Československá psychologie 2025 / ročník LXIX / číslo 3
DOI: 10.51561/cspsych.69.3.176

BRIEF ADAPTIVE NONVERBAL TEST (BANT), THE DIAGNOSTIC TOOL EVALUATING NONVERBAL COGNITIVE SKILLS IN PRESCHOOL-AGED CHILDREN

SIMONA PEKÁRKOVÁ[1], MARTINA ŠVANDOVÁ[1], MATĚJ SEIFERT[2], HYNEK CÍGLER[3], JIŘI ŠTIPL[4,,5], FILIP SMOLÍK[6]

[1] *Department of Psychology, Faculty of Arts, Charles University, Prague*
[2] *National Pedagogical Institute of the Czech Republic, Prague*
[3] *Department of Psychology, Faculty of Social Studies, Masaryk University, Brno*
[4] *Educational and Psychological Counselling Center STEP, Kladno*
[5] *Graduate School of Psychology, University of Amsterdam, Research Master's in Psychology*
[6] *Institute of Psychology, Czech Academy of Sciences, Prague*

ABSTRACT

Objectives. This study presents the development and validation of the diagnostic tool Brief Adaptive Nonverbal Test (BANT). The tool is intended to evaluate nonverbal cognitive skills in preschool-aged children, with a special focus on school readiness use.

Method. BANT was developed in a two-stage process involving item calibration and subsequent validation on independent samples of preschool children. In the first calibration phase, test items were administered on tablets and their IRT parameters were estimated, with the objective of developing an adaptive test. In the validation phase, the resulting adaptive test was administered to an independent sample of preschool children in order to determine the test's reliability, validity and investigate the test's internal consistency. In addition to the BANT, children in the validation sample were administered various tasks included in a preschool diagnostic application Myška, which were used to obtain validity estimates.

Sample and settings. Data collection for the development of the method followed two steps. Both took place in kindergartens and aimed at children aged 4 to 7 years. Calibration sample consisted of N_1=302 children from 17 different kindergartens. Validation sample with prepared computer adaptive testing algorithm comprised N_2=507 children (girls 49.4%, boys 50.6%), examined in 37 kindergartens from 11 regions. All children were examined in their preschools, using tablets to administer the BANT and the battery of preschool diagnostic tasks Myška.

Analyses and results. Item calibration, check of CAT mechanism functioning, different item functioning analysis, reliability analyses, ability estimates and norms development has been performed using appropriate packages in R (R Core team, 2023). Testing with the CAT version of the test is significantly less time consuming in comparison with the fixed version of the test (m=2.78 mins compared to m=7.62 mins with fixed test; items administered m=X items compared to 59 items with fixed test). Reliability over 0,80 has been achieved approx. for -3 to +1 theta estimates, with a peak over 0,95 at approximately -1,75 theta estimate. Correlations with other tests in Myška application support the interpretation of the test results, according to which test aims at visual processing and nonverbal reasoning using visual material. Overall, BANT appears to be a practical tool for evaluating cognitive development and school readiness in the general cognitive domain in pre-school children. Its fast adaptive administration using tablets is a practical utilisation in pedagogical and clinical practices and is especially beneficial to children with language barriers.

Limitations. In subsequent development, the results from the task should be validated against a dedicated nonverbal cognition tool such as SON-R.

key words:
computerized test,
preschool age,
non-verbal thinking,
computer adaptive testing (CAT),
school readiness

Submitted: 9. 5. 2025; S. P., Charles University, Ovocný trh 560/5, 110 01 Prague, e-mail: simona.pekarkova@ff.cuni.cz

The authors wish to express appreciation to the schools, teachers, and children who participated in this research. We would also like to acknowledge the helpful contributions of Adam Tápal with item calibration and creation of CAT algorithm.

INTRODUCTION

The development of new tools for psychological assessment in early childhood is important for many reasons. Such instruments assist in identifying developmental disorders in children, detecting irregularities in development, and help in establishing children's potential. Of particular importance is the age range before and around the school entry, i. e. approximately 4 to 7 years, when the tools can identify the likelihood of successful participation in formal schooling and identify children who need support in this process.

The developmental assessment tools aimed at this purpose and age group often take one of two perspectives: either they focus on specific components of school readiness or examine general cognitive skills. The perspective of school readiness typically involves assessments of some or many specific areas of knowledge, cognitive skills, or executive abilities, such as pre-reading skills, graphomotor skills, early quantitative skills, language skills, as well as social and self-control skills essential for successful adaptation to the school environment. The perspective of cognitive skills is based on general cognitive assessments, focusing on the core cognitive ability that underlies many academic skills and identifying children who need special support or, on the other hand, are exceptionally gifted.

School readiness methods

There are currently only a few efficient but fairly comprehensive tools for school readiness assessment in the Czech Republic, and their use is not systematic. These tools are usually used for school readiness assessment in counselling services or preschools (e.g., Bednářová & Šmardová, 2015; Jirásek, 1992; Poláková & Vlčková, 2013). In primary schools and preschools in the Czech Republic, similar tools for educational assessments are used infrequently and not comprehensively due to time constraints– more often for identifying risks in the development of reading and writing skills (e.g. Švancarová & Kucharská, 2001; Sindelarova, 2008). There are no consensual guidelines about which of these tools should be used, and under which conditions. In part, this is because the number and variety of available instruments is limited, and tools for various purposes are missing, calling for approximate and ad-hoc solutions. At the same time, the lack of consensus reflects the broader situation in the field.

Internationally, the concept of school readiness does not have universal interpretation. Sometimes it includes multiple skills that support learning to read, write, and do arithmetic, which corresponds to the approach in Czechia. Alternatively, one factor serves as the main indicator of school readiness, whether this comprises general cognitive skills, self-regulatory skills (attention and motivation), or child behavior (e.g., Dockett & Perry, 2013; Vidmar et al., 2017). Many educational professionals focus predominantly on the child's social-emotional competencies and self-help skills (Cuskelly & Detering, 2003; Kagan, 2003; Lin et al., 2003). Although there is still no consensus on the elements and dimensions of school readiness (Snow, 2006), there is a growing emphasis on the importance of education in Early Childhood Education and Care (ECEC) settings where school readiness is an integral part (e.g., Niklas et al., 2017; Tayler, 2016). Effective and reliable measurement of performance domains is crucial for setting development goals (Józsa et al., 2022).

While the methods focusing on specific school readiness skills provide valuable information, the focus on a single main dimension of general skills provides background information that is often the key to interpreting findings from more specific methods. In an environment where only a limited set of tools is available, it is meaningful to provide such general measures first.

The perspective of general cognitive skills

There are many instruments available for testing children's general cognitive skills, especially nonverbal skills, in the international environment. Such tasks try to avoid too strong dependency on a particular cultural, linguistic or educational background, which makes them well-suited to evaluate children's general cognitive preparedness for the educational process. They typically focus visual-spatial processing (Gv) (Flanagan & Dixon, 2014), which is a crucial broad ability within the Catell-Horn-Carroll theory (McGrew, 2009), and evaluate individuals' ability to manipulate, organize, and recognize visual and spatial information, which is fundamental for tasks requiring pattern recognition and understanding spatial relationships (Uttal et al., 2013). Examples of such tasks include NNAT, Raven's progressive matrices or SON-R (Tellegen & Laros, 1998).

In the Czech Republic, the instruments available for early general cognitive ability include Raven's Colored Progressive Matrices (Raven, 1998) and SON-R (Tellegen & Laros, 1998). While the former instrument is still in use internationally, the Czech adaptation and norms are old and only available in the paper-based form. The latter is a complex battery that also relies on physical materials and includes six subtests with administration time around 50 minutes. A modern tool for brief assessment is thus missing. To fill this gap, one way would be to adapt internationally available tools such as NNAT (Naglieri, 2018) or develop original methods. Relying on original methods can be of advantage when previous data and materials on the performance of Czech children are available as resources for addressing this task. This was the case in the present project, and the existing tasks formed the core for the development of the Brief Adaptive Nonverbal Test BANT (Pekárková et al., 2025).

Brief Adaptive Nonverbal Test

The development of BANT was started as an extension of the preschool educational assessment battery iSophi (Pekárková & Švandová, 2019) and its tablet-based variant Myška (Pekárková & Švandová, 2024). This battery is aimed at helping preschool teachers with mapping the profile of children's skills. It is still undergoing psychometric evaluation, validation and refinement but it is not intended to provide a fully psychometrically validated tool, rather a guidance for preschool teachers' practice. While evaluating the components of the battery, the visuospatial items showed particularly good measurement properties. Their similarity to some of the existing tools for the assessment of intellectual abilities, and the need to create a shorter and unidimensional screener for school readiness, led to the decision to develop a separate instrument based on these types of items. The instrument was intended for tablet presentation because it makes the administration easier and more consistent across children, contributing to standard procedure. At the same time, the tablet format made it possible to create an adaptive task that would present items around the estimated level of the proband's skills, shortening the procedure and decreasing the load for the child.

The items used for the test are based on the "odd one out" format where the probands view an array of 5–6 figures that have some systematic relations, except for one. The probands are required to identify this odd figure. This task is relatively easy to explain and understand, does not require complex responses and fits well into the tablet-administered form. Another appealing feature of the task format is that it is well suited for adaptive testing, which is another novel aspect of the BANT instrument.

Adaptive testing in children

Computer-adaptive testing is an approach to measurement that avoids presenting the same set of items to all individuals undergoing an assessment. Rather, the items are selected automatically so that they are close to the estimated level of the trait in a given individual. In cognitive testing, this means that it avoids presenting items that are too easy or too difficult for the given individual and rather focuses on more precise determination of the skill level. In practice, this means that the continuation of the test depends on the previous responses, so that participants who initially respond correctly are given items that are equally or more difficult, and vice versa. The algorithm for choosing the upcoming items and calculating the estimated value of the trait is based on the item-response theory and requires relatively complex calculations after responding to each item. It is thus impossible or impractical to use the approach in tests administered using traditional paper-and-pencil methods, but the wide availability of tablets and laptop computers means that the adaptive approach can be used widely in practice.

Various tools for school readiness or early cognitive assessment that include aspects of adaptive testing are available internationally but not for the Czech population. For example, PIPS-BLA (Performance Indicators in Primary Schools–Baseline Assessment) is an interactive test focused on reading, mathematics, vocabulary, and phonological awareness used in five jurisdictions: England, Scotland, New Zealand and one State and one Territory in Australia (Wildy & Styles, 2008). EDI (Early Development Instrument), developed in Canada, is nowadays used in a wide range of countries include Australia, the United States, Scotland, Jamaica, Ireland, Peru, Chile, Mexico, Brazil, Mozambique, Vietnam, Hong Kong, China, Sweden, Estonia, Kosovo, Kyrgyzstan, Indonesia, Philippines, South Korea, and Jordan (Janus & Offord, 2007). The EDI assesses children's school readiness across five domains–physical, social, emotional, language and cognitive development, communication, and general knowledge. Its electronic version, e-EDI, is already being used in Canada and shows a strong correlation between cognitive skills and performance in reading, writing, and mathematics (Kokkalia et al., 2019).

In Hungary, the standardized DIFFER test, which assesses preschool skills such as sound discrimination, relational reasoning, and basic numerical skills, is common. The digital version of this test also includes tasks on inductive thinking, focusing on the ability to identify similarities and differences between objects (Csapó et al., 2014).

In the Czech Republic, the adaptive approach is used in some tests focused on adolescents or adults, e. g. Vídeňský maticový test (2002), e-psycholog (Cígler et al., 2024), but not in children's assessment. BANT is thus the first method that utilizes this approach in the Czech Republic for early cognitive and school-related assessment. The approach is highly useful for the assessment of children because it makes possible to shorten the task and also helps avoid the boredom associated with presenting too many too easy tasks, and the frustration due to presenting tasks that are too challenging. The decision to use the odd-one-out type of items also makes the task well suited for the adaptive format because the order, number and composition of items for each proband are free to vary, not requiring different or additional instructions.

METHOD

Preparation, item development, the application

As described above, the development started with existing components of the iSophi battery that showed good psychometric properties and allowed easy creation of ad-

ditional items. The visuo-spatial perception items in iSophi included different task formats and various types of visual stimuli. The most consistent results were achieved using letter-like shapes of varying complexity, so this type was chosen for creating additional items. The format of "odd-one-out" was chosen to unify the type of the tasks; probands in this format are asked to search for one in a series of visual stimuli that does not fit the others. The format has been used in some influential and widely used instruments, such as the CMMS - Columbia Mental Maturity Scale (Burgemeister et al., 1972),

The task was developed in two steps. In the first, calibration step, children were administered a set of 58 potential items in a non-adaptive way, with two practice items besides this bank. This step was used to obtain IRT estimates of item properties that served as the input for the adaptive algorithm. In the subsequent, "validation" step, the full adaptive algorithm was tested, and we examined the resulting scores, their measurement properties, and the numbers of items and time needed to achieve adequate standard errors of measurement.

The task is implemented in a tablet computer application that was developed for the diagnostic system Myška (Pekárková & Švandová, 2024). The task was presented as one of the subtasks in the set of Myška, which also means that some data from the diagnostic system are available for validation. Due to the background in the development of Myška, the tool benefited from a skilled team of programmers and educational specialists who designed the original as well as the novel BANT items. The application includes pictures created by a professional artist, and a professional voice-over for instructions. In the beginning of the game, children are introduced to a mouse character who acts as a guide. This character encourages the child to collect droplets in a watering can, which are later used to water flowers. A droplet is awarded to the child for each task completed, regardless of its success. The opening scene also serves to screen for motor skills by tracking finger movement across the tablet screen. After the opening scene, training items are presented, first as an animation that explains the task to the child and shows how to solve it. Two active training items follow, in which the child responds to the task on her or his own and receives feedback. After incorrect training responses, an animation with the correct solution and an explanation is shown. The training items familiarize the child with the format of the tasks and the required response behaviors.

They also allow the examiner to observe if the child has problems dealing with the tablet or otherwise using the technology or the task; in that case, the examiner may decide to postpone the administration. After training, the child proceeds to the main task where he or she responds independently and receives uninformative but generally positive feedback after each item. After successfully completing the items assigned by the adaptive algorithm, children are invited to use the collected droplets to water a selected flower. This provides a natural conclusion to the game. After that, a performance report is immediately generated. The reports are in a form of simple charts supplemented with essential information for the professional test administration. The final design of the charts and the exact information content in the reports is currently being finalized, based on the results of the norming study.

Participants

In both stages of data collection, convenience sampling has been used. All parents in selected classrooms have been offered the possibility to participate, and children of those who gave consent were included in the study. The sample may thus not be con-

sidered fully representative, but we believe it reflects well the demographic diversity of the Czech population, including preschools from municipalities of different sizes, and from different regions of the country. In both steps, only children whose parents provided a signed confirmation of consent participated.

In the calibration step of the development, the difficulty and discrimination parameters of the items were obtained based on a sample of 302 children aged 4.0–6;11, with a mean age of 5 years; 9 months and standard deviation of 5,22 months. The data collection in this group was conducted in October and November 2021. These children from 17 different kindergartens were examined with the non-adaptive version of the test.

The data collection for the second step took place from November 2022 to March 2023. We had to exclude some of the 641 observations obtained: in 1,8% of cases the testing was interrupted by the respondent and completed on another day, and 19% of administrations were terminated early due to an unforeseen technical error in the application.

The validation group thus consisted of 507 children who completed the task, 49% (250) girls. The average age was 5 years and 8 months (SD=6 months), and the data were collected in 37 kindergartens from 11 of the 14 Czech administrative regions ("kraj").

Procedure

The data collection was conducted using the tablet application Myška (Pekárková & Švandová, 2024), with preschool teachers supervising the testing process. The teachers received a short training on how to administer the task. In the second, validation step of the development, data collection was conducted along with 18 other tasks included in the development version of the Myška application. The order of the tests was fixed, with the BANT being third tested. Both preceding tests were short, and the test protocol also allowed for a pause before start of testing, if the child was tired or conditions in the preschool prevented continuation. Parents were provided with simple results for all administered tests, based on item success rates and speed of response, and were also informed of the provisional nature of these results.

Item bank

At the beginning of the method development, 85 items were available, two of which were designed as practice items. After a qualitative pilot with a smaller group of children, 25 items were dropped from further collection. Thus, an item bank of two practice items and 59 test items was prepared for the initial assessment of item parameters. These are always items with a choice of one correct response, scored 0 or 1. The instruction for the test items is: „Dotykem obrať kartičku, která nepatří mezi ostatní. Až budeš mít hotovo, zazvoň na zvonek dole v rohu." („Tap to flip over the card that doesn't belong with the others. When you're done, ring the bell in the bottom corner.")

In 30 cases, the items consist of 5 figures (one correct and four distractors), and in the remaining cases, there were 6 options. The position of the correct option in each item is fixed and varies randomly between items. For the items, a uniform time of no more than 20 seconds has been set for solving the problem. After solving the item, the participant either moves on to the next item by pressing a button (the procedure is subject to practice and occasional repeated instructions in the test), or the application moves on to the next item by itself after the time limit has expired. Children can also skip an item by pressing the same button.

Calibration of items

A test with 59 items was administered to the calibration sample (N_1). The order of the items was fixed in the calibration; all items of the test were administered to the children. Using the mirt package (Chalmers, 2012), we fitted several Item Response logistic test models, namely the 2-parameter(2PL) and 3-parameter(3PL) model. In addition, we also fitted the 3PL model, where the lower asymptote was fixed to the initial value, i.e.,1/5 or 1/6 depending on the number of response options (2PLl, two-parameter model with lower asymptote). As the research sample was relatively small, the 3PL and 2PLl parameters estimates were unstable, and they did not fit the data better compared to the initial 2PL model. Therefore, we selected the 2PL model for further analyses, and its parameters (item difficulty and discrimination) were used for the adaptive testing procedure. Other analyses performed were related to the test reliability, information function of the items and test, reliability for sub-age groups and test invariance across age groups. The time taken to administer the whole test and the relationship between the time taken and the difficulty of the item were also assessed.

Simulation of adaptive testing

Based on latent trait variance, we estimated the RMSE (root mean-square error) needed to achieve reliability of .8 or higher to be ≤ .45. We used this threshold for estimation error as the stopping rule for adaptive testing. We also used a graph to balance test length and estimated reliability based on this ending rule, see Figure 1. The other stopping rules were the minimum of 10 and maximum of 40 administered items.

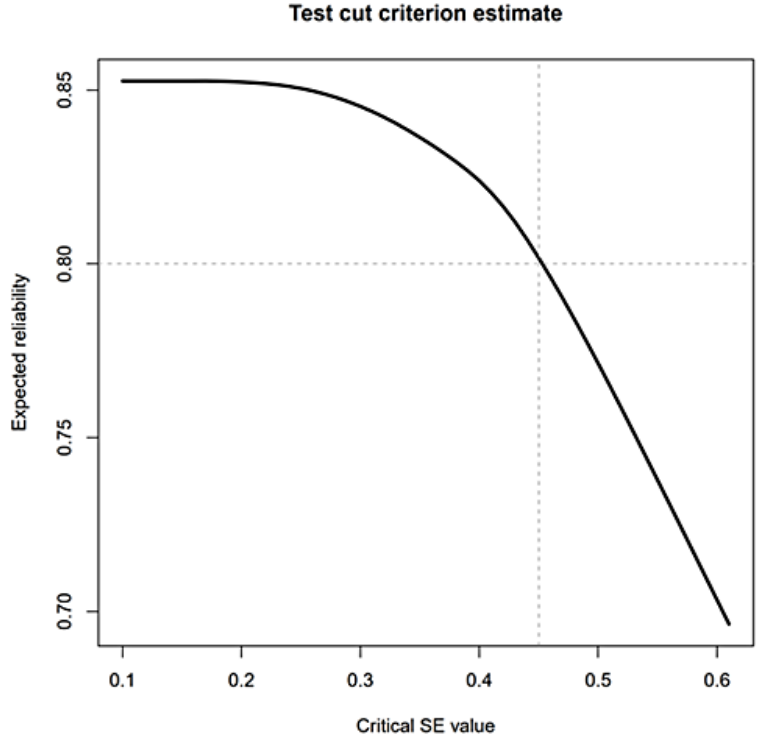

Figure 1 Test cut criterion estimate

Using these parameters, we ran simulation studies using the original data as if the test were administered adaptively, focusing on the number of administered items, achieved errors of estimation (and, subsequently, test reliability), and the estimation bias. Test scores (θ) were estimated using expected a-posteriori method (EAP).

Validation of the CAT mechanism and construction of the final version of the test

An adaptive algorithm was then programmed in the Myška application to calculate the parameter θ and the standard error of estimation after each iteration, using the expected a-posteriori (EAP) estimator. The baseline in testing is always the first item in the set. Next item administered is chosen from the remaining items in the item bank so that the information obtained for the estimated ability is maximized.

The completed adaptive test was then deployed along with other tasks included in the Myška application. For the BANT test, this testing provided a large dataset that allowed for a comprehensive assessment of test and item functioning, and for correcting any deficiencies in the application prior to further test deployment. The features that were examined included variability in test passes, consistency of behavior when respondents gave the same answers, assessment of the conditional reliability of the test (in relation to the child's ability level), the number of items solved in relation to the person's ability level, the time spent solving items, accuracy of sequential item-by-item estimations and flow of ability estimates in individual observations (checking for undesirable skips in the CAT algorithm), and assessment of differential item functioning (DIF) for age groups and for items administered in non-adaptive and adaptive testing conditions. Opportunities to optimize the performance of the application were sought. For statistical processing, the catR (Magis & Barrada, 2017; Magis & Raiche, 2012) and mirt (Chalmers, 2012) packages and, in general, the R system (R Core Team, 2023) was used. Subsequently, the final version of the test was compiled.

RESULTS

Item calibration

Item difficulties b were found to be in the interval <-4.03; 13.71> with a median of -0.58; item discrimination parameter a was found to be in the interval <0.10; 9.19> with a median of 1.43. The 2PL model fits the data significantly better than the Rasch model (χ^2=484.3, *df*=57, *p*<0.001, *RMSEA*=0.022, $_{90\%}CI$ [0.017;0.027], *SRMSR*=0.059, *TLI*=0.986, *CFI*=0.986). The fit indices for the three-parameter model are ambiguous, and the difference in fit is not statistically significant when compared to the 2PL model ($\Delta\chi^2$=69.30, *df*=58, *p*=0.147). For this reason, analyses were continued with the 2PL model. The time to solve the full test was at least 3.54 minutes, at most 13.18 minutes, with the average *M*=7.62 (*SD*=1.64). One item was dropped from the analysis due to lack of variability. Two items were not included in part of analyses due to extreme and potentially misleading estimates of their parameters but were retained in the item bank for future testing.

Adaptive testing simulation

A total of 308 simulated passes were generated, tied to the estimated ability level of the probands from the fixed-variant testing. In the simulated passes, the average number of items solved was 15.74 (*SD*=9.16), and the median number of items solved was 11. The simulation-estimated savings in testing time for individuals was *M* = 4.58 min

(*SD*=1.33; *min*=1.32, *max*=8.71), while the average test-solving time was estimated to be M=2.0 min (SD=1.33), which would mean in average only 26% of time needed for fixed test.

Adaptivity of the test in the validation sample

There were 3 items in the bank that the adaptive algorithm has never selected in the validation sample; these had a low discrimination parameter or an extreme difficulty parameter. Only 13 of the 58 items have been administered to more than 50% of the respondents, on the other hand seven items (in addition to the three that were skipped completely) have been administered to less than 10 % of participants. There were 399 variations of unique test passages observed. Participants were presented 10 to 40 items in total, median *Md*=13, *M*=18.16, *SD*=9.58. The testing time was *M*=2.78 min, *SD*=1.45, *min* = 0.66, *max*=7.82. A total of 11 participants had administration time below one minute, all at a low or very low level of ability. The average testing time in the calibration phase was statistically significantly higher than in the adaptive version of the test ($p<0.001$). On average the adaptive version of the test used 36.48% of time needed for testing in the calibration sample.

The lowest number of items is administered to children with below-average levels of measured ability; in the case of standard scores below 90, a maximum of 11 items were administered, whereas in the case of children with above-average levels of ability (standard scores above 120), the maximum set of 40 iterations was always administered.

Differential item functioning (DIF) in the fixed and adaptive versions of the test

Item functioning was assessed for the original calibration sample, in the validation sample, and after the two sets of data were merged into one. Because some items were only administered rarely or only to a specific subset of participants in the validation step, the data to estimate DIF reliably were not available for all items, and thus we only report a subset where the estimation was possible. A scalar-invariant multiple group 2PL model was estimated. Then, both parameters (discrimination and difficulty) were released, and the new model was compared to the original scalar-invariant model using likelihood ratio test (LRT) and information criteria (*BIC*, Bayesian information criterion, and *SABIC*, sample-adjusted *BIC*). Finally, we visually compared empirical and model implied item characteristic curves.

A significant amount of DIF was observed for 10 items. After detailed examination, it was decided that for four items, the final parameter estimation would only be based on data from adaptive testing, and for one item, only data from the calibration step were retained. Four items were discarded entirely. One item was left in the item bank but marked to be recalibrated in the future. This led to an item bank of 54 items for which we had in total between 302 and 759 individual observations at the item level. This item bank has been used for final estimation of test parameters and construction of test norms.

Test validity

The fit of a one-factor Rasch model, a two-parameter logistic model, and a two-parameter logistic model with a fixed choice of the guessing parameter was validated over the full data set combining the calibration and validation testing data (*n*=809). The 2PL model shows the most favorable fit in terms of *SABIC*, *BIC*, and *LRT* test (*AIC* 22340.97, *SABIC* 22506.08; *BIC* 22849.04). The two-parameter model fit the

data significantly better than the Rasch model ($\Delta\chi^2$=520.3, df=53, p<0.001). We were not able to estimate and compare approximate fit indices based on M_2 statistics in this step, because the M_2 statistics was not defined when the analyses were performed for the missing data (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2006). Standardized discrimination parameter (factor loading) smaller than 0.3 occurred in a single item (0.258). For all other items, loadings were in the interval 0.311 to 0.939.

In cognitive developmental tests, one indicator of validity is that the performance increases with age. This was the case, with correlation between age and estimated ability level θ being r=0.209 with $_{95\%}CI$ [0.142; 0.274], p<0.001. The age invariance of the test was tested by comparing the younger and older half of the validation group of children. The goodness-of-fit indicators are most favorable for the scalar invariant model. Five items out of 54 retained in the item bank showed significant differences in functioning between the age groups; thus, the effect of age on the functioning of the test as a whole is minimal and the interpretation for different age groups may be identical.

Validity is further supported by relations to other tasks presented in the Myška application. A set of 18 additional tasks was presented in the validation sample. All these other tests were administered with fixed order of items, and were focused on visual and auditory perception, understanding time concepts, verbal reasoning, receptive language and quantitative skills. Children's latent abilities as measured by each task were estimated using 2PL IRT model, and we then conducted correlational analyses across tests (see Figure 2).

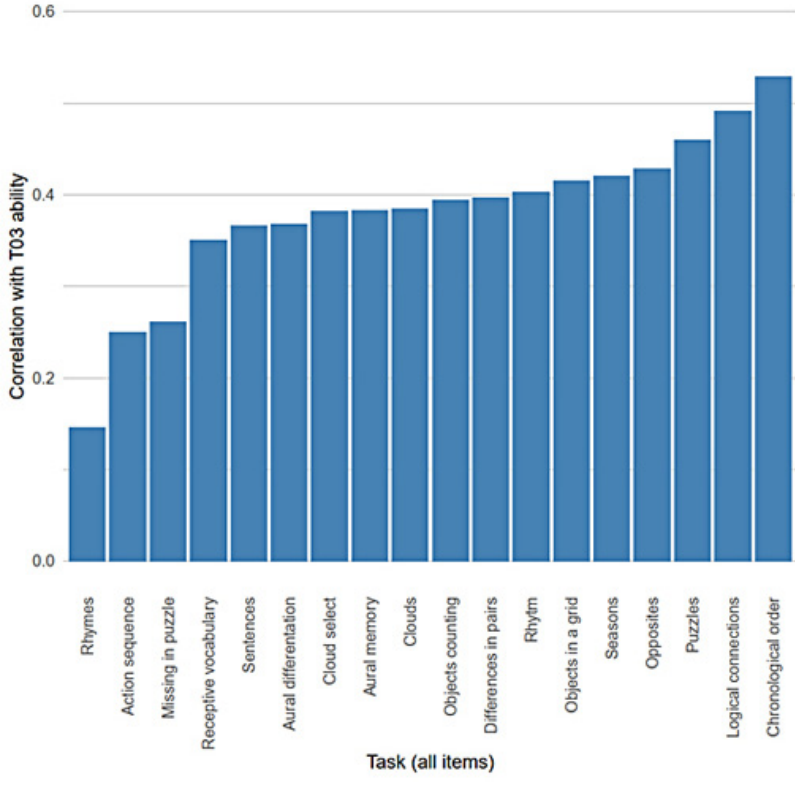

Figure 2 Correlation of BANT to other tests in Myška standardization sample

The strongest correlation (r>0.5) was observed to the Picture arrangement test where children arrange pictures in sequence to correctly capture a particular ongoing event. BANT also has correlations above r=0.4 with the subsequent six tests. The first of these tests is the *Seasons* task where 3–4 pictures are presented, and children must select the one that relates to a daily activity or an activity typical for a specific season. The question is played automatically after the picture set appears on the screen. In the next *Opposites task,* children are shown a stimulus card with an image. From three other pictures, children choose the one that represents the opposite meaning of the given image.

The third is the test *Logical Connections* which involves two rows of 3–5 pictures shown. The child identifies which pictures from each row logically belong together and connects them by drawing a line (on the tablet). In the fourth test *Composition from Pieces* three geometric shapes appear on the screen. Children are given two additional pieces and must determine which of the offered shapes would be formed by combining those two pieces. The next test *Objects in a Grid* includes searching for an object based on a specified location in a grid containing 6–12 compartments with various items. All these tasks require spatial imagination and arrangement, rely on perceptual material and are based mainly on nonverbal reasoning, which supports the validity of BANT as a measure of nonverbal, visuo-spatial reasoning. The sixth test correlated at the same level (above 0.4) was also observed for the Rhythm task which asked for identifying syllables in auditorily presented words by tapping a drum on the screen. The relation to visuo-spatial reasoning is not quite clear but it may be related to the sequences and counts.

In contrast, correlations below 0.3 were between the results of the *Differences in pairs* test where presenting 4 to 6 pairs of pictures and children identify and select the ones that are not identical, *Rhymes* test involving presenting the child with a series of word pairs as auditory stimuli, requiring the child to identify whether the words rhyme and *Action sequence* test which is based on a verbal instruction containing several commands based on which children locate specific objects within the picture, where it is acceptable given the different nature of the material. The lower correlation with the *Puzzle completion test* where images of objects composed of 4 to 9 pieces are shown, with one piece missing and children must find the correct missing piece among four offered cards and drag it into the puzzle. This test is likely due to the very low difficulty and resulting poor discrimination of this task, which is expected to be redesigned in the future. Overall, based on this result, the test is related to the ability to process organized and sequenced visuo-spatial stimuli, relate visual stimuli to each other and reason about their differences, and identify their significant and less significant features. This is in line with the intended main content of the BANT test.

Reliability of the test

Test reliability was assessed as conditional within the IRT 2PL model. Acceptable levels of local reliability of 0.8 and above were achieved for approximately <-3, 1> standard deviations from the mean of ability level θ (see Figure 3). Reliability is highest at about two standard deviations below the θ mean. Overall reliability of the test is $r_{xx'}$=0.853. For reliability of BANT in four age groups see Table 1.

Adaptive testing increases the estimation errors in comparison with fixed calibration test, but this goes hand in hand with a significant reduction in testing time. In the case of the calibration sample, the estimation error averaged *M*=0.295, *SD*=0.089; *min*=0.164, *max*=0.604. In the case of the validation sample, the estimation error averaged *M*=0.414, *SD*=0.073, *min*=0.212, *max*=0.614. Highest estimation errors were observed for children with above average ability as the test is rather easier.

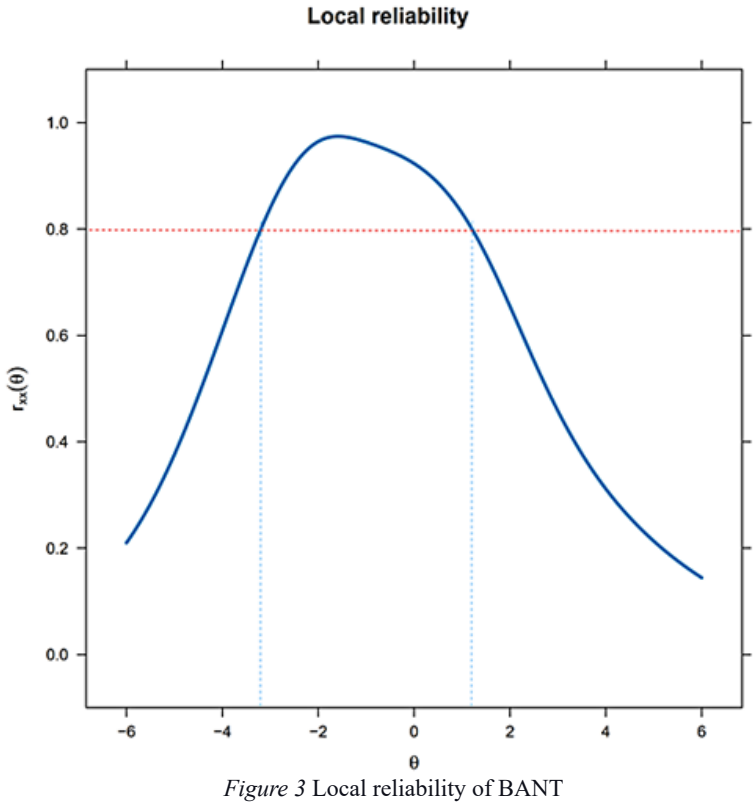

Figure 3 Local reliability of BANT

Table 1 Reliability of BANT for three age groups

Age group	Reliability
4;0–4;11	0.902
5;0–5;11	0.860
6;0–6;11	0.825

Effects of item time limitations

BANT was developed with a fixed maximum of the time available per item. Empirical findings from standardization showed that item solving time varied significantly with item difficulty as well as with the level of measured ability of participants. After evaluating times on the items, we extended the time limits by 2 to 10 seconds on 35 items so that the speed aspect was minimized and the time did not expire during item solving even with participants, who needed more time. For nineteen items time could be the same or even two seconds shorter. Generally, the influence of this step should be marginal, because only less than 5% of participants used possible maximum time on more than 3 items in their testing, 38% of participants have not run out of time on any item and another 48% of children have run out of time on one or two items. Generally, it would be advisable to study the time spent on items during the pilot study. Also, the final standardization should include an analysis of DIF by gender for the time aspect of responding.

DISCUSSION

We described the development of a novel adaptive instrument for assessing nonverbal cognitive skills in children before school entry. We show the rationale, the process of test development, the properties of the adaptive algorithm, and information about test reliability and validity. While the construction and psychometric evaluation of the method is complete, some additional steps are necessary or desirable to deploy the method in practice. However, the current data confirms the feasibility and utility of the method. The use of tablet computers and the adoption of the adaptive approach results in a method that has high potential for efficient, fast and highly useful screening for cognitive skills in practice. The interactive digital format is attractive for children, thus increasing their engagement and motivation to complete tasks. Adaptive testing provides measurement of individual abilities that does not burden the child with unnecessary tasks while maintaining chosen levels of precision. This approach helps overcome some of the barriers associated with traditional testing, such as test anxiety and the time burden.

The electronic mode of delivery using BANT also increases practical reliability of the testing process, avoiding a complicated scoring process and calculations. This can contribute to rapid and targeted educational interventions. Additionally, the test can enhance digital competencies among selected groups of professionals.

BANT has a good potential as a tool for assessment in preschool and around school entry, especially in the context of identifying developmental risks in children. Due to its nonverbal nature, BANT is useful for assessing children who may face language barriers, such as children with limited knowledge of the local language or children with certain types of communication and speech disorders. BANT is also suitable for children from various cultural and ethnic backgrounds because it minimizes cultural bias by not requiring verbal responses and relying on a relatively simple verbal instruction.

The evaluation of the calibration and validation step in the development shows good indicators of reliability especially in the range between two standard deviations below the mean and one standard deviation above the mean. This is in line with the intended purpose of the instrument, i. e. screening for children's readiness for school education in the domain of general nonverbal cognition. For this purpose, it is important to differentiate children who are around the mean and considerably below the mean, but it is not critical to accurately assess in the range of intellectual impairment (more than 2 standard deviations below the mean) or among the children who perform well or in the gifted range (1 SD above the mean or above).

Before the instrument is fully deployed in practice, the following steps need to be completed:

1) Preparing dynamic age norms, i. e. calculating the expected distribution of the ability estimates θ as a function of age and providing percentile ranking of the child among children of the same age. In addition to age norms, criterion percentile norms will be created using the age of 6;0 as the reference. This age is the lowest age for children to enter school, and the minimum level of cognitive skills should thus be evaluated against this benchmark.

2) Concurrent criterion validity study proof using established measurement tools. Data collection for this study has already begun using the non-verbal components of the WISCV (Wechsler, 2024). The anticipated sample size is 60 children.

3) Developing the summary reports for communicating the results. Two types of reports are under development. First one is designed for the non-specialist users in the educational system, i. e. preschool teachers and special educators, while the other

is intended for psychologists, who have a background in psychological measurement theory. The reports will be generated based on the results of the BANT test and are structured to meet the specific needs of professionals and reflect the level of their expertise in data interpretation. The goal is for the outputs to effectively aid in further developing interventions focused on supporting child development and minimizing delays in initiating necessary support measures.

Limitations

Currently, BANT focuses on a selected dimension of nonverbal thinking, mainly on visual-spatial processing, and avoids examining other cognitive skills. The use of tablet computers is required for test administration, which may be a limitation in some educational or clinical settings. Considering risks such as potential attention distraction and data analysis complexity is important.

Given that BANT focuses on visual-spatial processing, this test may disproportionately discriminate against children with various degrees of visual impairments or specific needs in visual perception. For children with visual impairments, the accuracy and fairness of the test can currently be quite limited but further work is needed to establish its validity in such populations.

The test is designed and calibrated specifically for children aged 4–7 and conducted within the Czech Republic's cultural and educational context. Although the method relies on nonverbal material, it may nevertheless put emphasis on tasks that are common for Czech children and perhaps less common elsewhere: validation in additional environments is thus desirable.

The characteristics associated with the number of administered items indicate that it will be appropriate to add additional items of greater difficulty to calibrate and subsequently expand the available item bank to enhance the correct adaptive function in children with above-average levels of measured ability. Subsequently, consideration can be given to possible future restandardization once the test is completed.

REFERENCES

Bednářová, J., & Šmardová, V. (2015). *Diagnostika dítěte předškolního věku: Co by dítě mělo umět ve věku od 3 do 6 let* (2. vydání). Edika.

Burgemeister, B. B., Hollander Blum, L., & Lorge, I. (1972). *Columbia Mental Maturity Scale (CMMS).* The Psychological Corporation.

Cai, L. & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66,* 245–276.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06.

Cígler, H., Jabůrek, M., & Ťápal, A. (2024). ePsycholog: Psychometrický manuál (1.0.0). ePsycholog online. https://epsychology.cz/files/53/technicky_manual.pdf

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, *106*(3), 639.

Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences, 19,* 209–218. https://doi.org/10.1016/j.lindif.2009.01.002

Cuskelly, M., & Detering, N. (2003). Teacher and student teacher perspectives of school readiness. *Australian Journal of Early Childhood*, *28*(2), 39–46.

Dockett, S., & Perry, B. (2013). Trends and tensions: Australian and international research about starting school. *International Journal of Early Years Education*, *21*(2-3), pp. 163–177.

Flanagan, D. P., & Dixon, S. G. (2014). The Cattell-Horn-Carroll theory of cognitive abilities. In C. R. Reynolds, K. J. Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education* (pp. 368–376). John Wiley & Sons. https://doi.org/10.1002/9781118660584.ese0431

Formann, A. K. (2002). Vídeňský maticový test (Nové přepracované vydání). Testcentrum.

Janus, M., & Offord, D. R. (2007). Monitoring the development of all children: The Early Development Instrument. *Early Education and Development*, *18*(3), 599–624. https://doi.org/10.1080/10409280701610796

Han, K. C. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions, 15,* 7.

Jirásek, J. (1992). *Orientační test školní zralosti: Příručka*. Psychodiagnostika.

Kagan, S. L. (2003). Children's readiness for school: Issues in assessment. *International Journal of Early Childhood, 35*(1-2), p. 114.

Kokkalia, G., Economou, A., & Roussos, P. (2019). School readiness from kindergarten to primary school. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(11). https://doi.org/10.3991/ijet.v14i11.10090

Lin, H. L., Lawrence, F. R., & Gorrell, J. (2003). Kindergarten teacher's views of children's readiness for school. *Early Childhood Research Quarterly, 18*(2), 225–237.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the Package catR. *Journal of Statistical Software, Code Snippets, 76*(1), 1–19. https://doi.org/10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R Package catR. *Journal of Statistical Software*, *48*(8), 1–31. https://doi.org/10.18637/jss.v048.i08.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713–732.

McGrew, K. S. (2005). The Cattell–Horn–Carroll Theory of Cognitive Abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, Tests, and Issues* (pp. 136-182). The Guilford Press.

Mhic Mhathúna, M., Ring, E., Moloney, M., Hayes, N., Breathnach, D., Stafford, P., Carswell, D., Keegan, S., Kelleher, C., McCafferty, D., O'Keeffe, A., Leavy, A., Madden, R., & Ozonyia, M. (2016). *An examination of concepts of school readiness among parents and educators in Ireland*. Department of Children and Youth Affairs. www.dcya.ie

Naglieri, J. A. (2018). *Naglieri Nonverbal Ability Test–Third Edition (NNAT3)*. Pearson.

Niklas, F., Tayler, C., & Cohrssen, C. (2017). What is 'school readiness' and how are smooth transitions to school supported? In T. Bentley, & G. C. Savage (Eds.), *Educating Australia: Challenges for the decade ahead* (pp. 117–132). Melbourne University Press.

Pekárková, S., Smolík, F., Švandová, M., & Štipl, J. (2025). *Diagnostická aplikace Brief Adaptive Nonverbal Test (výzkumný prototyp)*. iSophi Education.

Pekárková, S. & Švandová, M. (2019). *iSophi Pedagogická diagnostika 5-7*. iSophi Education.

Pekárková, S., & Švandová, M. (2024). *Herně-diagnostická aplikace Myška (výzkumný prototyp)*. iSophi Education.

Poláková, S., & Vlčková, H. (2013). *Test mapující připravenost pro* školu *(MaTeRS)*. Národní ústav pro vzdělávání.

R Core Team (2023). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/

Raven, J. C. (1998). *Raven's Coloured Progressive Matrices*. Oxford Psychologists Press.

Sabol, T. J., & Pianta, R. C. (2017). The state of young children in the United States. In E. Votruba-Drzal, & E. Dearing (Eds.), *The Wiley Handbook of Early Childhood Development Programs, Practices, and Policies* (pp. 1–17) John Wiley and Sons, Inc.

Sindelarová, B. (2008). *Deficity dílčích funkcí*. Nakladatelství D+H.

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development*, *17*(1), 7–41. https://doi.org/10.1207/s15566935eed1701_2

Švancarová, D., & Kucharská, A. (2001). *Test rizika poruch čtení a psaní pro rané školáky.* Scientia.

Tayler, C., Cloney, D., Adams, R., Ishimine, K., Thorpe, K., & Nguyen, T. K. C. (2016). Assessing the effectiveness of Australian early childhood education and care experiences: Study protocol. *BMC Public Health*, *16*, 352.

Tellegen, P. J., & Laros, J. A. (1998). *SON-R 2½-7: Nonverbal intelligence test*. Swets & Zeitlinger.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

Vidmar, M., Niklas, F., Schneider, W., & Hasselhorn, M. (2017). On-entry assessment of school competencies and academic achievement: A comparison between Slovenia and Germany. *European Journal of Psychology of Education*, 32(2), 311–331.

Wechsler, D. (2024). *Wechsler Intelligence Scale for Children* (5th ed.). Pearson.

Wildy, H., & Styles, I. (2008). Measuring what students entering school know and can do: PIPS Australia 2006–2007. *Australasian Journal of Early Childhood*, *33*(4), 43–52. https://doi.org/10.1177/183693910803300407