" true negatives, "FP" false positives, and "FN" false negatives.

3 Results

3.1 IRT feature selection results

Table 1 shows the parameter estimates for each feature for the 3-PL dichotomous IRT model. Some of the features have high slope parameters (above 3) indicating their potential utility in discriminating along the benign-malignant continuum. Some features have high pseudochance parameters (c). For example, feature 2, texture mean, has a pseudochance value of 0.30. Thus, even in cases with very low levels of Θ, there is a reasonable probability that they would fall into the "1" classification on this feature. This feature would be expected to reduce the specificity of the overall classification. For a

Table 1 Slope, difficulty, and pseudochance feature parameters for dichotomously coded data using a three-parameter logistic model

Feature number	Feature name	Slope	Difficulty	Pseudochance
Feature 1	Radius mean	6.76	0.51	0.22
Feature 2	Texture mean	1.21	0.77	0.30
Feature 3	Perimeter mean	6.92	0.48	0.21
Feature 4	Area mean	6.89	0.51	0.22
Feature 5	Smoothness mean	0.91	0.54	0.22
Feature 6	Compactness mean	2.11	0.11	0.06
Feature 7	Concavity mean	3.38	0.11	0.04
Feature 8	Concave points mean	4.37	0.17	0.08
Feature 9	Symmetry mean	1.23	0.74	0.30
Feature 10	Fractal mean	0.53	1.14	0.30
Feature 11	Radius SE	2.11	0.61	0.28
Feature 12	Texture SE	2.20	1.96	0.48
Feature 13	Perimeter SE	2.09	0.47	0.23
Feature 14	Area SE	3.90	0.53	0.24
Feature 15	Smoothness SE	1.86	1.81	0.48
Feature 16	Compactness SE	1.24	0.18	0.10
Feature 17	Concavity SE	1.43	0.13	0.08
Feature 18	Concave points SE	1.46	0.21	0.11
Feature 19	Symmetry SE	2.11	1.84	0.48
Feature 20	Fractal SE	0.86	0.32	0.15
Feature 21	Radius worst	9.14	0.44	0.19
Feature 22	Texture worst	1.16	0.88	0.33
Feature 23	Perimeter worst	8.38	0.37	0.18
Feature 24	Area worst	9.20	0.44	0.19
Feature 25	Smoothness worst	0.86	0.53	0.22
Feature 26	Compactness worst	1.64	0.16	0.09
Feature 27	Concavity worst	2.43	0.15	0.08
Feature 28	Concave points worst	3.62	0.19	0.09
Feature 29	Symmetry worst	1.50	0.93	0.35
Feature 30	Fractal worst	0.86	0.36	0.17

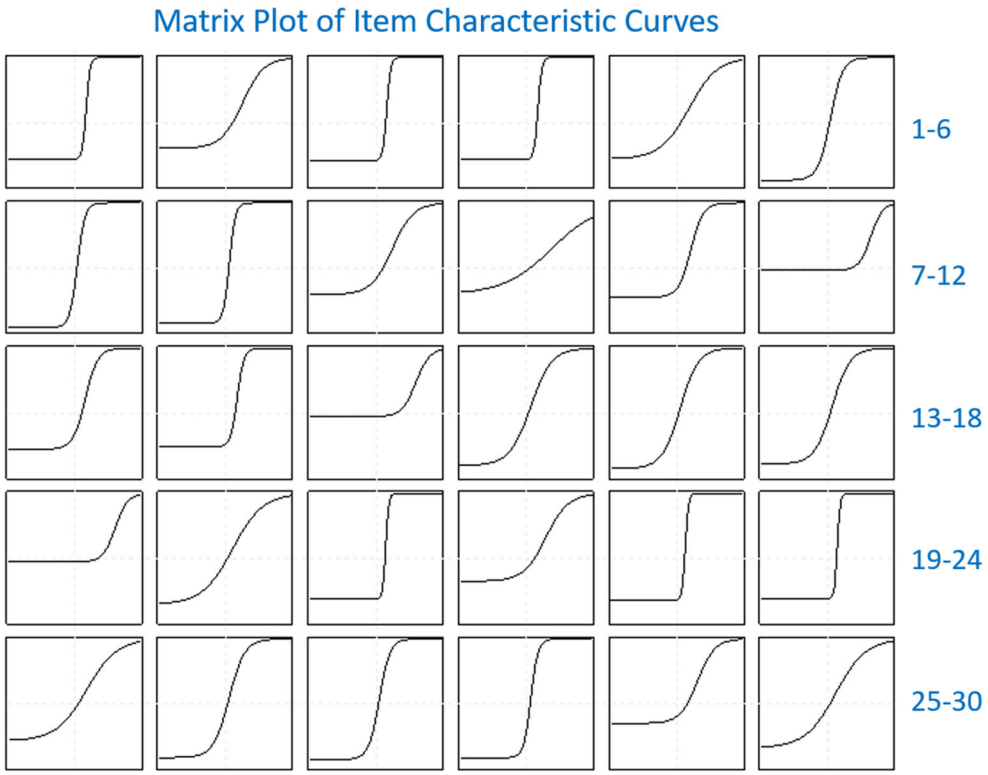

Fig. 5 Matrix of the 30 dichotomously coded feature functions under the three-parameter logistic model; *X*-axes = Θ level and *Y*-axes = probability of belonging to the malignant group

dichotomous variable of presence cancer or not, clearly, a pseudochance parameter of 0.5 would be no better than chance, so acceptable pseudochance parameters will be inevitably influenced by the number of classification categories. It is important to note that while not present in our data set, there may exist a trade-off between slope and the pseudochance parameter in other data sets where alternative features may be selected preferentially despite having a less discriminating slope.

Figure 5 shows a matrix of the feature functions. Features that have a strong sigmoidal function, with a steep slope are features that are most likely to be helpful in making discriminations between those with low and high Θ levels (i.e., those with benign versus malignant tumors). An examination of these functions in combination with a perusal of the information in Table 1 helps to identify those features with both high discrimination parameter (slope) and low pseudochance. Doing so highlights the particular utility of feature 7, concavity mean, feature 8, concave points mean, and feature 28, concave points worst. The point-biserial correlation between case scores on Θ and their actual benign-malignant classifications was 0.76 ($p < 0.001$). This information is useful in that it indicates whether or not the underlying Θ continuum actually reflects the assumed construct.

Table 2 shows the parameter estimates for each feature using the polytomous IRT model. Similar to Table 1, high slope parameters (above 2) are noted for feature 7, concavity mean, feature 8, concave points mean, and feature 28 concave points worst.

The category step parameters are used to create the feature functions. Note that they are the same for all items, but the location at which they start is unique to each feature. For example, for feature 1, radius mean, the location (b) parameter is − 0.64. The category steps are 0.93, 0.04, and − 0.97, respectively. Using the subtraction approach, the shift from option 1 to 2 will occur at Θ level − 1.57 (− 0.64–0.930); the shift from option 2 to 3 will occur at Θ level − 0.68 (− 0.64–0.04); and the shift from option 3 to 4 will occur at Θ level 0.33 (− 0.64–(− 0.97)). Figure 6 shows a matrix of the 30 feature functions, and the upper left first panel (radius mean) verifies the category step values calculated above.

Examining these functions, feature 1, feature 6, feature 7, feature 8, feature 27, and feature 28 all have functions that differentiate clearly along the benign-malignant continuum, with clear breaks between the four option categories. The options of 2 and 3 are subsumed under the options of 1 and 4 for most of the other features. This indicates that while the feature differentiates those cases at the lower and upper 25th percentiles, but not those with more moderate values on that feature. This analysis sheds a somewhat different perspective on the features than did the dichotomous analysis. The point-biserial correlation between case scores on Θ and their actual benign-malignant classifications was 0.81 ($p < 0.001$).

Table 2 Slope, location, and category step feature parameters for polytomously coded data using a graded response model

Category step parameters (SE)		0.93 (0.011), 0.04 (0.010), − 0.97 (0.011)	
Feature number	Feature name	Slope	Location
Feature 1	Radius mean	2.31	− 0.64
Feature 2	Texture mean	0.65	− 0.32
Feature 3	Perimeter mean	0.88	− 0.96
Feature 4	Area mean	0.47	2.32
Feature 5	Smoothness mean	0.70	− 0.28
Feature 6	Compactness mean	1.79	− 0.35
Feature 7	Concavity mean	2.85	− 0.38
Feature 8	Concave points mean	2.87	− 0.36
Feature 9	Symmetry mean	0.66	− 0.49
Feature 10	Fractal mean	0.46	− 0.20
Feature 11	Radius SE	0.93	− 0.47
Feature 12	Texture SE	0.46	− 0.23
Feature 13	Perimeter SE	1.09	− 0.43
Feature 14	Area SE	1.16	− 0.40
Feature 15	Smoothness SE	0.41	− 0.30
Feature 16	Compactness SE	0.90	− 0.27
Feature 17	Concavity SE	1.03	− 0.29
Feature 18	Concave points SE	1.02	− 0.30
Feature 19	Symmetry SE	0.43	− 0.54
Feature 20	Fractal SE	0.69	− 0.39
Feature 21	Radius worst	1.38	− 0.08
Feature 22	Texture worst	0.64	− 0.42
Feature 23	Perimeter worst	1.47	− 0.15
Feature 24	Area worst	1.46	− 0.75
Feature 25	Smoothness worst	0.68	− 0.48
Feature 26	Compactness worst	1.38	− 0.32
Feature 27	Concavity worst	1.82	− 0.36
Feature 28	Concave points worst	2.49	− 0.39
Feature 29	Symmetry worst	0.61	− 0.26
Feature 30	Fractal worst	0.69	− 0.35

The selection of features using the IRT results is a somewhat subjective process. However, the best performing features are those with high slope values and low pseudochance parameters based on the dichotomous results, as well as high slope values that provide good separation for all four of the quartile groups based on the polytomous analysis. These three dimensions were used to optimize the decision of which features to select using the weighted sum method for multi-objective optimization [23]. Although this approach allows for differential weighting of the dimensions, we have no particular rationale for doing so in this study, and thus, all dimensions were unit-weighted. The data points within each dimension were standardized to the normal distribution, and the pseudochance values were multiplied by − 1 so that all values would be on the same metric and higher values indicate better performance. Values for these transformed data points were then summed across features (see Table 3). The results showed that concavity mean, concave points mean, and concave points worst were the features that showed the highest promise as being useful. The equation for calculating the multi-objective optimization (MOO) value is:

$$U = \sum_{i=1}^{n} w_i F_i(x)$$

This metric allowed us to objectively select the top 3 features using the IRT approach, where *wi* is the unit weight applied, and $F_i(x)$ is the function multiplied by the individual weights in this case, slope from dichotomous variables, pseudochance parameter, and slope from polytonomous data. The variable U is synonymous with the MOO value numeric reported in Table 3.

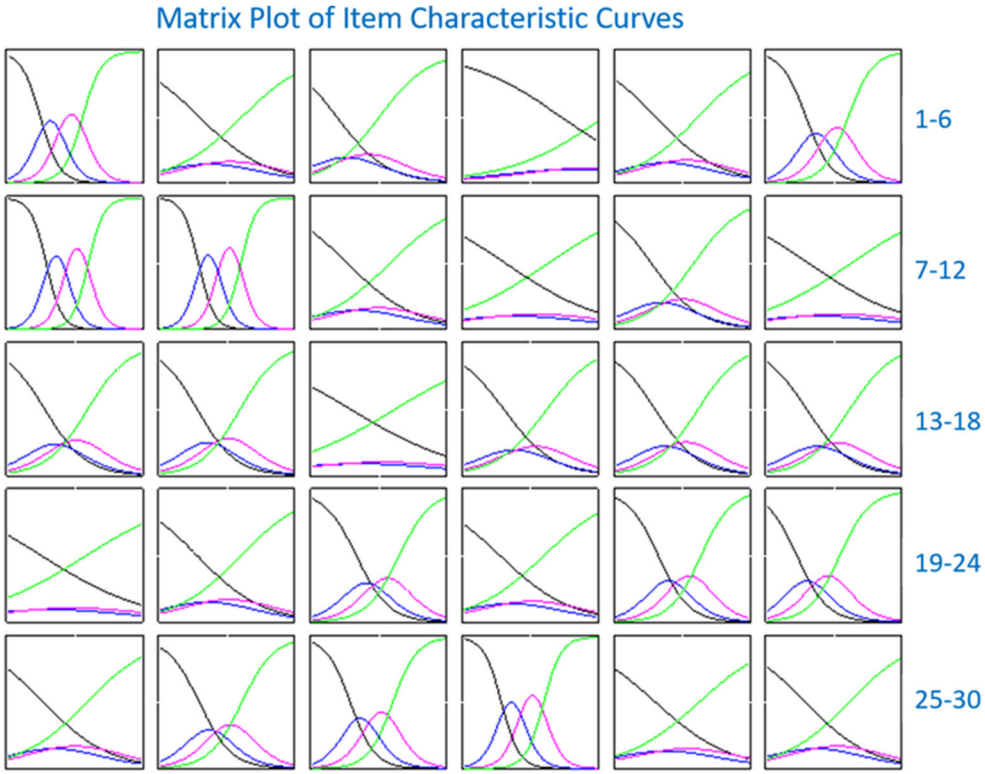

Fig. 6 Matrix of the 30 polytomously coded feature functions under the graded response model; *X*-axes = Θ level and *Y*-axes = probability of belonging to the malignant group

3.2 Machine learning feature selection results

RFE led to the selection of the top three features of concavity worst, concavity mean, and radius worst. Ridge regression's three features with the highest positive coefficients were as follows: smooth worst, concave points worst, and symmetry worst.

3.3 Classification results

Table 4 shows the results of the test classification analyses using the features based on the different methodological approaches. Evaluation of feature selection is reported on the basis of accuracy. None of the accuracies was statistically different from each other ($p > 0.05$).

4 Discussion

There were several purposes of this study. First, was to demonstrate the similarities and differences between IRT models that utilize dichotomous versus polytomous integer-scaled features. Second was to use the characteristics of the features to identify those that would be most useful in the input phase of a machine learning system. Third was to compare the features identified using IRT with two machine learning feature selection methods. Fourth was to assess the various feature combinations' utilities in an ANN classification problem.

The first two purposes were accomplished by demonstrating why both dichotomous IRT and polytomous IRT provide insight as to the inherent characteristics of the features as they related to an underlying construct—in this case, degree of breast cancer malignancy. The slope, threshold, and pseudochance parameters of the dichotomous IRT show which features have high discriminability, at what point along the construct feature discriminate best, and the extent to which cases that have very low levels on the degree of malignancy may be inadvertently assigned to be in the "malignant" category based on that feature. The polytomous IRT highlighted how well the feature was able to discriminate at all levels of the construct. Features that had no meaningful discrimination for the middle 50% of the cases were not very useful. It was found that there was a higher correlation between the actual classification and the underlying "degree of malignancy" (Θ) scores using the polytomous IRT (0.81) than the dichotomous IRT (0.76). This is not surprising, as there is more information available in the former than the latter in generating the individual case scores.

IRT-based feature selection resulted in similar outcomes as that of two different machine learning feature selection (RFE and ridge regression) approaches. Two of the three features selected using IRT results (concavity mean, concavity points worst) overlapped with the RFE and ridge regression methods. This suggests that the IRT as a feature selection tool is quite comparable to using machine learning algorithms in terms of final outcomes. There were no significant differences

Table 3 Feature parameters used to create IRT-based multi-objective optimization (MOO) values

Feature number	Feature name	Standardized slope (dichotomous)	Standardized pseudochance (*− 1) (dichotomous)	Standardized slope (polytomous)	MOO value
Feature 1	Radius mean	1.39	− 0.04	1.63	2.99
Feature 2	Texture mean	− 0.70	− 0.68	− 0.70	− 2.08
Feature 3	Perimeter mean	1.45	0.05	− 0.37	1.12
Feature 4	Area mean	1.44	− 0.04	− 0.95	0.45
Feature 5	Smoothness mean	− 0.82	− 0.04	− 0.63	− 1.48
Feature 6	Compactness mean	− 0.37	1.26	0.90	1.80
Feature 7	Concavity mean	0.11	1.42	2.39	3.93
Feature 8	Concave points mean	0.49	1.10	2.42	4.00
Feature 9	Symmetry mean	− 0.70	− 0.68	− 0.68	− 2.06
Feature 10	Fractal mean	− 0.96	− 0.68	− 0.96	− 2.61
Feature 11	Radius SE	− 0.37	− 0.52	− 0.30	− 1.19
Feature 12	Texture SE	− 0.33	− 2.14	− 0.96	− 3.43
Feature 13	Perimeter SE	− 0.37	− 0.12	− 0.08	− 0.57
Feature 14	Area SE	0.31	− 0.20	0.02	0.13
Feature 15	Smoothness SE	− 0.46	− 2.14	− 1.03	− 3.63
Feature 16	Compactness SE	− 0.69	0.94	− 0.35	− 0.10
Feature 17	Concavity SE	− 0.62	1.10	− 0.16	0.31
Feature 18	Concave points SE	− 0.61	0.86	− 0.18	0.07
Feature 19	Symmetry SE	− 0.37	− 2.14	− 1.00	− 3.51
Feature 20	Fractal SE	− 0.84	0.53	− 0.64	− 0.94
Feature 21	Radius worst	2.28	0.21	0.33	2.82
Feature 22	Texture worst	− 0.72	− 0.93	− 0.71	− 2.36
Feature 23	Perimeter worst	2.00	0.29	0.45	2.74
Feature 24	Area worst	2.31	0.21	0.44	2.96
Feature 25	Smoothness worst	− 0.84	− 0.04	− 0.65	− 1.52
Feature 26	Compactness worst	− 0.54	1.02	0.33	0.80
Feature 27	Concavity worst	− 0.24	1.10	0.95	1.80
Feature 28	Concave points worst	0.20	1.02	1.89	3.11
Feature 29	Symmetry worst	− 0.59	− 1.09	− 0.75	− 2.43
Feature 30	Fractal worst	− 0.84	0.37	− 0.64	− 1.11

between testing accuracies using the different features. However, an important difference in the techniques is that there is a far better understanding of why these features are likely to perform well by using the IRT approach. It is obvious from the features selected that the degree of nuclei concavity is indicative of malignancy.

This study has highlights the strengths of using IRT models within the context of machine learning in terms of feature selection, a problem that has been identified and into which much research effort has been expended. This study provides an additional piece of that important puzzle, with an added aim of improving our understanding of the characteristics of the feature as it relates to an underlying construct. Its methodological approach in feature selection offers a high degree of visual interpretability over the distribution of the data points—in this case, patients with possible malignant breast cancer. This holds promise in more effectively integrating machine learning findings with clinical practice, a problem that has plagued big data analytics [29].

Another benefit of using IRT for feature selection is that it is independent of the classification outcome. Many machine learning technologies use the entire data set of features as well as the classification outcome to select features. Then, the machine learning algorithm uses those features to correctly classify the cases. This methodological circularity is problematic. While cross-validation attempts to correct for this underlying accuracy bias, performing hyperparameter selection prior to testing is inappropriate as this cannot be remedied by the cross-validation. This feature subset selection bias engenders models that overfit the data, have poor generalizability to new information, and result in inflated accuracy rates [16].

Table 4 Test sample classification accuracies (%) for neural network analyses with 4, 8, and 12 hidden neurons using features selected via IRT and two machine learning–based methods (RFE and ridge regression)

Approach	Features	Accuracy (%) ± (SE) 4 hidden neurons	Accuracy (%) ± (SE) 8 hidden neurons	Accuracy (%) ± (SE) 12 hidden neurons
IRT	Concavity mean Concave points mean Concave points worst	93.3 ± 0.8	91.5 ± 0.2	91.9 ± 0.3
RFE	Concavity worst Concavity mean Radius worst	94.5 ± 0.5	93.8 ± 0.4	93.7 ± 1.1
Ridge regression	Smooth worst Symmetry worst Concave points worst	91.2 ± 0.9	90.5 ± 0.4	90.7 ± 0.6

There are several limitations of this study. The first is that the data set used carried continuous values. The high level of precision of this type of data was truncated when it was converted to dichotomous- and quadrature-level data for use in the IRT models. However, many of the variables in data sets are in dichotomous or multi-category format to begin with, so demonstrating how IRT works with continuous values that can be converted, was a worthwhile exercise. Second, the splits of the data into the dichotomous and 4-level categories were based solely on the percentiles of the distribution. It might have been more useful to split the data based on values that were of known clinical relevance, or verification of a 4-category system over an "*n*" number of categories (i.e., 3, 5). However, it is important to note that as the number of categories increases, the likelihood of exclusivity to that one category decreases. Again, however, the way the data were split in this example is easy to translate into other contexts. Third, IRT is univariate approach and does not consider combinations of features, meaning weak features that become discriminative in the presence of the other cannot be investigated. Future research in feature selection using the suggested IRT approaches demonstrated in this study would give insight into the maximum number of categories appropriate for the classification task. Fourth, we restricted the number of features to three. While using more features from all feature selection approaches might increase the classification accuracy, this was not the aim of the study. The aim was to demonstrate IRT's utility in feature selection and as an explanatory method as to "why" the underpinning feature was selected without it being contingent on the classification algorithm.

Future work into using this methodology warrants investigation using alternative data sets. Follow-up work likely in the form of a Monte-Carlo study design is necessary to determine the "break even" point between the pseudochance parameter and the slope in feature selection.

5 Conclusion

There are a number of reasons for limiting the selected number of features for machine learning. One of the more obvious is that often sample sizes of cases are relatively small compared to the number of possible features that can be used to classify/cluster the data. In instances like this, overfitting the data and non-generalizability of the findings is problematic. Thus, effective culling of features is needed. Other issues include that some features may be more expensive, time-consuming and/or invasive to collect relative to others. Thus, feature selection decisions often need to be carefully considered. Each method to doing so has strengths and weaknesses. The proposed IRT approach introduced in this study claims as a strength, clarification of why some features might perform more effectively than others based on their internal characteristics that they share with an underlying construct—in this instance, the degree of malignancy of a breast cancer tumor.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D.E. Pritchard. Model-based collaborative filtering analysis of student response date: machine learning item response theory. In Proceedings of the 5th International Conference on Educational Data Mining, 95-102. Chinia, Greece, June 19-21, 2012.
2. J. Brownee. A gentle introduction to the rectified linear unit (relu). In Better Deep Learning, Retrieved from: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks. January 9, 2019.
3. Cai J, Luo J, Wang S, Yang S (2018) Feature selection in machine learning: a new perspective. Neurocomputing 300:70–79
4. Chandrashekar F, Sahin G (2014) A survey on feature selection methods. Computers and Electrical Engineering 40:16–28
5. De Vlaming R, Groenen PJF (2015) The current and future use of ridge regression for prediction in quantitative genetics. BioMed Research International:1–18
6. Deo RC (2015) Machine learning in medicine. Circulation 132(20): 1920–1930

7. D. Dua and C Graff. In UCI machine learning repository. irvine, California: University of California, School of Information and Computer Science [http://archive.ics.uci.edu/ml], 2019.
8. S.E. Embretson and S.P. Reise. Item response theory for psychologists. 2000.
9. M.A. Hall and L.A. Smith. Practical feature subset for machine learning. In C.McDonald (Ed.), Computer Science '98 Proceedings of the 21st Australian Computer Science Conference, 181-191. Springer, Perth, 1998.
10. K.T. Han. Parscale. In BB. (Ed.) Frey, editor, The SAGE encyclopedia of educational research, measurement, and evaluation, 1208-1210. Sage, Thousand Oaks, 2018.
11. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, Lee MJ, Asadi H (2019) Peering into the black box of artificial intelligence: evaluating metrics of machine learning methods. American Journal of Roentgenology 212(1):38–43
12. A. Jovic, J. Prados, and M. Hilario. A review of feature selection methods with applications. In 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 1200-1205. IEEE, doi: https://doi.org/10.1109/MIPRO.2015.7160458, Opatija, Croatia, 2015.
13. Kalousis A, Prados J, Hilario M (2007) Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems 12(1):95–116
14. D.P. Kingma and J.L. Ba. Adam: A method for stochastic optimization. In 3rd International conference on Learning Representations (ICLR). San Diego, California, May 7-9, 2015. Retrieved from https://arxiv.org/pdf/1412.6980.pdf.
15. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artificial Intelligence 97(1-2):273–324
16. Krawczuk J, Lukaszuk T (2016) The feature selection bias problem in relation to high dimensional gene data. Artificial Intelligence in Medicine 66:63–71
17. J.P. Lalor, H. Wu, and H. Yu. Building an evaluation scale using item response theory. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 648-657. Austin, Texas, November 1-5, 2016.
18. Lee CH, Yoon HJ (2017) Medical big data: promise and challenges. Kidney Research and Clinical Practice 36(1):3–11
19. H. Liu and R. Setiono. A probabilistic approach to feature selection - a filter solution. In Proceedings of 9th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, 319-327, Fukuoka, Japan, 6 1996.
20. Liu C, Wang W, Zhao Q, Shen X, Konan M (2017) A new feature selection method based on a validity index of feature subset. Pattern Recognition Letters 92:1–8
21. Mangasarian OL, Street WN, Wolberg WH (1995) Breast cancer diagnosis and prognosis via linear programming. Operations Research 43(4):570–577
22. Marinez-Plumed F, Pudencio RBC, Martinez-Uso J, Hernandez-Orallo A (2019) Survey of multi-objective optimization methods for engineering. Artificial Intelligence 271:18–72
23. Marler RT, Arora JS (2004) Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization 26(6):369–395
24. M. Mohri, A. Rostamizadeh, and A. Talwalker. Foundations of machine learning. MIT press, 2012.
25. E. Muraki and D. Bock. PARSCALE 4. Lincolnwood, IL: Scientific Software Inc., 2003.
26. Parmar C, Gossman P, Bussink J, Lambin P, Aerts HJWL (2015) Machine learning methods for quantitative radiomic biomarkers. Scientific reports 5(13087):1–11
27. Pliakos K, Seang-Hwane J, Park JY, Cornillie F, Vens C, Van den Noortgate W (2019) Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. Computers and Education 137:91–103
28. G. Rasch. Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen, Denmark, 1960.
29. Rumsfeld JS, Joynt KE, Maddox TM (2016) Big data analytics to improve cardiovascular care: promise and challenges. Nature Reviews Cardiology 13(6):350–359
30. Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement 34(4):100
31. Wolberg WH, Street WN, Heisy DM, Mangasarian OL (1995) Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology 26(7):272–796
32. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H (2018) Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Annals of Translational Medicine 6(11):216–226
33. Zimowksi MF (2018) BILOG-MG. In: Frey BB (ed) The SAGE encyclopedia of educational research, measurement, and evaluation. Sage, Thousand Oaks, California, pp 199–202
34. Zimowski M, Muraki E, Mislevy R, Bock D (2003) BILOG-MG, vol 3. Scientific Software Inc., Lincolnwood

Adrienne Kline completed her PhD in the Biomedical Engineering Program and is now working on completing a Doctor of Medicine at the University of Calgary, Canada.

Theresa Kline is a professor emeritus in the Department of Psychology at the University of Calgary and author of the book "Psychological Testing: A Practical Approach to Design and Evaluation".

Joon Lee is the Director of the Data Intelligence for Health Lab and an Associate Professor of Health Data Science in the Department of Community Health Sciences, University of Calgary.