Viability Of Machine-Learning Strategies To Solve Psychometric Problems

Arthur Trognon (arthur.trognon@gmail.com)
Clinicog (185 rue Gabriel Mouilleron, 54000 Nancy)
Youssouf Cherifi
INSERM, CNRS, Institut de la Vision, Sorbonne Université, 17 Rue Moreau, 75012 Paris)
Loïs Demange
Lorraine University (23 Boulevard Albert 1er, 54000 Nancy)
Cécile Prudent
BePsyLab, Angers University (5bis Boulevard Lavoisier, 49045 Angers)

Research Article

Keywords: psychometrics, validation, scale, machine-learning

Posted Date: February 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1314080/v1

Abstract

Validating scales for clinical purposes is a common process in medicine and psychology. Using machine-learning and statistics, we revalidated the Fenigstein & Vanable Paranoia Scale and showed that these kinds of approaches could be used both to achieve construct validity and criterion validity and could thus add an additional layer of evidence to traditional validation approaches. However, there is still a lot of work needed in order to evaluate the whole range of applications, disadvantages, and potential limitations of these approaches when applied for psychometric purposes.

Introduction

Validating scales to discriminate between clinical populations is a relatively common procedure in psychology and medicine. The current validity framework postulates that validity corresponds to the level of evidence and theoretical justification that supports the interpretation and use of scores given by a scale. Thus, it is not the scale itself that is formally validated, but rather the interpretation of the scores it generates[1–3].

The concept of validity generally crosses three major aspects: construct validity, content validity, and criterion validity. Construct validity is a central concept in psychology that allows us to identify the extent to which the proposed test allows us to identify the construct being measured, is generally obtained by correlating the measure with a certain number of other measures and whose correlation patterns are theoretically predictable[4]. In contrast to construct validity, content validity allows us to verify the representativeness of the items on a given instrument with regard to the construct under study, and the degree to which this instrument has an appropriate sample of items for the measurement of the construct under study[5]. Finally, criterion validity aims to measure the performance predicted by the measurement tool. In particular, it is underpinned by predictive validity, which aims to verify the performance of the test with respect to the criterion taken as the object of study, for example by predicting a diagnosis from the set of variables (i.e. items) considered[6].

Computational approaches, such as machine learning, have shown spectacular results on many numerical problems. Thus, a growing number of research fields are beginning to use these types of strategies to analyze data such as in physics[7], chemistry[8], ecology[9], and neuroscience[10,11], where machine-learning techniques have been applied to discriminate between a normally hard-to-identify clinical population (suicide ideators) and neurotypicals, using neural signatures of conceptual representations captured in fMRI and following the presentation of valence-bearing words on a screen.

In this paper, we postulate that machine-learning strategies may represent a promising avenue for solving psychometric problems, or, more generally, may allow for an additional layer of evidence in approaches to scale validity assessment.

In general, machine-learning methods are composed of two phases: a first phase that aims to create a model from a finite number of available data (i.e., observations), and then to evaluate the power of the model by solving a practical task, such as estimating a probability density, or assigning a class to an observation, on a different sample than the one used to build the model[12].

With regard to construct validity in particular, regressive approaches would seem to be particularly appropriate for solving this type of task. From a mathematical point of view, regression is a particular category of analysis that allows one variable to be approximated by a set of other variables that are correlated to it[13]. In a construct validation approach, we could transpose this approach by verifying the capacity of a regression model to reconstruct the variance observed in a sample on a scale measuring a construct, from samples coming from other scales measuring other constructs that are correlated to it.

In the same way, concerning criterion validity, and in particular predictive validity, we believe that supervised classification approaches could represent a candidate of choice. In machine learning, supervised classification is defined as an algorithmic categorization of objects, allowing to assign a class to each observation of a sample based on statistical data[14]. In a criterion validity approach, we could transpose this approach by verifying the capacity of a classification model to correctly identify a clinical group based on the self-reported answers of subjects to a questionnaire.

In order to validate our hypotheses, we worked on archival data from the University of Lorraine having studied the construct of paranoia through Fenigstein & Vanable's Paranoia Scale, designed to identify paranoid thinking, defined by the author as a style of thinking with exaggerated self-referential biases (e.g. suspiciousness, external locus of control feelings of resentment and ill will, mistrust...), and occurring in everyday life (Fenigstein and Vanable, 1992); this scale having been designed to identify paranoid thinking. suspiciousness, external locus of control feelings of resentment and ill will, mistrust...), and occurring in everyday life[15]; this scale having been cross-referenced with the Minnesota Multiphasic Personality Inventory 2 Restructured form (MMPI-2-Rf - RC6) Paranoia scale, designed to measure persecutory ideation (e.g., seeing others as threats to one's own life). e.g., seeing others as threats, representing oneself as a victim of others, being suspicious...[16]). Furthermore, and given that paranoia is equated in the clinical literature with delusions of anticipation of a threat or negative future events[17], and the existence of Bayesian reasoning biases in the cognitive science literature[18,19], these scales have been cross-referenced with two other clinical scales: the Neuroticism Revised Scale (NEGE-r - MMPI-2-Rf), which measures negative emotions such as anxiety, insecurity, preoccupation, as well as general tendency toward dramatization and negative anticipation[16]; and the Response Bias Scale (RBS - MMPI-2-Rf), designed to detect symptoms associated with cognitive response biases[20].

We conducted two experiments on these data: first, we assessed construct validity using regression techniques, as described above, by attempting to reconstruct the observed variance in the Fenigstein & Vanable scale from the available samples in the other scales (i.e., RC6; NEGE-r; and RBS) and observed the differential performance of the scales in this task. We then assessed criterion validity using

classification techniques, as described above, by attempting to automatically predict the clinical category associated with the observations (i.e., psychotypic or paranoid), from the different scales available, and again observing the differential predictive abilities of these tools.

For each of these experiments, control experiments were conducted. Control experiments ensure that the observed results are not random events, thus allowing for the distinction between signal and background noise[21]. In general, a positive control is a group that is known to definitely have an event when exposed to a stimulus; whereas the negative control is a variable where no response is expected when exposed to a given stimulus, thus allowing for a baseline value for background noise[22]. In the present study, we used the MMPI-2-Rf Paranoia scale as a positive control. Indeed, based on the same construct as the Fenigstein & Vanable scale, it should be able to reconstruct the variance of the Fenigstein & Vanable scale in the regression experiment. Similarly, given that the diagnosis of Paranoia is made on the basis of the MMPI-2-Rf T-scores, one would expect a ceiling effect to be reached when the diagnosis is predicted with this scale (i.e. the maximum performance that can be expected). More generally, this step will verify the possibility of the operation. In addition, we used two parameters as negative controls: age and gender of the respondent. Indeed, these variables should not be able to reconstruct the variance of the Fenigstein & Vanable scale, as these data are not informative about the paranoia construct in the regression experiment, nor should they be able to provide the information necessary to correctly identify paranoid individuals in the sample. More generally, this step will allow us to note the impossibility of achieving accurate discrimination in the absence of sufficient information about the measured construct.

Experiment 1: Performing Construct Validity Of Fenigstein & Vanable (1992) Scale Using Random Forest Regression

Material and Method :

Subjects :

Four hundreds and seventy five not preselected adults (mean age = 25.33 years, SD = 11.22 ; male n = 145 ; females n = 330) from the general French population participated in this study. Full measures were available for all subjects.

All participants received detailed information about the study purpose and objectives, and provided online informed consent to participate in the study. All procedures were conducted in accordance with the Declaration of Helsinki and the study protocol was approved by the Institutional Review Board Commission Nationale de l'Informatique et des Libertés (registration n°2225110v0).

Psychometric materials :

MMPI-2-RF: The Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) is a standardized psychological instrument for comprehensive assessment of psychopathology and/or personality assessing the subject’s psychological dynamics (e.g., psychopathological trouble, behavioral

tendencies…). We used the French version of this tool which consisted of 51 clinical scales including a scale for diagnosing paranoia (Restructured Clinical 6 - RC6) by implying questions about auto-justice moral, interpersonal sensitivity and mistrust. This tool was used as a positive control, measuring the same construct as the Fenigstein & Vanable scale. We also used two clinical scales from the MMPI-2-Rf, the Responses Biases Scales (RBS) and the Neuroticism Scale (NEGE), to test the construct validity of the Fenigstein & Vanable scale.

Fenigstein & Vanable (1992) Scale : This Paranoia Scale was designed to detect paranoia-related symptoms in nonclinical populations examining paranoid thinking[15]. In this work, we translated this tool to French and used it to verify its construct validity.

Machine-learning model :

The machine-learning approach consists of training a model with a subset of the provided dataset and then testing the trained model on an unseen subset. In this study, we used a Random Forest Regression model to assess the differential predictive power of participants' responses to the different clinical subscales of the MMPI-2-Rf. The aim was to reconstruct the variance of these same participants' responses to the Fenigstein & Vanable scale. Random forest is a “meta-estimator” that fits a number of decision trees on several sub-samples of the dataset in parallel. It uses averaging of multiple decision trees to both improve its predictive accuracy and control for over-fitting[23,24]. In this study, only the number of generated trees and the loss function were set by the experimentator, while the remaining hyperparameters were used as defaults (i.e. they were not explicitly set by us). The number of generated trees was set 10000 (In accordance with (Oshiro et al., 2012)[25]) and for the loss function, we used the “mean-squared error” (MSE).

Procedure :

Data were collected online through word of mouth and social media. The obtained dataset was then split randomly to constitute an independent training set (n=380) and a testing set (n=95) with the help of the scikit-learn library on Python.

To evaluate the differential performance of participants' response matrices to different clinical scales (RBS and NEGE; including all items) we used Machine-Learning regressors, this was done with the assumption that the constructs underlying these scales share informatively useful dimensions that will allow for automatic inference of participants' responses to the Fenigstein & Vanable scale.

Additionally, we used similar regressors to compare our results to those of two control factors: the inference of participants' responses to the Fenigstein & Vanable scale from the RC6 scale of the MMPI-2-Rf, designed and validated to measure the paranoia construct (i.e. This experiment measures the feasibility of the process); and the inference of participants' responses to the Fenigstein & Vanable scale from their age and gender (i.e., two variables). This was done with the assumption that these two factors should make participants' responses vary, when paranoia is not considered. In other words, we assessed

predictions under six conditions: RC6 alone, RBS alone, NEGE alone, RBS and NEGE together, RBS and NEGE together. We carried this for a reduced version (10 random items each) in order to fit the number of items with Fenigstein & Vanable scale ; and Age+Gender.

To compute similarity between predicted Fenigstein & Vanable full participant's responses (the whole 20 items) and the observed responses, we built a cosine similarity matrix for all conditions. This way, each predicted item is compared to its real counterpart and the similarity between the two is represented using a scaler between 0 and 1. Additionally, we used the Kruskal-Wallis test (similarity ~ prediction source ; according to the non-normal distribution of data) to assess if the power varies across all prediction sources, and a non-parametric Dunn-Bonferroni post-hoc procedure to analyze the interactions between variables.

Finally, we performed six two-way ANOVA (score ~ source*item ; i.e. one ANOVA for each prediction source) with regard to value source (two levels : predicted or observed) and item (one level per Fenigstein & Vanable item). This was done to statistically evaluate the differential performance of the different scales in the reconstruction of the variance of the Fenigstein & Vanable scale. The assumption here is that the capacity of a machine-learning algorithm to predict the observed variance of a sample of participants' answers to a questionnaire varies and that the answers to the other questionnaires provide sufficient information on the studied construct to infer its properties. The different interactions between variables studied were analyzed using the Tukey procedure, the machine learning models were coded in Python using functions from the scikit-learn library, and for statistical analysis we used R with RStudio v1.0.143 as the main interpreter.

Results and discussion

Results of the analysis of cosine similarity matrices (see "Procedure" above) can be observed in Figure 1 and Table 1, and results about the analysis of variance are shown in Table 2.

Similarity :

Concerning prediction from RC6 to Fenigstein & Vanable Scale, mean similarity was .99 (SD±.008), with minimal similarity at .961 and max similarity at 1, indicating a perfect prediction of participant's auto reported values by the algorithm. Then, from NEGE+RBS to Fenigstein & Vanable, mean similarity was .99 (SD±.009), with minimal similarity at .94 and maximal similarity at 1. Then, from RBS to Fenigstein & Vanable Scale, mean similarity was .98 (SD±.01), with minimal similarity at .94 and maximal similarity at 1. Then, from NEGE to Fenigstein & Vanable, mean similarity was .98 (SD±.01) with minimal similarity at .92 and maximal similarity at 1. Then, from the reduced version of NEGE+RBS, mean similarity was .98 (SD±.01). Finally, from Age+Gender to Fenigstein & Vanable scale, mean similarity was .94 (SD±.03) with minimal similarity at .82 and maximal similarity at 1.

Table 1 : Kruskal-Wallis test results. [n.s. : not significant ; RC6 (PC) : Restructured Clinical 6 (Positive Control) ; NEGE : Neuroticism Scale ; RBS : Responses Biases Scales ; R+N(R) : RBS+NEGE (Reduced) ; A+G (NC) : Age+Gender (Negative Control)]

Scale	RC6 (PC)	RBS+NEGE	NEGE	RBS	R+N(R)	A+G
RC6 (PC)	-	***p*<.001**	***p*<.001**	***p*<.001**	***p*<.001**	***p*<.001**
RBS+NEGE	***p*<.001**	-	n.s.	n.s.	***p*<.001**	***p*<.001**
NEGE	***p*<.001**	n.s.	-	n.s.	n.s.	***p*<.001**
RBS	***p*<.001**	n.s.	n.s.	-	n.s.	***p*<.001**
R+N(R)	***p*<.001**	***p*<.001**	n.s.	n.s.	-	***p*<.001**
A+G	***p*<.001**	***p*<.001**	***p*<.001**	***p*<.001**	***p*<.001**	-

These findings suggest that when the same dimensions are measured across two scales (e.g. between RC6 and Fenigstein & Vanable), similarity between real and predicted matrices will be high with low standard-deviation. Moreover, if one or more dimensions are shared between two or more scales (e.g. between Fenigstein & Vanable and RBS+Nege ; Fenigstein & Vanable and RBS ; Fenigstein & Vanable and NEGE), similarity between real and predicted matrices will be remain high with low-standard deviation, with a positive association between the number of shared dimensions and the predicting power (which can be suggested by the significant differences between RBS+NEGE predictions and NEGE alone). Furthermore, the prediction power will be higher when more items are measuring a shared dimension between two constructs can be shown by the significant difference between RBS+NEGE (Full) and RBS+NEGE (Reduced). Notably, these results also suggest that when the factor associated with clinical condition is controlled for, only Age+Gender would produce variation in the participants' responses.

Variance analysis of predicted responses :

The results of the analysis of variance (see "Procedure" above) are available in Table 2. The data suggest that all scales sharing common dimensions with paranoia are able to reconstruct the variance of participants' responses to the Fenigstein & Vanable scale.

Table 2 : Analysis of Variance test results. [n.s. : not significant ; RC6 (PC) : Restructured Clinical 6 (Positive Control) ; NEGE : Neuroticism Scale ; RBS : Responses Biases Scales ; R+N(R) : RBS+NEGE (Reduced) ; A+G (NC) : Age+Gender (Negative Control)]

Scale	Item	Source	Item:Source
RC6 (PC)	$F_{(19,3640)}$=61.60, p<.001	$F_{(1,3640)}$=1.49,p=.22	$F_{(19,3640)}$=0.13,p=1
RBS+NEGE	$F_{(19,3640)}$=55.67,p<.001	$F_{(1,3640)}$=1.38,p=.24	$F_{(19,3640)}$=0.19,p=.999
RBS	$F_{(19,3640)}$=55.27,p<.001	$F_{(1,3640)}$=17.82,p=.26	$F_{(19,3640)}$=0.17,p=.999
NEGE	$F_{(19,3640)}$=54.64,p<.001	$F_{(1,3640)}$=0.004,p=.95	$F_{(19,3640)}$=0.17,p=.999
R+N(R)	$F_{(19,3640)}$=55.67,p<.001	$F_{(1,3640)}$=1.38,p=.24	$F_{(19,3640)}$=0.19,p=.999
A+G (NC)	$F_{(19,3640)}$=71.35,p<.001	$F_{(1,3640)}$=7.75,p=.0054	$F_{(19,3640)}$=1.005,p=.45

These results suggest that all scales that share dimensions with the paranoia construct are able to automatically reconstruct the observed sample variance. In general, the only variables that failed to do so was the negative control (Age+Gender), which behaved opposite to all other conditions (i.e. they behaved like the positive controls). We thus postulated that this highlighted difference could be interpreted as a factor that does not depend on either age nor gender, and that the algorithm is not able to capture: it lacks the dimensions underlying paranoia. We also believe that predictability remains high when predictions are made using the "Age+Gender" factor since, if the "paranoia" factor is removed, then only age and gender will vary participants' responses (i.e., these factors predict neurotypical participants' responses).

In sum, the main findings indicate that when two scales are assessing the same dimensions, a prediction from one to the other will result in a high similarity matrix between real values and predicted ones, with low standard-deviation. Furthermore, when two or more scales sharing dimensions, so similarity between matrices will remain high, but in a lesser extent and with higher standard-deviation, and with a relation where the fewer dimensions you share between scales, the less similarity there will be and the higher the standard deviation will be (see “Similarity” section). In the same way, the more dimensions you share between scales, the more the algorithm would be able to reproduce original sample variance (see “Variance analysis of predicted responses” section) and sample characteristics.

Experiment 2 : Performing Criterion Validity Of Fenigstein & Vanable Scale Using Random Forest Classifiers.

Material and Method :

Subjects :

A subset (paranoiac : n=131 ; controls : n = 131) of the initial sample was used for this experiment. We assumed that in classification study, experimenters would have initial knowledge about clinical conditions and would perform classification using two balanced groups.

Psychometric materials :

MMPI-2-RF : We used the same Restructured Clinical 6 (RC6) to test the criterion validity of the Fenigstein & Vanable scale.
Fenigstein & Vanable (1992) Scale : We used the same Paranoia Scale (Fenigstein & Vanable, 1992).

Machine-learning model :

To assess if the Fenigstein & Vanable scale is sufficient to diagnose paranoia, we constructed a Random Forest Classifier. The classification performed in this study consists to predict MMPI-2-Rf paranoia diagnostic (i.e. T-score of MMPI-2-Rf Paranoia Scale > 80, using a binary variable : 0 or 1) using only participant's responses to Fenigstein & Vanable scale. In this study, only the number of generated trees and the criterion were set by the experimentator, while all other parameters were set to default. 10000 trees were used (accordingly with Oshiro et al. (2012) [8]) and the criterion was set as "entropy".

Procedure :

MMPI-2 was scored automatically using a Python program to avoid human-error. Normative data were encoded in Python to automate the transformation from raw data to T-score, then from T-score to diagnostic, according to MMPI-2-Rf cutoff (Paranoiac if T-score > 80). Then, 131 control individuals were selected randomly using Python script in order to have n=131 individuals in each group (i.e. $n_{paranoiac}$=131 and $n_{controls}$=131). The dataset was then split randomly to constitute an independent training set (n=209) and test set (n=53). A k-Fold (with k = 10) cross-validation was performed to assess the differential performance of the scales in automatically assigning participants' class from participants' responses to the respective scales, and accuracy, specificity, and sensitivity metrics were generated for all conditions.

Finally, two ANOVAs were performed: first, a one-way ANOVA (accuracy~scale), to check whether significant differences in performance exist according to the scale used to perform the class assignment (4 levels: RC6 (positive control); F&V; RBS+NEGE; A+G (negative control)). Then, a two-way ANOVA (score~scale*metric) was performed to control the performance of the scale in terms of sensitivity and specificity. All significant interactions were controlled using a Tukey post hoc procedure. Machine-learning algorithms were coded in Python using Sci-Kit Learn libraries and statistical analysis were coded in R and interpreted in RStudio v1.0.143.

Results and discussion

The cross-validation results for all conditions are presented in Figure 2. Random Forest classifiers that were trained on Fenigstein & Vanable scale data from 209 of 262 selected participants identified the home group of the remaining participants with an average accuracy of .81(±.4), with an average sensitivity of .50(±.15) and an average specificity of .93(±.05) (Figure 2). In contrast, classifiers trained on the positive control data (RC6) had a mean accuracy of .92(±.03), with a mean sensitivity of .78(±.13)

and a mean specificity of .96(±.05); whereas classifiers trained on the negative control data (ge+Genre) had a mean accuracy of .67(±.04), with a mean sensitivity of .37(±.2) and a mean specificity of .93(±.06).

Accuracy

Results for the accuracy measures are presented in the left panel of Figure 2 and Table 3. The scales showed differential performance in identifying participants' class of membership.

Table 3 : Tukey post-hoc test results for the accuracy. [RC6 (PC) : Restructured Clinical 6 (Positive Control) ; F&V : Fenigstein & Vanable ; R+N : RBS+NEGE (Combined) ; A+G (NC) : Age+Gender (Negative Control)]

Scale	RC6 (PC)	F&V	R+N	A+G (NC)
RC6 (PC)	-	***p*<.001**	***p*<.001**	***p*<.001**
F&V	***p*<.001**	-	*p*=.39	***p*<.001**
R+N	***p*<.001**	*p*=.39	-	***p*<.001**
A+G (NC)	***p*<.001**	***p*<.001**	***p*<.001**	-

Regarding accuracy, the one-way ANOVA (accuracy~scale) revealed a significant effect of the scale used to perform the automatic classification of the participants' category of belonging [$F_{(3,36)}$=36.45, p<.001]. Further post-hoc analysis revealed differential performance between all conditions (Table 3, Figure 2 (Left)), except between the F&V and R+N conditions, which showed similar performance.

Detection metrics

Results for the detection metric measures are presented in Table 4 and Figure 2. The scales showed differential performance in identifying patients, but similar performance in identifying control subjects.

Table 4 : Tukey post-hoc test results for sensitivity and specificity. [RC6 (PC) : Restructured Clinical 6 (Positive Control) ; F&V : Fenigstein & Vanable ; R+N : RBS+NEGE (Combined) ; A+G (NC) : Age+Gender (Negative Control)]

Metric	Sensitivity				Specificity			
Scale	RC6	F&V	R+N	A+G	RC6	F&V	R+N	A+G
RC6	-	***p*<.001**	***p*<.001**	***p*<.001**	-	*p*=.96	*p*=.96	*p*=.97
FV	***p*<.001**	-	*p*=.10	***p*<.001**	*p*=.96	-	*p*=1	*p*=1
R+N	***p*<.001**	*p*=.10	-	***p*<.001**	*p*=.96	*p*=1	-	*p*=1
A+G	***p*<.001**	***p*<.001**	***p*<.001**	-	*p*=.97	*p*=1	*p*=1	-

Regarding the detection measures, the two-way ANOVA (score~scale*metric) revealed a significant effect of scale [$F_{(3,72)}$=51.21, p<.001] and metric [$F_{(3,72)}$=487.54, p<.001], with a significant interaction between the scale used to generate the prediction and the type of metric used to measure performance [$F_{(3,72)}$=41.20, p<.001]. Further post-hoc Tukey procedure revealed significant differences between all scales regarding sensitivity measures, except between the F&V and R+N conditions (Table 4, Figure 2 - Middle); but no differences between different conditions regarding specificity measures (Table 4, Figure 2 - Right).

These results show that only the scales specifically measuring the paranoia construct are able to correctly attribute the clinical condition to paranoid participants, whereas all are able to correctly identify control individuals. However, a significant difference was observed between the performance of classifiers trained on the RC6 scales and those trained on the Fenigstein&Vanable scales. This could be explained by the fact that the diagnosis was produced from the normative data of the MMPI-2-Rf, from which the RC6 scale is derived. This bias should give it an advantage in performing the classification task. However, it should not be a problem in future work, given that in clinical trials the assignment of clinical condition to participants is made by medical and/or psychological diagnosis prior to the scales being administered, the controls of which will be able to play their role more objectively and with less bias.

General Discussion

Using machine-learning approaches, we were able to perform two common validation operations in psychometrics, namely construct validity and criterion validity.

First, we performed a construct validation operation using a Random Forest regression algorithm, showing that it was possible to reconstruct the observed variance of participants' responses to a questionnaire from responses to a questionnaire measuring the same constructs, or from responses to a set of questionnaires measuring constructs sharing dimensions with the construct of interest.

We then performed a predictive criterion validation operation using a Random Forest type classification algorithm, showing that it was possible to automatically identify the class to which the participants belonged, in relation to two clinical groups (paranoid and neurotypical), using scales specifically measuring the construct of interest, or a set of scales sharing dimensions with it.

To our knowledge, and even though machine-learning paradigms have previously been employed to discriminate between different clinical groups using Likert scales and classification algorithms[26], we believe that this work is the first to investigate the viability of these approaches within the overall framework of the clinical scale validation approach. As such, it could achieve a proof of concept that big data approaches could be applied to solve psychometric problems, and allow for an additional layer of evidence when employed in concert with traditional validation approaches.

In light of these experiments, we can produce some recommendations for further research in machine learning applied to psychometrics: first, concerning construct validity, it is necessary to show that the regression algorithm is able to (1) capture the responses of participants with high accuracy, (2) reproduce the variance of the response of the participants of the scale we seek to validate, and (3) show that this operation is not feasible using another scale that does not share any dimension with the construct of interest. Concerning the criterion validity, it is necessary to show that the classification algorithm is able to (1) capture the membership classes of the participants with a high sensitivity and specificity, (2) to show that this operation is feasible using other scales measuring dimensions shared with the construct of interest and (3) that this operation is not feasible in the opposite case

Finally, we believe that beyond the framework of validation, and even beyond the general framework of the clinic, these types of approaches could have interesting perspectives in the research domains. Indeed, pre-trained classification models on clinical populations could perform automatic predictions on new and much more massive unlabeled datasets, and thus allow for large-scale multidimensional psychometric studies. One could also imagine linking dimensions of different natures, for example by performing predictions of psychophysical data from cognitive and/or psychometric measures, thus allowing the creation of synthetic data sets to train theoretical models. Further research will therefore be needed to further evaluate the potential of these approaches in research settings and especially in terms of clinical translation possibilities.

Data Availability Statement

The dataset analysed during the current study is available in (Supplementary File 1).

References

1. Gonzalez, O., MacKinnon, D. P. & Muniz, F. B. Extrinsic convergent validity evidence to prevent jingle and jangle fallacies. *Multivar. Behav. Res.* **56**, 3–19 (2021).

2. Messick, S. Validity. em r. linn (org.), educational measurement.(13-103). *N. Y. NY Am. Counc. Educ. Macmillan Publ. Co.* (1989).

3. Schmeiser, C. B., Welch, C. J. & Brennan, R. L. *Educational measurement*. (American Council on Education and Praeger Publishers, Westport, CT, 2006).

4. Westen, D. & Rosenthal, R. Quantifying construct validity: two simple measures. *J. Pers. Soc. Psychol.* **84**, 608–618 (2003).

5. Shi, J., Mo, X. & Sun, Z. [Content validity index in scale development]. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* **37**, 152–155 (2012).

6. Taherdoost, H. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *Test Valid. Quest. Res. August 10 2016* (2016).

7. Moreno, J. R., Flick, J. & Georges, A. Machine learning band gaps from the electron density. *Phys. Rev. Mater.* **5**, 083802 (2021).

8. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 1–10 (2019).

9. Howell, O., Wenping, C., Marsland III, R. & Mehta, P. Machine Learning as Ecology. *ArXiv190800868 Cond-Mat Stat* (2019).

10. Just, M. A., Cherkassky, V. L., Buchweitz, A., Keller, T. A. & Mitchell, T. M. Identifying Autism from Neural Representations of Social Interactions: Neurocognitive Markers of Autism. *PLOS ONE* **9**, e113879 (2014).

11. Just, M. A. *et al.* Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat. Hum. Behav.* **1**, 911–919 (2017).

12. Serdyukov, P. Machine Learning Powered A/B Testing. in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* 1365–1365 (2017).

13. Maulud, D. & Abdulazeez, A. M. A Review on Linear Regression Comprehensive in Machine Learning. *J. Appl. Sci. Technol. Trends* **1**, 140–147 (2020).

14. Michie, D., Spiegelhalter, D. J. & Taylor, C. C. Machine learning, neural and statistical classification. (1994).

15. Fenigstein, A. & Vanable, P. A. Paranoia and self-consciousness. *J. Pers. Soc. Psychol.* **62**, 129–138 (1992).

16. Tellegen, A. *et al.* MMPI-2 Restructured Clinical (RC) scales: Development, validation, and interpretation. (2003).

17. Freeman, D. & Garety, P. A. *Paranoia: The psychology of persecutory delusions*. (Psychology Press, 2004).

18. Moutoussis, M., Bentall, R. P., El-Deredy, W. & Dayan, P. Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cognit. Neuropsychiatry* **16**, 422–447 (2011).

19. Rossi-Goldthorpe, R. A., Leong, Y. C., Leptourgos, P. & Corlett, P. R. Paranoia, self-deception and overconfidence. *PLoS Comput. Biol.* **17**, e1009453 (2021).

20. Gervais, R. O., Ben-Porath, Y. S., Wygant, D. B. & Green, P. Development and validation of a Response Bias Scale (RBS) for the MMPI-2. *Assessment* **14**, 196–208 (2007).

21. Torday, J. S. & Baluška, F. Why control an experiment? *EMBO Rep.* **20**, e49110 (2019).

22. Ghosh, R., Gilda, J. E. & Gomes, A. V. The necessity of and strategies for improving confidence in the accuracy of western blots. *Expert Rev. Proteomics* **11**, 549–560 (2014).

23. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

24. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).

25. Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. How Many Trees in a Random Forest? in *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition* 154–168 (Springer-Verlag, 2012). doi:10.1007/978-3-642-31537-4_13.

26. Duda, M., Haber, N., Daniels, J. & Wall, D. P. Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Transl. Psychiatry* **7**, e1133 (2017).

Figures

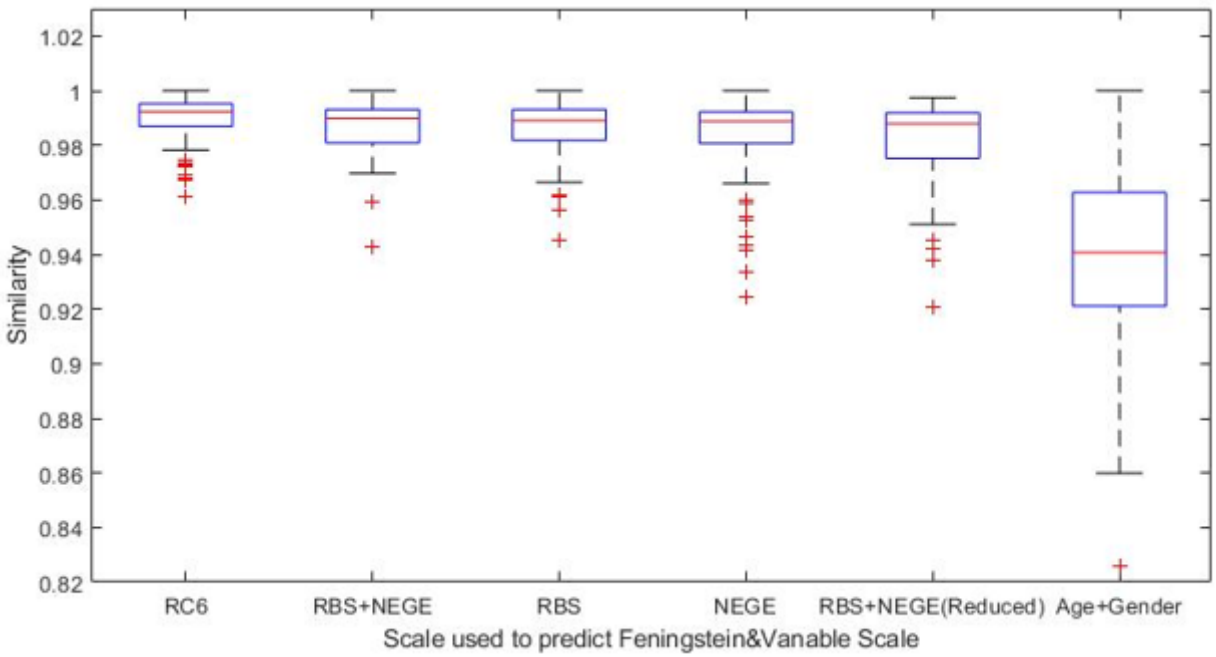

Figure 1

Mean similarity between true and predicted matrices for prediction from RC6, RBS & NEGE (Combined), RBS, NEGE, RBS + NEGE (Reduced), and Age+Gender (Negative control)

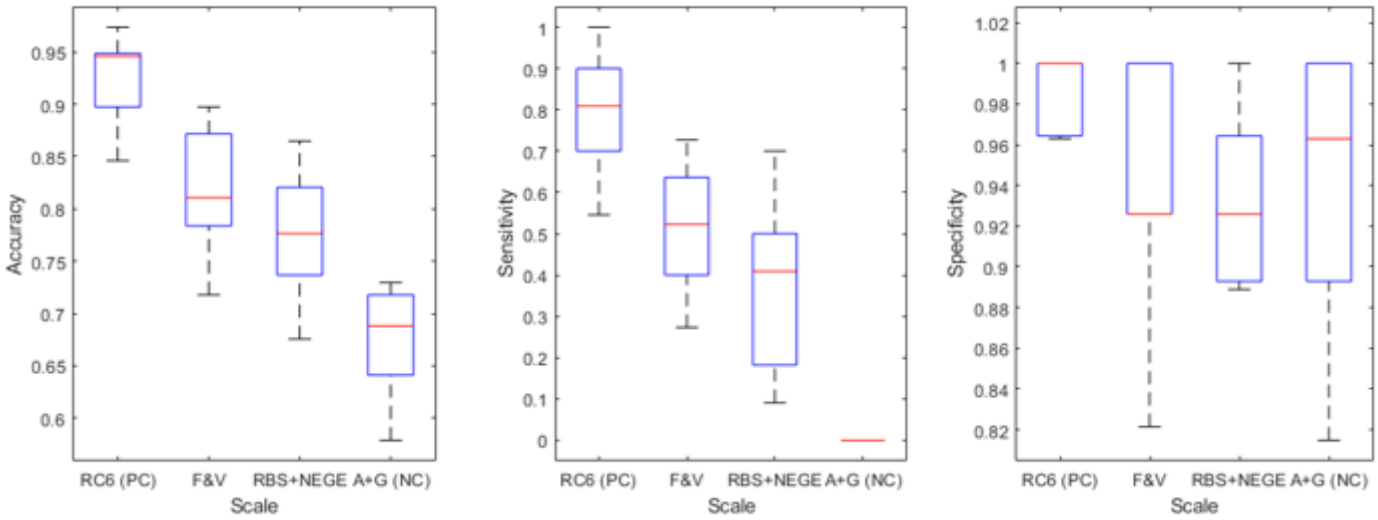

Figure 2

Results of detection metrics in 10-fold cross-validation. [RC6 (PC): Restructured Clinical 6 (Positive Control); F&V: Fenigstein & Vanable; A+G (NC): Age+Gender (Negative Control). Left: Accuracy; Middle: Sensitivity; Right: Specificity]

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- trognon2022supplementaryFile1.csv