Evaluating the Predictive Power of Tasks and Items in IQ Tests

Joshua Blickle (joshuablickle@gmail.com)
Klinikum Chemnitz gGmbH, Germany

Sara Todorovikj (sara.todorovikj@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

Intelligence tests are used in various scenarios in order to assess individuals' cognitive abilities. As these tests are typically resource-intensive and quite lengthy, we propose a predictive analysis paradigm with the aim of most effectively predicting IQ scores and thus shortening and optimising tests by identifying the most predictive test components. Using the Berlin Intelligence Structure Test for Adolescents (BIS-HB) as an example, we apply machine learning models and successfully predict IQ scores at the individual level. In addition, we identify non-significant and potentially redundant tasks and items and exclude them from the analyses, while maintaining the same satisfactory predictive results. A new direction of research in this area will allow not only the inductive optimisation of intelligence tests, but also the improvement of knowledge and understanding of intelligence in general.

Keywords: Intelligence, Predictive Analysis, Test-Optimisation, Prediction of the Individual, BIS-HB

Introduction

Intelligence encompasses the essence of our intellectual and cognitive abilities. In a broader sense, it can be understood as an individual's ability to "learn, reason and solve problems" (Plomin & Von Stumm, 2018, p. 148) and is a significant predictor of many important outcomes in life, including educational and occupational success (Deary, Strand, Smith, & Fernandes, 2007; Gottfredson, 1997), health, longevity (Deary, Hill, & Gale, 2021) and general well-being (Pesta, 2022). However, because the number and nature of testable intellectual abilities subsumed under the term "intelligence" are disputed, many theories vie for primacy, such as the Theory of primary mental abilities (Thurstone, 1938), or the Cattell-Horn-Carroll Model (McGrew, 2005), resulting in the lack of a universally accepted definition.

Intelligence is measured via a variety of theory-dependant and appropriately constructed tasks and items that cover the predefined range of cognitive abilities, forming a test. Due to the previously mentioned theoretical variations, the type of assessment, the abilities required, and the general task design vary considerably from test to test, with some emphasising school and culture-based content, such as mathematics and vocabulary tasks (e.g., Stanford-Binet Intelligence Scales; Grob, Gygi, & Hagmann von Arx, 2019), and others using exclusively abstract content, such as figural matrices (e.g., Raven's Progressive Matrices; Raven, 1940). Thousands of working hours and large sums of money and expertise are invested into the creation of intelligence tests, making them rare and at times expensive. More importantly, the length of testing, up to 2 hours for traditional assessments, coupled with potential biases (Nenty & Dinero, 1981) and participant fatigue (Ackerman & Kanfer, 2009), raises questions about the fairness and practicality of traditional intelligence tests.

In this paper, we focus on the "Berlin Intelligence Structure Test for Adolescents: Diagnosis of High Abilities and Giftedness" (BIS-HB; Jäger et al., 2006), based on Jäger's (1982, 1984) Berlin Intelligence Structure (BIS) model and its predictive power concerning the individual as well as possible methods of improvement to combat the aforementioned cons to intelligence testing inductively, largely free from theoretical assumptions. To do so, we explore the potential for identifying the optimal combination of test components. So, rather than relying on extensive assessments and theoretical arguments, we ask whether a more concise set of tasks and items could achieve comparable accuracy within acceptable margins of error. With that sentiment in mind, we are interested in taking the first steps towards assessing the predictive power of individual tasks and items that make up intelligence tests, utilizing simple machine learning (ML) approaches, leading to our first research question:

[RQ1] To which extent can we predict an individual's IQ score using machine learning models?

Moreover, by focusing on possible reductions in test length, we will take advantage of the ability of ML approaches to uncover novel patterns from complex data. By learning more about the relationships between tasks and items in an intelligence test and by recognizing the most predictive ones, we can subsequently identify the smallest set of tasks that predicts an individual's IQ score within comparable error margins. Accordingly, we pose our second research question:

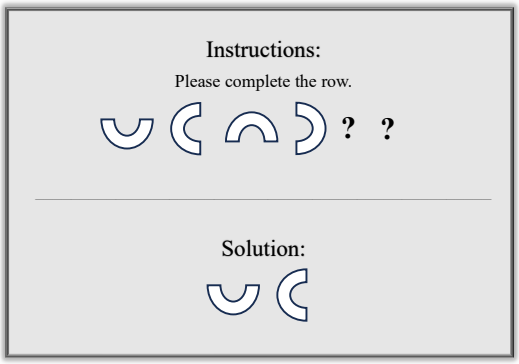
[RQ2] Which BIS-HB tasks and items are the most influential predictors of an individual's IQ score?

Our paper is structured as follows - in the following section we present necessary theoretical background regarding the BIS-HB intelligence test, followed by a brief description of the used data and the applied ML approaches. Afterwards, we present the results of our analysis and conclude with a discussion of our outcomes and findings.

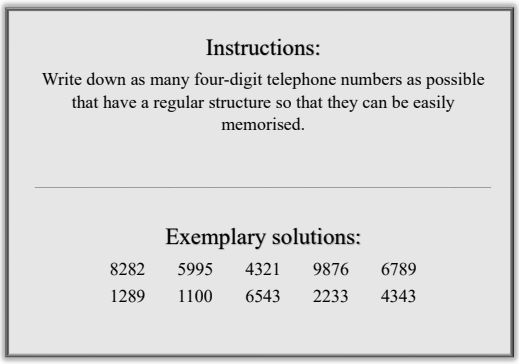
BIS-HB

The BIS-HB's wide measurement range (up to IQ=145) distinguishes it from other comparable instruments, avoiding

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*



(a) *Figural reasoning*: Individuals must identify the rule, 90 degree clockwise rotation, used to construct a sequence of drawings in order to determine the last items.



(b) *Numerical creativity*: Individuals are prompted to construct easily memorised x-digit numbers, applying as many different rules as possible.

Figure 1: Illustrative representations of a selection of BIS-HB items (Reasoning and Creativity). Individuals provide their answers by writing or drawing the solutions by hand. *Note:* These are not original items and figures as they would appear in the BIS-HB test, but only an analogy created by the authors for illustrative, exemplary purposes only.

"ceiling effects", whereby test takers consistently achieve the highest possible scores and therefore the difference in their potential is not accurately reflected in their IQ scores, a common problem when testing gifted (IQ>130) individuals (Preckel, 2003). In addition, the BIS-HB aims to assess a wide range of predefined abilities, all of which are thought to reflect aspects of *g*, a general form of intelligence.

Most intelligence tests are based on theoretical assumptions that are reviewed and revised before the test is finally normed and subsequently published. The standard review process hereby involves the development of a battery of similar tasks and items that are assumed to assess the same constructs. A task consists of a number of items which all adhere to the same theoretical principles and therefore possess a comparable structure. Furthermore, items are the smallest possible unit of a test.

The BIS-HB (Jäger et al., 2006) distinguishes between the four operations: *reasoning* (*r*), *creativity* (*c*), *memory* (*m*), and *speed* (*s*). They can be assessed via the three content formats: *figural* (*f*), *numerical* (*n*), and *verbal* (*v*) skill. The nomenclature is adopted from Breit, Scherrer, Blickle, and Preckel (2023), all terms are translations from the original German. Multiplying four operations with three types of content results in 12 abilities, accessed via 45 tasks, comprised of 278 items in total. According to its definition by Jäger's (1982, 1984) BIS, all items possess the characteristics required to measure intelligence. The individual intelligence score, as a result of a completed BIS-HB, is expressed in form of an IQ value ($\mu = 100, SD = 15$; Frenzel & Nett, 2008). A short form of the BIS-HB exists and takes ca. 55 instead of the usual 140 minutes to complete, but it consists *solely* of reasoning items (Vogl, Vogl, & Preckel, 2015), abandoning other intelligence facets and therefore the BIS model's core idea of intelligence as a multifaceted construct, distributed across a variety of operations and content domains.

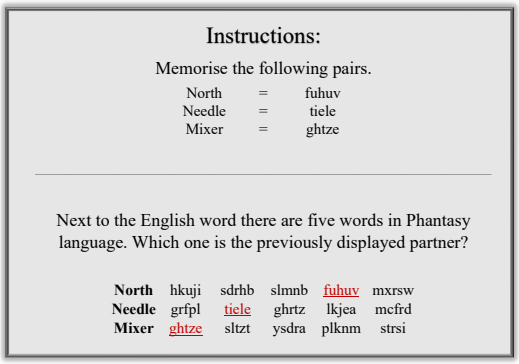
Content Formats

The four operations, detailed below, are examined using *figural*, *numerical* and *verbal* contents. Figural content requires test-takers to visually imagine spatial elements. Both numerical and verbal content utilise test-takers capabilities in dealing with complex symbol-systems (numbers or written language), which entails the use of advanced skills such as inference and the application of rules (Jäger et al., 2006).

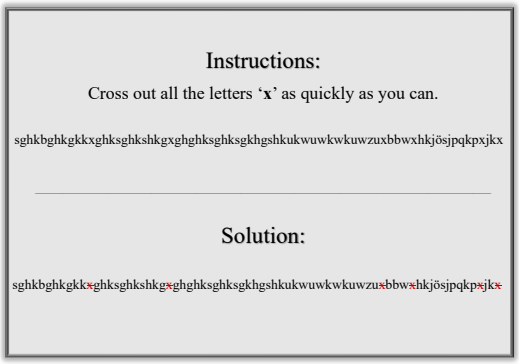
Operations

Reasoning Reasoning equates to drawing conclusions or inferences from information or more broadly speaking to how people solve problems and make decisions (Lakin & Kell, 2020; Leighton, 2004). In the BIS-HB, reasoning describes the ability to process complex information in tasks that cannot be solved immediately but require drawing on information, establishing a variety of relationships and applying formal logic accurately and appropriately (Jäger et al., 2006). Fig. 1a illustrates a figural reasoning item that requires individuals to derive a rule describing the spatial transformation of an object and draw a conclusion. In total, there are 15 reasoning tasks, 5 per content format.

Creativity The underlying processes of generating novel ideas or concepts are captured by creativity. This operation is measured via tasks that require the active generation of ideas rather than, e.g., simple identification, recognition, or selection of material. Therefore, test takers are instructed to provide a variety of responses under comparatively unstructured conditions. According to Runco (1986), this assesses divergent thinking – the ability to generate numerous and different ideas as e.g. responses to questions with no obvious, singular answer. Fig. 1b presents a numerical creativity item challenging individuals to generate as many different numeric sequences as possible, following a specific rule. In total, there are 12 creativity tasks, 4 per content format.



(a) *Verbal memory*: Given German (English) - Phantasy word pairs, individuals need to memorise them and then correctly recognise the Phantasy words.



(b) *Figural speed*: Given a sequence of random letters, individuals need to detect and cross out a specific letter as fast as possible.

Figure 2: Illustrative representations of a selection of BIS-HB items (Memory and Speed). Individuals provide their answers by writing or drawing the solutions by hand. *Note:* These are not original items and figures as they would appear in the BIS-HB test, but only an analogy created by the authors for illustrative, exemplary purposes only.

Memory This operation includes the active memorisation and short-term recognition or reproduction of various types of contents. The BIS-HB defines it as a measure of short-term acquisition and retention performance (Jäger et al., 2006). It should be noted that longer-term retention processes, i.e., the activation of long-term memory structures through the retrieval of crystalline knowledge, are not captured by the BIS-HB. Typical memory tasks require the retention of pairs of stimuli, the free recall of previously viewed content, or the reproduction of visual figures. In Fig. 2a an example of a verbal memory item is illustrated. First, individuals are required to memorize a pair of words, one being in German (English in the example) and the other one in an imaginary fantasy language. After the time has elapsed, they are asked to recall the correct fantasy word, out of a set of 5, corresponding to the German word. In total, there are 9 memory tasks, 3 per content format.

Speed Cognitive processing speed is typically captured via tasks that require rapid information processing on easy task material – due to the easy task-difficulty, adherence to the given time limits is essential. Typical speed-tasks require test-takers to locate and mark certain stimuli within the test material, to complete a pattern, or to make pairwise assignments. As the time taken to solve items is predetermined for all of the items in the BIS-HB, not just those capturing the operation speed, it is unlikely that the measures of the other factors (operations and content-domains) are completely independent of the variance explained by processing speed. Such a dependence has already been demonstrated for the BIS reasoning tasks (Wilhelm & Schulze, 2002). Figure 2b shows an example of a figural speed item. Individuals are given a sequence of letters and asked to detect the positions of a specific letter and cross out all occurrences as fast as possible. In total, there are 9 speed tasks, 3 per content format.

Methods

Data

The data used in the following analyses is the result of two BIS-HB surveys (LUMI Project; Breit, Scherrer, & Preckel, 2020-2021). The time frame between the two surveys spans 6 months. The tested subjects, in both samples, are pupils attending four high schools in Rhineland-Palatinate, Germany. A unique feature is the inclusion of children enrolled in special classes for the 'gifted'.

Sample 1. 424 individuals were tested, including 138 gifted children. 87 are excluded due to being too young or too old to obtain norm-referenced IQ scores and 34 are excluded due to missing data. That leads to a total of $N = 303$ subjects with a mean age of $M = 13.66$ years.

Sample 2. 387 individuals were tested, including 128 gifted children. 66 are excluded due to their age and 22 due to missing data. Ultimately, this sample is comprised of $N = 299$ subjects with a mean age of $M = 13.70$ years. 76.82% of the subjects from the second sample were retested.

Each subject was assessed via all 45 BIS-HB tasks, comprising a total of 278 items. The number of items per task varies between 5, (or 1 in cases where task and item were synonymous) and 22. The items of 23 tasks possess interval scaled responses, while the remainder are dichotomously coded in the data as correct or incorrect. The task features provide the overall score for all items in the respective task. Additionally, we acknowledge that Breit et al. (2023) found a large variance between the two samples. Due to that and the fact that over 75% of the subjects are retested, we decided to treat the samples separately.

Analysis

To address our research questions, we define a predictive task utilising models applied to data describing individual performance in tasks or, when appropriate, items, to predict IQ

Table 1: Predictive performance of models given all tasks.

S	Model	R^2	RMSE	MAE
1	Linear Regression	.96	2.51	2.07
	Random Forest	.88	5.56	4.27
2	Linear Regression	.97	2.51	2.03
	Random Forest	.91	5.41	4.19

Annotation: S – Sample; RMSE – Root Mean Squared Error; MAE - Mean Absolute Error

scores. We use two machine learning methods, linear regression and random forest, to assess the importance and predictive power of individual tasks and items in the BIS-HB, allowing for comparative analysis of their performance.

Linear Regression is a statistical method used to establish the relationship between one or more independent variables/predictors and a dependent variable. Giving a concise representation of the connection between the variables, it facilitates predictions of the dependent variable given new data points. We fit the linear regression model by minimising the sum of squared distances between the true and the predicted values. Due to the dichotomous nature of some of the items, we only applied linear regression to the tasks.

Random Forest is a robust machine learning model, an ensemble of decision or regression trees. For our task particularly, we use regression trees, generated by recursively splitting the data based on feature values while minimising the error of the target variable. By using randomly selected subsets of the data, we generate a number of regression trees, which store dataset features and splitting criteria in their nodes. The final result is the average predictions of all trees (Burger, 2018; Rebala, Ravi, & Churiwala, 2019). The randomness introduces a degree of diversity within the ensemble, making random forests less likely to overfit the data, while being able to capture complex relationships. The most valuable aspect of random forests for our task is the fact that they have an inherent feature selection process, when determining the best split at every node, which can help identify the most important predictors in a model (Horning, 2010). In order to evaluate the importance of the variables we examine the node purity, which is a measure of how much the model error changes when a particular variable is randomly shuffled or permuted (Tohry, Chelgani, Matin, & Noormohammadi, 2020). With it we obtain a quantification of the increase in model error should a particular variable have no information. No norm values exist, as it is a purely relative measure of the individual variable importance, specific to each random forest model (Dewi, 2019; Kuhn, 2008a).

Feature Importance *Recursive Feature Elimination* (RFE) is a popular predictor selection technique that works by iteratively using machine learning algorithms to identify and remove the least important predictors. This is achieved by fitting the selected machine learning algorithm, ranking all the predictors based on their importance, discarding the least important one, and re-fitting the model with the new, reduced number of predictors. This process is repeated until a specified number of predictors is obtained. Possible hyperparameters to explore are the number of predictors and the choice of algorithm used to aid feature selection (Brownlee, 2020). We apply RFE to determine the most prominent predictors among the tasks and items while using random forest. In order to evaluate the performance of each feature subset, we performed k-fold cross validation, with $k = 10$ – a technique used to assess the generalisability of models by splitting the dataset in k subsets and repeatedly validating a model trained on $k-1$ subsets on the remaining one, ensuring that each subset has been used as both training and validation data among all iterations.

Predicting the Individual In the case of predictive models, we aim to evaluate their predictive performance of an *individual's* IQ score. Following approaches in predictive modeling of deductive reasoning (e.g., Todorovikj & Ragni, 2021; Riesterer, Brand, & Ragni, 2020), we evaluate the models' predictive abilities using a *leave-one-out cross validation* (LOOCV) approach. With LOOCV, we fit a model with known data to all participants except the one, whose score is to be predicted based on the scores/correctness of their task and item answers. The same process is repeated for every participant in the data set. LOOCV has several advantages over other validation approaches, including a more robust estimate of model performance because each observation is given the opportunity to represent the entire test data set, resulting in lower bias and overall variability, providing a reliable estimate of how well a particular model generalises.

The software used for all of the analyses featured in this paper is R-Studio, version 2023.03.0, Build 386 "Cherry Blossom" (RStudioTeam, 2023). The basis to R-Studio is R, version 4.2.3 "Shortstop Beagle", released on the 15th of March 2023 (RCoreTeam, 2023). Additionally, the following R-packages were used: AICcmodavg (Mazerolle, 2023), broom (Robinson, Hayes, & Couch, 2023), car (Fox & Weisberg, 2019), caret (Kuhn, 2008b), corrr (Kuhn, Jackson, & Cimentada, 2022), dplyr (Wickham, François, Henry, Müller, & Vaughan, 2023), doParallel (Daniel, 2022), effsize (Torchiano, 2020), haven (Wickham, Miller, & Smith, 2023), and multcomp (Hothorn, Bretz, & Westfall, 2008).

Results

In the following, we present the results of our analysis. As mentioned above, we apply the models to the two samples separately. For each sample and feature scenario we report the R^2 value – the coefficient of determination, a statistical metric quantifying the proportion of variance captured by the model. Additionally, we report two errors measurements: the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). Between the two, RMSE is a strict metric that penalises large errors more, making it useful in cases when data contains outliers. MAE, on the other hand provides a

Table 2: Predictive performance of models given a reduced number of tasks. The total number of tasks is 45. In the case of Linear Regression, the number of tasks was reduced based on their lack of statistical significance as predictors. For Random Forest the amount is reduced through Recursive Feature Elimination.

S	Model	R^2	RMSE	MAE	#Tasks
1	Linear Regression	.95	2.79	2.22	28
	Random Forest	.88	5.53	4.23	43
2	Linear Regression	.96	2.64	2.08	35
	Random Forest	.91	5.33	4.09	35

Annotation: S – Sample; RMSE – Root Mean Squared Error; MAE - Mean Absolute Error; #Tasks – Number of tasks

Table 3: Predictive performance of Random Forest given all items and a reduced number of items. The total number of items is 278. When reduced, the number of items was derived through Recursive Feature Elimination.

S	R^2	RMSE	MAE	#Items
1	.85	6.10	4.70	278 (all)
	.86	5.78	4.47	72
2	.86	6.13	4.71	278 (all)
	.88	5.85	4.53	99

Annotation: S – Sample; RMSE – Root Mean Squared Error; MAE - Mean Absolute Error; #Items – Number of items

more intuitive and interpretable value, which is especially valuable in our task. We report all three metrics to provide a comprehensive assessment of model performance.

The results of the models applied to the tasks are shown in Table 1. For both samples, linear regression captured a very high variance percentage (.96 and .97, respectively). Additionally, evaluating the MAE, we can see that the model is able to predict individual IQ scores with a low margin of error – approximately 2 IQ points. In the case of random forest, we notice a slight decline in performance, yet still a large portion of variance is accounted for (.88 and .91, respectively), with a margin of prediction error being approximately 4 IQ points.

As stated before, we did not perform RFE using linear regression, however, we examined the statistical significance of the tasks as predictors of the target variable – the IQ score. For Sample 1, we found that 17 out of 45 tasks failed to reach statistical significance ($p > .05$) and for Sample 2, 10 out of 45. There was an overlap of five non-significant tasks: four reasoning and one speed task. Overall, the ratio of non-significant tasks in Sample 1 is: 47% reasoning, 44% speed, 22% memory and 33% creativity, and in Sample 2: 27% reasoning, 33% speed, 11% memory and 17% creativity. Ultimately, we fitted a linear regression model using only the statistically significant tasks, the results are presented in Table 2. With 17 tasks less in Sample 1, the R^2 value remains almost unchanged, while the error metrics suffer a slight increase, however negligible. Similarly, for Sample 2, after removing 10 tasks, the performance remains practically identical.

In order to select the best predicting tasks for random forest, we performed RFE using 10-fold Cross-Validation. For Sample 1, we found that the optimal number of tasks is 43 out of 45 (R^2 = .81, RMSE = 5.36, MAE = 4.23) and for Sample 2, 35 out of 45 (R^2 = .85, RMSE = 5.29, MAE = 4.13). In Sample 1, one reasoning and one memory task are excluded. The same ones are excluded in Sample 2, where the overall ratio is: 27% reasoning, 11% speed, 33% memory and 17% creativity. Despite the similar percentage distribution among excluded tasks in Sample 2 with RFE to the non-significant predictors in linear regression, most concerned tasks are different. After determining the subset of best predictors, we used them to fit a random forest using LOOCV and the results can be seen in Table 2. If we compare the results with those obtained using a full set of tasks, we see that the proportion of variance explained remains the same, while the errors decrease slightly, although negligible. In fact, after reducing the number of tasks, the random forest is still able to perform as well as before.

By focusing on the items, the smallest unit in the BIS-HB test, we examine their performance exclusively using random forest analysis, as linear regression is not an appropriate analytical tool for this analysis due to the dichotomous nature of some items.

Table 3 reports the results. We notice a slightly worse performance than when the model is fit to the tasks, with an explained variance of .85 for Sample 1 and .86 for Sample 2 and an error margin of more than 4.5 IQ points. Similarly to before, we performed RFE in order to determine the best predicting items using 10-fold Cross-Validation. For Sample 1, the optimal number of items is 72 out of 278 (R^2 = .79, RMSE = 5.93, MAE = 4.43) and for Sample 2, 99 out of 278 (R^2 = .99, RMSE = 5.93, MAE = 4.45). Given the large amount of individual items, we will analyze the exclusion in detail based on which tasks they belong to. In Sample 1 there are 3 reasoning, 1 speed, 4 memory and 2 creativity that are *entirely* excluded and 12 reasoning and 3 memory tasks that have 2 or more items excluded. On the other hand, in Sample 2, there are 2 reasoning and 2 memory entirely excluded, 12 reasoning and 4 memory tasks with 2 or more items excluded and 1 reasoning task with 1 item excluded. Comparing these numbers and tasks to those excluded when performing RFE with random forest on tasks, we observe an overwhelmingly larger number of tasks that are affected when discriminating based on the items' relevance. With the exception of four tasks in Sample 2, all excluded tasks in the previously reported analysis are also encountered in the item-based exclusion. Between-sample comparison shows that the tasks whose items were excluded are the same for the rea-

soning tasks. Similarly for the memory tasks, with the exception of one. Finally, we evaluate the individual predictive performance of the substantially smaller item subsets using LOOCV and report the results in Table 3. Once again, we notice a slight improvement in the explained variance and error margin allowing us to conclude that a significant decrease of the items still allows for the same, satisfactory performance.

Discussion

Given the considerable time and resources required to create, prepare and administer intelligence tests, this paper proposes a novel approach to assessing the relevance of different tasks and items other than through theoretical and/or deductive approaches. Motivated by cognitive load, fatigue and loss of concentration, we applied machine learning methods to analyse potential opportunities to reduce the number of tasks and their smaller units, items. The main goals of our work were summarised in our two research questions. In **RQ1** we asked how well we can predict an individual's IQ. Using data from the BIS-HB IQ test, we tested the predictive performance of the linear regression and random forest models given previously obtained task scores of the individual. We achieved highly satisfactory results within a margin of error of 2 and 4 IQ points respectively, giving a positive answer to our question. Typically, random forests are rather robust and likely to outperform other models, however here we witnessed a better performance by the linear regression, though marginal. One plausible explanation lies in the relationships between the variables. While random forests have demonstrated their ability to deal with complex problems, a simple linear relationship may be better explained by linear regression. Another point is that random forest regressors lack the ability to extrapolate to values beyond the limited intervals they are exposed to during training - something a linear regression is certainly able to do. Additionally, we evaluated the performance of a random forest when fitted on all 278 items, observing only a minor increase in error. In **RQ2** we ask which BIS-HB tasks and items are the most predictive. To answer that, we a) identified and recomputed the linear regression models including only significant tasks, maintaining a similar performance. Similarly, b) for random forest, we performed RFE and determined a reduced subset of tasks and items, the latter being substantially smaller (from 278 to 72 for Sample 1 and 99 for Sample 2), all while the predictive performance remained unharmed, if not slightly improved. We identified tasks and items with limited impact, indicating the need for further analysis to understand their differences and the factors contributing to their relevance. We have previously recognised an existing short form of the BIS-HB test that consists *only* of reasoning tasks. Through our analysis, we show that the other facets need not be abandoned, as they still hold valuable predictive power. As even the 'downsized' models, i.e., those consisting of only the most predictive test components, still assessed intelligence using all operations across all content-domains we partially confirmed the theoretical concept of intelligence as defined by the BIS-HB. When analysing the results, a natural question to ask is how much of a deviation from the true value are we willing to accept and still be satisfied with the performance? Given that in our article we propose and take the first step towards this approach, setting such a cut-off value would not be sensible. We are acknowledging and presenting the potential trade-off between predictive accuracy and amount of tasks/items, whose worth is decided on a case-by-case basis.

In our exploration of the principles of artificial intelligence and machine learning, we can distinguish between deductive reasoning and inductive learning (Aggarwal, 2021). The deductive approach starts with a hypothesis based on theoretical assumptions, while the inductive approach derives patterns from observations. We suggest that both approaches should be used to optimise intelligence testing. While hypothesis and tasks derived from a strong theoretical background are undoubtedly valuable, it would be tremendously beneficial to take steps in a different direction which would facilitate not only making intelligence tests more accessible and minimize testing of same factors, but also enrich our knowledge regarding intelligence and its assessment. Naturally, a finding that a specific task does not have a high predictive power does not immediately negate the theoretical significance of the related component as an aspect of intelligence. By observing patterns in the data, even finding individual clusters and applying predictive models, we open the doors to new findings and understandings. Naturally, the methods that we presented, along with many other approaches can be applied to other tests as well and help us identify the importance of certain task features.

A limitation of the used dataset that is worth discussing is the lack of the *exact* responses that participants provided. For many of the items we only had dichotomous information on the answer correctness. If specific answers are provided, that would allow us to develop concrete, detailed models that incorporate solving-rules and facilitate error-analysis, which in turn would allow to even simulate the steps that individuals take. With that, the prediction task would develop further to predict the exact answer. Another point that should be acknowledged is the difference in ratios of non-significant tasks in Samples 1 and 2. The reason behind that likely lies in the general difference in results, even between the same individuals. However, this point is beyond the scope of our article, as it has been discussed appropriately in Breit et al. (2023).

In conclusion, our work has successfully demonstrated a predictive approach to intelligence tests. Not only did we predict individuals' IQ scores, but we also identified less predictive tasks and items whose exclusion did not significantly affect performance. In doing so, we motivate the development of a new predictive analysis paradigm in the field of intelligence testing, with the aim of optimising existing and future psychometric tests.

Acknowledgements

This project has been partially funded by a grant to MR in the DFG-projects 529624975 and 283135041. We extend a sincere gratitude to Prof. Dr. Franzis Preckel from the university of Trier for her collaboration and providing us with the necessary data (Breit et al., 2020-2021) for our analyses.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *12*(2), 163–181. Retrieved from https://doi.org/10.1037/a0015719 doi: 10.1037/a0015719

Aggarwal, C. (2021). *Artificial Intelligence: A Textbook*. Springer International Publishing.

Breit, M., Scherrer, V., Blickle, J., & Preckel, F. (2023). Students' intelligence test results after six and sixteen months of irregular schooling due to the COVID-19 pandemic. *PLOS ONE*, *18*(3), 1–25. doi: 10.1371/journal.pone.0281779

Breit, M., Scherrer, V., & Preckel, F. (2020-2021). *Längsschnittliche Untersuchung von Motivation und Intelligenz im Schulalter - LUMI.* FB I - Psychologie, AE Hochbegabtenforschung und -förderung. Trier: Universität Trier.

Brownlee, J. (2020). *Recursive feature elimination (rfe) for feature selection in python.* MachineLearningMastery. Retrieved from https://machinelearningmastery.com/rfe-feature-selection-in-python/ (Retrieved May 6th, 2023)

Burger, S. V. (2018). *Introduction to machine learning with r: Rigorous mathematical analysis*. O'Reilly Media, Inc.

Daniel, F. (2022). doParallel: Foreach Parallel Adaptor for the 'parallel' Package [Computer software manual]. (R package version 1.0.17. Retrieved March 12th, 2023, from https://CRAN.R-project.org/package=doParallel)

Deary, I. J., Hill, W. D., & Gale, C. R. (2021). Intelligence, health and death. *Nature Human Behaviour*, *5*(4), 416–430. doi: 10.1038/s41562-021-01078-9

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13-21. doi: https://doi.org/10.1016/j.intell.2006.02.001

Dewi, C. (2019). Random forest and support vector machine on features selection for regression analysis. *International journal of innovative computing, information & control: IJICIC*, *15*, 2027–2037.

Fox, J., & Weisberg, S. (2019). *Car: An R-Companion to Applied Regression (3rd ed.)*. Sage.

Frenzel, A. C., & Nett, U. (2008). Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB) from A. O. Jäger et al., 2006. *Diagnostica*, *54*(4), 221–225.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79–132. doi: 10.1016/S0160-2896(97)90014-3

Grob, A., Gygi, J. T., & Hagmann von Arx, P. (2019). *Stanford-Binet Intelligence Scales, Fifth edition (SB5)–German adaptation*. Hogrefe Verlag GmbH & Co. KG.

Horning, N. (2010). Random forests: An algorithm for image classification and generation of continuous fields data sets. In *International conference on geoinformatics for spatial infrastructure development in Earth and allied sciences* (Vol. 911).

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. doi: 10.1002/bimj.200810425

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzlestungen: Experimentell kontrollierte Weiterenwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, *28*(3), 195–225.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, *35*, 21–35.

Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., & Süß, H.-M. (2006). *BIS-HB: Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik – Manual*. Hogrefe Verlag GmbH & Co. KG.

Kuhn, M. (2008a). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1 – 26. doi: 10.18637/jss.v028.i05

Kuhn, M. (2008b). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. doi: 10.18637/jss.v028.i05

Kuhn, M., Jackson, S., & Cimentada, J. (2022). corrr: Correlations in R [Computer software manual]. (R package version 0.4.4. Retrieved April 4th, 2023, from https://CRAN.R-project.org/package=corrr)

Lakin, J. M., & Kell, H. J. (2020). Intelligence and reasoning. In R. J. Sternberg (Ed.), *The cambridge handbook of intelligence* (pp. 528–552). Cambridge University Press. doi: 10.1017/9781108770422.023

Leighton, J. P. (2004). The assessment of logical reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 291–312). Cambridge University Press.

Mazerolle, M. J. (2023). AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c) [Computer software manual]. (R package version 2.3.2. Retrieved April 4th, 2023, from https://cran.r-project.org/package=AICcmodavg)

McGrew, K. S. (2005). The cattell-horn-carroll theory of cognitive abilities: past, present, and future. , 136–181.

Nenty, H. J., & Dinero, T. E. (1981). A cross-cultural analysis of the fairness of the cattell culture fair intelligence test using the rasch model. *Applied Psychological Measurement*, *5*(3), 355–368. Retrieved from https://doi.org/10.1177/014662168100500309 doi:

10.1177/014662168100500309

Pesta, B. J. (2022). Updated IQ and well-being scores for the 50 u.s. states. *Journal of Intelligence*, *10*(1), 15. doi: 10.3390/jintelligence10010015

Plomin, R., & Von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics*, *19*(3), 148–159. doi: 10.1038/nrg.2017.104

Preckel, F. (2003). *Diagnostik intellektueller Hochbegabung: Testentwicklung zur Erfassung der fluiden Intelligenz*. Hogrefe Verlag GmbH & Co. KG.

Raven, J. C. (1940). *Progressive matrices*. H. K. Lewis.

RCoreTeam. (2023). R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.R-project.org/ (Retrieved March 15th, 2023)

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An introduction to machine learning*. Springer Interational Publishing. doi: 10.1007/978-3-030-15729-6

Riesterer, N., Brand, D., & Ragni, M. (2020). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*(3), 960–974. doi: 10.1111/tops.12501

Robinson, D., Hayes, A., & Couch, S. (2023). broom: Convert Statistical Objects into Tidy Tibbles [Computer software manual]. (R package version 1.0.4. Retrieved April 4th, 2023, from https://CRAN.R-project.org/package=broom)

RStudioTeam. (2023). RStudio: Integrated Development for R [Computer software manual]. Boston, MA. (Retrieved February 8th, 2023, from http://www.rstudio.com/)

Runco, M. A. (1986). Divergent thinking and creative performance in gifted and nongifted children. *Educational and Psychological measurement*, *46*(2), 375–384.

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

Todorovikj, S., & Ragni, M. (2021). How good can an individual's conclusion endorsement be predicted? In T. C. Stewart (Ed.), *Proceedings of 19th international conference on cognitive modeling* (pp. 275–281).

Tohry, A., Chelgani, S. C., Matin, S. S., & Noormohammadi, M. (2020). Power-draw prediction by random forest based on operating parameters for an industrial ball mill. *Advanced Powder Technology*, *31*(3), 967–972. doi: https://doi.org/10.1016/j.apt.2019.12.012

Torchiano, M. (2020). effsize: Efficient Effect Size Computation [Computer software manual]. (R package version 0.8.1. Retrieved April 7th, 2023, from https://CRAN.R-project.org/package=effsize)

Vogl, E., Vogl, K., & Preckel, F. (2015). *Rezension BIS-HB – Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik [Review BIS-HB - Berlin intelligence structure test for adolescents: Diagnosis of high abilities and giftedness]*. Karg-Stiftung. Retrieved from https://www.fachportal-hochbegabung.de/oid/85004/ (Retrieved March 20th, 2023)

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation [Computer software manual]. (R package version 1.1.2. Retrieved April 7th, 2023, from https://CRAN.R-project.org/package=dplyr)

Wickham, H., Miller, E., & Smith, D. (2023). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files [Computer software manual]. (R package version 2.5.2. Retrieved March 3rd, 2023, from https://CRAN.R-project.org/package=haven)

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence*, *30*(6). doi: https://doi.org/10.1016/s0160- 2896(02)00086-7