What the New Measure of Thinking in School Students Has to Offer to Contemporary Education

I. L. Uglanova, I. N. Pogozhina

Received in June 2021

Irina L. Uglanova, Junior Research Fellow, Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. Email: iuglanova@hse.ru (corresponding author)
Irina N. Pogozhina, Doctor of Sciences in Psychology, Associate Professor, Department of Psychology of Education and Pedagogics, Faculty of Psychology, Lomonosov Moscow State University. Address: Bld. 9, 11 Mokhovaya St, 125009 Moscow, Russian Federation. Email: pogozhina@mail.ru

Abstract

For the pedagogical principle of assigning comprehensible and adequate tasks to be implemented, allowance should be made for students' individual levels of logical reasoning, which requires diagnostic measures for objective and quick assessment. Today, the "clinical method" allows the most comprehensive assessment of logical thinking within the Piagetian framework. However, this diagnostic measure is extremely resource-consuming, hence unsuitable for large-scale testing. An overview of literature shows that the existing standardized diagnostic measures require a great number of highly-qualified experts to review the scores and prepare feedback for teachers, instructional designers, practicing psychologists and researchers.

The article describes design methodology of an instrument to evaluate levels of logical reasoning that will allow automated scoring without sacrificing score meaning, eventually facilitating and accelerating the diagnostic measurement procedure. Implementation of these principles is analyzed using the example of scenario-based tasks realized as computerized performance-based assessment in the form of stealth assessment of fifth- and seventh-grade pupils.

Keywords

automated scoring, critical thinking, diagnostic measurement, formal operations, logical reasoning, logical thinking, performance-based assessment, Piaget, psychometrics, scenario-based problems, stealth assessment.

For citing

Uglanova I. L., Pogozhina I. N. (2021) Chto mozhet predlozhit' novaya metodologiya otsenki myshleniya shkol'nikov sovremennomu obrazovaniyu [What the New Measure of Thinking in School Students Has to Offer to Contemporary Education]. *Voprosy obrazovaniya/Educational Studies Moscow*, no 4, pp. 8–34 https://doi.org/10.17323/1814-9545-2021-4-8-34

Assigning comprehensible and doable tasks is a fundamental pedagogical principle. "All the subjects that are to be learned should be arranged so as to suit the age of the students, that nothing which is beyond their comprehension be given them to learn." [Comenius

1939:151]. Overall, the design of modern school curricula makes allowance for stages of mental development, so an "average" student encounters little difficulty in learning. At the same time, underachievement in school education remains quite a pressing issue. One of its sources is that cognitive development rates of individual students are insufficient to enable them to learn what they are supposed to learn at a particular grade. There is empirical evidence of the relationship between academic gains in various disciplines and the level of intellectual development [Malhotra 2020; Watkins, Lei, Canivez 2007]. Comprehension of mathematical and physical concepts as well as of social norms and rules has been shown to require a certain level of logical reasoning [Inhelder, Piaget 1958; Piaget, Inhelder 2003].

Practices of the most successful education systems demonstrate that the problem of low academic achievement in school can be solved by personalizing the learning process, i. e. acknowledging that "every child is special" [Vainikainen et al. 2015; Hautamäki, Thuneberg 2019; Hienonen 2020]. Personalized learning requires, among other things, taking account of the individual levels of logical reasoning, indicated by students' performance on Piagetian tasks [DeVries 1974; Goldschmid 1967; Lawson, Renner 1975; Lovell, Shields 1967; Lawson, Blake, Nordland 1974]. Because conventional IQ test scores do not provide enough information on the composition and structure of logical operations to adapt curricula to individual levels of development, the modern school needs diagnostic measures that are suitable for mass application and objective, i. e. independent from subjective interpretations [Avila de, Pulos 1979; Hathaway, Hathaway-Theunissen 1975; Kaufman 1972]. A new diagnostic measure would help understand the cognitive sources of low achievement in individual children and adolescents so as to build personalized educational trajectories with due regard to individual psychological characteristics.

The present study seeks to identify and describe the stages in design of an automated-scoring instrument to measure school students' levels of logical reasoning.

1. The Construct of Logical Thinking: Composition and Structure

Before describing the instrument design methodology, it would be natural to define the construct that will be measured.

Logical thinking is not something we have by default; objectivity increases over the four stages of development between birth and maturity: (1) sensorimotor period; (2) pre-operational thought; (3) concrete operations; and (4) formal operations. From stage to stage, regular cognitive patterns that capture associations between various characteristics of reality are organized into increasingly generalized structures (logical operations). At the level of behavior, we observe changes in the individual's perceptions of the objects in the world around, their properties, space, time, motion, causality, etc. [Piaget 1994a; 1994b; 1994c].

A great contribution to success in middle and high school is made by logical structures that emerge at the stage of formal operations:

(1) (1) Propositional logic: ability to engage in hypothetical reasoning, i. e. identify (reflect) consistent patterns and associations not only between real-world objects and their visual representations but also between propositions of different language systems. Includes the following propositional operations: if … then (implication), or (disjunction (either or both)), and (conjunction), not (negation), if and only if (equivalence), etc.;
(2) (2) Combinatorial logic: conditional combination of various objects in all possible configurations, e. g. when any six (seven, ten, … *n*) objects are systematically combined into sets of two, three, … etc. in all possible ways without repetitions and with rigorous control of the outcome;
(3) (3) Synthesis of the two kinds of reversibility into a single cognitive structure of four transformations INRC (*I* is identity, or direct operation; *N* is negation of change, or inverse operation; *R* is reciprocity, or reciprocal operation, allowance for the impact of mutually related factors; and *C* is negation of reciprocity, the inverse of the reciprocal). The adolescent gains the ability to analyze problematic situations using all the four transformations at the same time.

At the latter level, the adolescent mind develops new operational schemata necessary for successful learning in school disciplines:

- Proportions (commensurability, equality of two or more ratios);
- Mechanical and homeostatic equilibrium (equality between action and reaction);
- Relative motion (motion in relation to a fixed system);
- Probabilities (odds of an event occurring in certain circumstances);
- Ability to go beyond the observed data, which involves hypothesizing (what if …?), constructing a system of probable regular patterns, etc. [Piaget 1994a; 1994b; 1994c; Piaget, Inhelder 2003; Piaget 2008].

As we can see, logical thinking represents a sophisticated system of interrelated structures.

2. Using the "Clinical Method" to Measure Logical Thinking

Logical thinking has been traditionally assessed using structured or semi-structured clinical interviews in which questions are asked by the interviewer (specifically trained expert) individually or in small groups. The purpose of a clinical interview is not restricted to scoring the product of problem solution as in standardized diagnostic measures; it also involves finding out how exactly the solution was reached, i. e. which cognitive processes were involved or not involved, and in what order.

Mistakes made while solving a problem indicate deficiencies in specific logical operations and thus are of high importance when interpreting the results. Using the clinical interview data, the expert makes inferences about the levels of specific logical operations and the overall level of logical thinking ability [Bringuier 2000; Piaget 1994a; 1994b; 1994c; Piaget, Inhelder 2003].

The disadvantage of the "clinical method" is its low level of resource efficiency, especially when using it on large samples or in large-scale assessments. First, there should be a sufficient number of highly-qualified experts capable of producing as unbiased diagnostic results as possible. Second, individually administered tasks make diagnosis essentially more time-consuming. In most contemporary educational situations, there are just not enough personnel and time for the administration of Piagetian tasks via the "clinical method" [Avila de, Pulos 1979; Meyer 1972].

Therefore, researchers face the need to develop a new measure of logical thinking that would be a decent alternative to clinical interview. It should be easy to use for teachers, instructional designers, practicing psychologists, and researchers doing large-scale data collection. It should also preserve construct specifics, i. e. allow observing how cognitive processes forming part of logical operations get involved in problem solving.

Scalability of the instrument can be achieved by standardized testing, where all behaviors observed in a test are interpreted using diagnostic criteria that are uniform, pre-determined, and objective (unaffected by expert bias).

3. Using Standardized Methods to Measure Logical Thinking

Since the 1960s, Piaget's followers have attempted to replace individual testing (clinical interviews) with group testing in order to check correspondence between mandatory school curricula and pupils' cognitive levels, e. g. in teaching life science disciplines [Lovell 1961; Shayer 1978; Shayer, Küchemann, Wylam 1976].

Researchers enquired whether group testing could yield results as good as those of clinical interviews. It was shown that clinical interview and frontal assessment (e. g. when stimulus material is displayed on a screen and respondents write down their answers) produce comparable outcomes [Faust 1983; Renner et al. 1978; Rowell, Hoffmann 1975; Shayer 1979]. Yet, frontal assessment is not equal to standardized testing as it allows variations in how stimuli are presented and how results are interpreted.

We analyzed a number of studies describing the use of standardized diagnostic measures of logical thinking in the period of transition between the Piagetian stages of concrete and formal operational thinking in large-scale group-administered quantitative assessments. The main objective of their authors was to ensure the same degree of interpretation as in clinical interviews, while reducing the time and re-

sources required for test administration. The standardized instruments analyzed differ by a few critical parameters:[1]

(1) Test content:
 a. Construct measured (logical operations);
 b. Scope of measurement: test outcomes (product) and/or reasoning (process);
(2) Response format:
 a. Selected response (multiple choice or yes/no (closed-ended) questions);
 b. Essay or short constructed response (open-ended questions);
(3) Scoring:
 a. Automated;
 b. Manual (expert);
(4) Stimuli: physical objects, pictures, text, video.

The tests differ both in the set of logical operations assessed and the scope of measurement. Despite some differences in the types of logical operations that they measure, all the tests diagnose both concrete and formal operational thinking. As for the scope of measurement, the tests can only measure participants' final scores (product) [Tisher 1971; Raven 1973; Milakofsky, Patterson 1979; Avila de, Pulos 1979; Roberge, Flexer, 1982; Bergling 1998; Bakken et al. 2001]. However, some authors developed diagnostic instruments that assess both the product and process [Longeot 1962; 1965; Staver, Gabel 1979; Lawson 1978; Tobin, Capie 1981; Roadrangka 1991]. The latter is diagnosed via open-ended questions, which require manual scoring by experts.

On the one hand, the obvious benefits of closed-ended questions are reduced testing time and lower student stress as writing skills are not involved. In addition, multiple-choice testing ensures score objectivity, whereas manual scoring of constructed responses by different experts may cause inconsistency of scores and reduce their reliability [Roadrangka 1991]. On the other hand, the closed-ended format does not allow to obtain justifications for choosing the specific answer, thereby curtailing the diagnostic potential of the instrument and leaving open the possibility of answering correctly by making a random choice.

Stimuli are presented not only as plain text but also as a combination of text and pictures [Longeot 1962; 1965; Tobin, Capie 1981; Bergling 1998], pictorial material alone [Milakofsky, Patterson 1979; Avila de, Pulos 1979; Bitner-Corvin 1988], a combination of text and video [Staver, Gabel 1979; Tobin, Capie 1981], a combination of textual de-

[1] The table at doi:10.17632/vxt3237yvt.1 presents a systematization of tests the psychometric characteristics of which are publicly accessible in peer-reviewed publications [Uglanova I. L., Pogozhina I. N. (2021) What the New Measure of Thinking in School Students Has to Offer to Contemporary Education. *Mendeley Data*, V1].

scriptions and physical objects [Roberge, Flexer, 1982], physical objects alone [Tisher 1971; Lawson 1978], etc.

Entirely verbal stimulus material is valid for assessing the development of formal operational thought; however, it implies a lot of reading, which can lead to semantic bias in the assessment of concrete operational structures. Furthermore, in the absence of physical objects as stimuli, students may stop perceiving the problem as significant [Lawson 1978]. Verbal stimulus materials and open-ended questions represent the greatest challenge for younger children [Lawson 1978; Roadrangka 1991].

On the whole, standardized tests as measures of the level of logical structures prove to be reliable and valid. As one of the ways of measuring their validity, results obtained in a standardized test are compared to results obtained in a clinical interview. Final reliability coefficients vary from insignificant on some scales [Staver, Gabel 1979] to significant at the level of 0.88 on others [Avila de, Pulos 1979; Bitner-Corvin 1988].

As we can see, the use of standardized measures allows to reduce testing time without sacrificing the scope of the logical reasoning construct to be measured. However, the problem of automated scoring has not been solved so far. In addition, there are no studies measuring logical operations in realistic situations, only in laboratory contexts.

Computerized performance-based assessment is a promising diagnostic measure of logical thinking (process and product) in technology-enhanced learning environments. This type of assessment allows to apply automated scoring, evaluate both the outcome and the reasoning process, and create real-life systems in every item [Wang, Shute, Moore 2015].

4. Requirements for the New Methodology

The new design methodology for diagnostic measures of logical thinking should, first of all, allow diagnosing the process of arriving at the solution and, second, provide feedback (which implies analysis of the results) for teachers, instructional designers, and practicing psychologists without the involvement of experts. Furthermore, the new measures should follow a principled assessment design framework [American Educational Research Association, American Psychological Association, National Council on Measurement in Education 2014; Messick 1992].

4.1. The process measurement requirement

Positively, "the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics" [Messick 1992:17]. Traditional multiple-choice items are not suitable for assessing the process as they only register the outcome (correct/incorrect) without making allowance for how it was achieved [Griffin, McGaw, Care 2012; Messick 1994; Razzouk 2011].

Prominent among the alternative modes of assessment are performance-based tasks [Messick 1994] which focus on both the product and process of solution. Today, computerized performance-based assessment is widely used to measure complex constructs in the form of games, simulations, and scenario-based tasks [Klerk de, Eggen, Veldkamp 2016; Graesser, Kuo, Liao 2017; Sun et al. 2020]. However, performance-based diagnostic measures of logical thinking within the framework of Piaget's theory have not been described in literature yet.

4.2. The automated scoring and feedback requirement

Transition from the "clinical method" to performance-based standardized testing provides a unique combination of benefits of both multiple-choice and constructed-response items: the analysis limitation is overcome by matching the examinee's product to the correct answer and assessing the solution process without involving experts.

Another advantage of performance-based testing is that it can be implemented in the format of stealth assessment, whereby test situations are embedded seamlessly into a computer-based learning or gaming environment such that the learner is unaware of being assessed [Wang, Shute, Moore 2015]. Stealth assessment helps to reduce test anxiety and maintain learners' engagement, while at the same time providing authentic contexts to make interpretation of assessment results as close as possible to everyday real-life problems.

Finally, the diagnostic framework should allow the use of test results to create psychological assistance programs for students whose levels of logical thinking are not enough to succeed in a discipline. The approach proposed in this article suggests decomposition of formal operational structures (activity-based diagnostic measurement, in some authors' terminology [Ilyasov 1986; Talyzina 2018]), which will allow assessment of not only the development of a logical operation as such but also the development of its constituent cognitive processes. Just as in clinical interviews, a lot of importance is assigned to analysis and interpretation of mistakes made during the solution process. Mistake is understood here as non-completion or incorrect completion of a cognitive task that is part of the logical operation measured and has to be completed to progress through the scenario. The developmental stage of the logical operation is determined based on the data collected.

In the future, the obtained diagnostic results can be used to elaborate an approximate action plan for promoting the development of the missing components of a "problematic" logical operation as well as to design and implement, within the activity-based approach to learning and socialization, a system of teaching practices required for such operation to be "internalized". The proposed methodology for diagnosing the composition and structure of logical operations will make it possible to accurately identify the "problem areas" in as many students as possible, so that each of them could be offered an appropriate personalized formative learning program.

5. The New Methodology and Instrument for Diagnosing Logical Thinking in School Students

The new methodology for diagnostic measurement of logical thinking proposed here involves the following:

- Identifying the logical operations as defined by Piaget [Piaget 1994a; 1994b; 1994c], their composition and structure as a diagnostic framework for activity-based diagnostic measurement and as the basis for design of formative learning programs in the future [Ilyasov 1986; Talyzina 2018];
- Analyzing the benefits and limitations of the currently applied measures of logical thinking;
- Applying modern psychometric techniques and digital technology.

To measure logical thinking in school students, we suggest using scenario-based tasks that put students into a technology-enhanced environment akin to learning and real-life contexts. In scenario-based tasks, examinees consecutively perform tasks that are interconnected within a context-enhanced story (scenario). The process and product of solution in such problems act as behavioral indicators allowing to assess the development of a specific cognitive process within a logical operation. Scenario-based tasks are similar in their format to computer games; unlike games, however, they involve less variation, which results in a more standardized assessment. Such an approach allows implementing the principles of stealth assessment for diagnostic measurement of logical thinking.

The diagnostic profiles constructed on the basis of test results describe levels of development of each logical structure (combinatorial logic, INRC group) measured. A teacher, psychologist, or parent can view information about the developmental stage for each logical operation separately to construct a comprehensive personal profile of logical thinking. As of now, scenario-based tests have been developed for fifth- and seventh-grade pupils.[2]

The new diagnostic methodology thus meets the design requirements for tasks measuring logical thinking and at the same time ensures that diagnostic measurement results can be scored without the involvement of experts.

6. Meeting the Process Measurement Requirement

6.1. Stimulus presentation

To measure concrete operational thinking, test items are supported with physical objects or their visual representations (videos, pictures), as pupils mostly think in images at this stage of development. Emergence of formal operations in the adolescent's mind enables them to

[2] The 4C's Project for assessment of 21st-century skills (critical thinking, creativity, communication, and collaboration) designed by the Centre for Psychometrics and Measurements in Education (Laboratory for New Construct Measurement and Test Design) of the HSE Institute of Education under an R&D contract with the Charitable Foundation "Investment to the Future".

build complex logical relations mentally without interacting with physical objects or their visual representations. That is why a verbal mode of stimulus presentation is more relevant than a pictorial one for diagnosing the developmental stage of formal operations [Avila de, Pulos 1979; Piaget 1994a; 1994b; 1994c].

The computerized scenario-based tasks that we propose allow to combine verbal and pictorial stimuli and simulate interactions with physical objects, which makes it possible to assess the examinee's ability to mentally manipulate various types of materials (physical, pictorial, symbolic) and ensure a more accurate measurement of the level of logical reasoning (concrete or formal operations stage).

6.2. Combinatorial problems

Combinatorial operations are a logical structure emerging within the stage of formal operations. An example of a combinatorial problem for fifth-grade pupils is presented in Figure 1. The problem offers an imaginary scenario, asking students to make fuel for a space ship. The instructions are partially verbal, while the stimulus material imitates interaction with real-life objects, allowing examinees to try out all possible combinations of ingredients one by one while avoiding repetitions and omissions.

In the course of the solution process, we systematically evaluate the development of cognitive processes within combinatorial operations: (1) identification of variables to be combined in the problem scenario; (2) combination of variables presented in the problem scenario with regard to instructions (into sets of two, three, etc.); (3) control of variable combinations (systematic combination of variables into all possible sets in accordance with the given conditions while avoiding repetitions).

As the first step, the examinee explores the planet and collects ingredients for the fuel (Figure 1a), i. e. identifies the variables. After that, the examinee is given illustrated instructions for research (testing of all possible combinations of the ingredients collected) to produce fuel. Particular focus is placed on making sure that students fully understand the instructions (Figure 1b), i. e. they realize that combination of variables should be part of the solution process. Next, the examinee solves the problem (Figure 1c) by combining the variables into groupings that satisfy the given conditions and exercising control over the combinations.

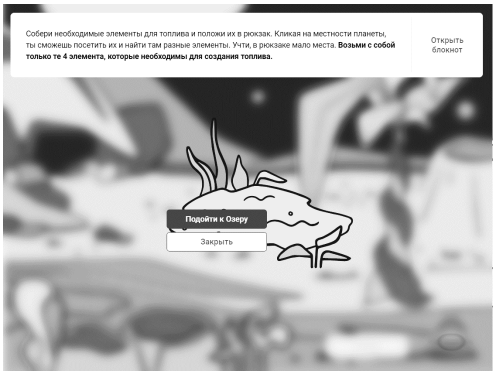
As students are asked to combine stimulus components systematically in search for candidate solutions and allowance can be made for mistakes that they make, not only the product but also the process of solution can be assessed.

The final assessment of the development of combinatorial operations is based on the scores assigned for the product (the number of correct variable combinations) and the process of solution (absence of repetitions). Three stages of combinatorial operations development are identified in the diagnostic profile:

Figure 1. **An example of a scenario-based problem measuring the development of combinatorial operations.**

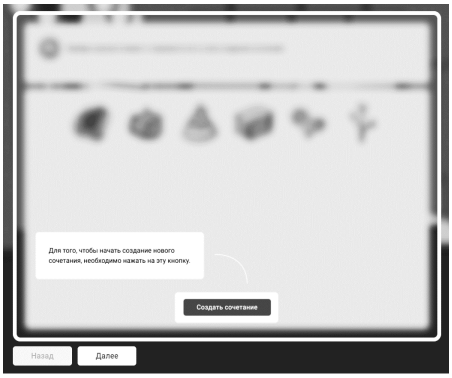
(a)

Collect the elements for making fuel and put them into the backpack. By clicking on planet locations, you can visit them and find various elements. Be mindful that there is little room in the backpack. Take only the four elements that are necessary to produce fuel.
Open notebook
Done

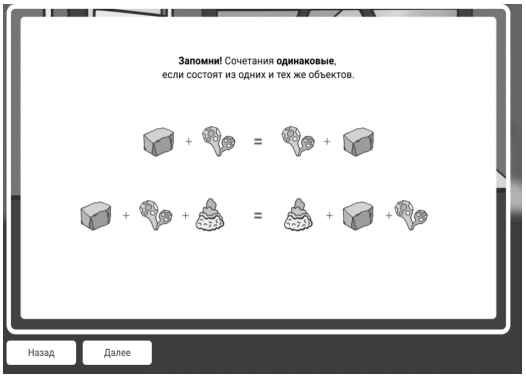


Collect the elements for making fuel and put them into the backpack. By clicking on planet locations, you can visit them and find various elements. Be mindful that there is little room in the backpack. Take only the four elements that are necessary to produce fuel.
Open notebook
Approach 'Lake'
Close

(b)

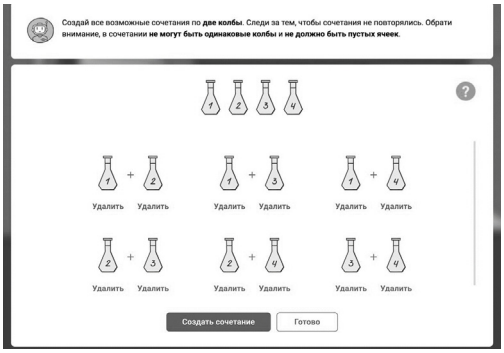


Click this button to start a new combination
Create a combination
Back
Next

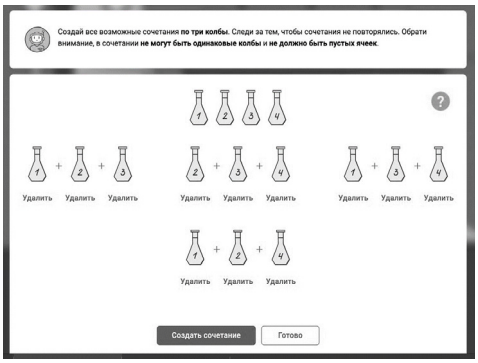


Remember! Combinations are duplicate if they consist of the same objects.
Back
Next

(c)



Create all possible combinations of two flasks. Make sure you have no duplicate combinations. Be mindful that there should be no identical flasks and no empty boxes in a combination.
Remove (×12)
Create a combination
Done



Create all possible combinations of three flasks. Make sure you have no duplicate combinations. Be mindful that there should be no identical flasks and no empty boxes in a combination.
Remove (×12)
Create a combination
Done

Stage 1: Combinatorial operations are not developed. The student is able to identify the variables to be combined in a problem scenario (1) but fails to produce combinations satisfying the given conditions (2) as well as to control the process of combination by avoiding repetitions (3);

Stage 2: Combinatorial operations are partially developed. The student is able to identify the variables to be combined in a problem scenario (1) but makes mistakes in the process or fails to combine all the variables (2) as well as to control the process of combination by avoiding repetitions (3);

Stage 3: All the cognitive processes within combinatorial operations are fully developed. The student systematically generates all the correct configurations satisfying the problem statement and avoids repetitions.

6.3. INRC problems

The emergence of the cognitive structure of four transformations INRC manifests itself in the ability to analyze the problem situation and examine the impact of all the given conditions one by one. By checking whether the situation is affected by a specific factor, the adolescent mentally performs two divergent operations, varying the factor characteristics (modifying the variables) and at the same time holding all the other factors constant (fixing the variables).

For instance, in a classic Piagetian task, the bending of rods with weights applied [Piaget, Inhelder 2003], to find out whether length, diameter, shape, and material have a difference in bending, the adolescent tests each variable one by one. To test for the effects of length, they select for comparison long and short rods that are alike in every other aspect: both are thick, round, and made of steel. Next, they bend the two rods and analyze the results. If the rods bend equally, a conclusion is drawn that length has no impact on bending, and vice versa. Effects of all the other variables in the scenario are tested in a similar way.

The same principle underlies the problem designed for seventh-graders (Figure 4). In a fantastic scenario, the examinee is asked to find out which residents of an old house have a chance of turning into a ghost.

Just as in combinatorial problems, this one allows to assess not only the outcome (correct/incorrect) but also the process of solution by measuring the development of processes within the logical operation of variable fixation as an indicator of INRC group [Ilyasov 1986; Baldina 1987; Pogozhina 2006]:

(1) Identify variables and their values in the problem scenario;
(2) Fix the variables: to determine the impact of variables on the outcome, the examinee has to compare situations in which all the variables except one are held constant;

Figure 2. **An example of a scenario-based problem measuring the development of INRC group.**

(a)

If you haven't sorted the evidence into three columns yet, it's just about time to do it. Classify all the pieces of evidence.

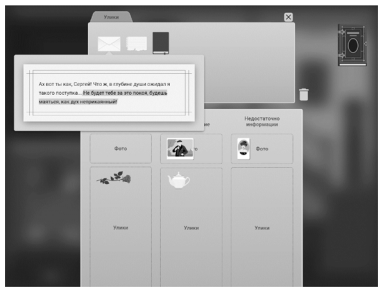
Got it

I'm sick and tired of this ghost! And, most importantly, why does it sneeze? Sneezing should be done quietly, into a tissue. I remember my late uncle Vladimir Alexeevich who coughed in a very cultured manner. He would wrap his scarf around his neck, brew some herbal tea, and sit there groaning. You know, he would always suffer some imaginary illnesses, but I never saw him actually sick in his whole life

At least this ghost doesn't do any damage. We have so many fragile things here, like this antique porcelain tea set... It belonged to Sergey Konstantinovich. He loved porcelain and was rarely seen without his favorite cup. Such a nice little thing

Evidence

(b)



Evidence

So that's what you're up to, Sergey! Well, deep down inside, I knew this would come... For what you've done, you will never find peace anymore and will be roaming forever like a restless ghost!

Insufficient evidence

Photo

(c)

Evidence

Not a ghost
Natalya Vasilyevna getting dressed up for the theater
Tea
Carries a tissue with a coat of arms
Visits a witch doctor

Not a ghost
Vanechka playing a ghost game
Tea
Carries a tissue with a coat of arms
Visits a witch doctor

Ghost
Sergey Konstantinovich
Tea
Carries a tissue with a coat of arms
Visits a witch doctor

Ghost
Arina Sergeevna
Tea
Carries a tissue with a coat of arms
Visits a witch doctor

Ghost
Vladimir Alexeevich at work
Tea
Carries a tissue with a coat of arms
Visits a witch doctor

No a ghost
Pavel Alexandrovich
Tea
Carries a tissue with a coat of arms
Visits a witch doctor

(3) Make a logical inference: if the effects of the variable are the same across the situations compared, it has no impact on the outcome; if the effects vary, the variable does have an impact.
(4) According to the scenario, the student first collects information on all the residents (those who have and have not turned into ghosts, Figures 2a and 2b), i. e. identifies variables and their values (1). Next, they fix the differing characteristics at constant levels (Figure 2c), i. e. perform variable fixation (2) in order to identify the factor affecting the probability of turning into a ghost—that is, to make a logical inference (3).

The results of systematic performance of all scenario-based tasks are used to construct diagnostic profiles, i. e. developmental levels of the logical operation of variable fixation as an indicator of INRC group:

Stage 1: The operation is not developed. The student is able to identify groups of variables described in the problem scenario (1) but cannot fix the differing factors (2) or make inferences about their effects (3);

Stage 2: The operation is partially developed. Having identified the right variables and their values (1), the student systematically fixes only some of them, making mistakes in variable fixation (2) and drawing invalid inferences (3);

Stage 3: All the cognitive processes within the logical operation of variable fixation are fully developed. The student identifies all the possible variables and their values in the problem scenario, systematically fixes all the variables except one, analyzes the effects of the non-fixed variable, and makes valid inferences.

7. Meeting the Automated Scoring and Feedback Requirement

The critical benefit of standardized testing is the easiness of scoring. Development of a scoring system that works without expert evaluation allows to maintain this benefit.

In the computerized scenario-based tasks that we designed, examinees have no opportunity to type their responses, but their actions can be interpreted based on the choices that they make, which allow to understand which cognitive processes are involved in solving the problem.

Feedback for examinees is prepared using quantitative data analysis and advanced psychometric techniques [Almond et al. 2015; Jeon et al. 2020]. The existing measures of logical thinking offer various interpretation frameworks, possible indicators of developmental gains including an increase in the total score on a test [Staver, Gabel 1979] or attainment of a cut-off score corresponding to the transition to formal operational thought [Lawson 1978; Roadrangka 1991]. The diagnostic system proposed here uses problem solution results to construct a personal profile of logical thinking, which reflects the development of cognitive processes within a specific logical operation.

The profile includes diagnostic measurement results for the logical operation of combinatorics (Stage 1, 2, or 3) and the logical operation of variable fixation as an indicator of INRC group (Stage 1, 2, or 3). Each stage reflects mastery of cognitive processes within a formal operational structure.

By contrast with the traditional approaches to diagnostic measurement that imply selecting cut-off values for total scores or specific scales, we suggest using modern psychometric techniques, i. e. treating the construct as a discrete variable and measuring the probability of being at a specific stage of development [Almond et al. 2015]. Such discrete arrangement and presentation of results allows automated feedback that can be easily used by teachers, parents, and psychologists. However, the staging procedure is yet to be validated in further research on the instrument's quality.

8. Conclusion

Comprehensibility and adequacy of school curricula can be provided by learning personalization, which requires performing large-scale objective assessments of logical reasoning and making allowance for the assessment results in teaching and in design of psychological assistance programs.

The Piagetian method of clinical interview allows the most comprehensive and accurate assessment of logical thinking at a specific stage of development. At the same time, the "clinical method" has a number of limitations when applied on a large scale, such as lack of highly-qualified experts, expert bias, and too much time per interview. One more essential limitation consists in that the examinee is always aware of being tested.

Standardized measures allow overcoming the limitations of clinical interviews. The existing group tests of logical thinking demonstrate sufficient reliability and validity, but they still involve expert evaluation. Correctness of answers to multi-choice questions can be assessed automatically, but experts are inevitably involved evaluate multiple choice justifications, i. e. the solution process. Besides, measures of logical operations in technology-enhanced learning environments have not been described in literature yet.

Overcoming of the existing limitations in diagnostic measurement of logical thinking development necessitates a new methodology that can be used in large-scale data collection and allows observing how logical operations get involved in problem solving.

The methodology proposed in this article makes allowance for the specific aspects of both logical thinking and large-scale testing with automated scoring. Furthermore, diagnostic results obtained with the proposed measure can be used for designing personalized psychological assistance programs for students within the framework of activity-based learning and socialization, as it meets the item design requirements imposed by complexity of the construct measured as well

as the interpretability requirement for large-scale standardized tests with automated scoring.

The item design requirements imposed by complexity of the construct measured—the composition and structure of logical thinking—are satisfied by the following:

- Computerized performance-based tasks, which allow selecting formats of stimuli (text, pictorial representations, object simulations) that contribute to validity of logical thinking assessment;
- Items requiring the use of combinatorial operations;
- Items requiring the cognitive structure of four transformations INRC, e. g. where examinees have to systematically fix all the variables except one to find the optimal solution under given conditions;
- Personal profiles of logical thinking for users of test results, providing scores on each logical operation.

The methodology proposed also satisfies the interpretability requirement imposed by the need to avoid expert evaluation in large-scale standardized testing. Feedback on test results is based on quantitative data analysis, allowing to avoid the costs associated with recruiting a number of highly-qualified experts, eliminate expert bias, and essentially reduce the overall testing time in case of large samples.

9. Recommendations on Using the Diagnostic Results in Educational Practice

Overall, the design of modern school curricula makes allowance for stages of cognitive development, so an "average" student encounters little difficulty in learning. However, findings show that the currently widespread perceptions of the age at which students develop formal operational thought appear to be overly simplified [Shayer, Küchemann, Wylam 1976]. Indeed, there is evidence that formal logical operations start to emerge in early adolescence, yet even college students sometimes reason at the concrete level [Lawson 1978; Tobin, Capie 1981; Tisher 1971]. Individually administered diagnostic measurement of logical thinking can help determine whether a particular adolescent is ready to learn the material adequate for their grade.

Diagnostic results will allow not only to establish whether or not specific logical structures are present but also to measure the stage of their development, thereby highlighting the cognitive processes that have not yet formed in a particular student. In the future, this diagnostic measure can be used for designing personalized psychological assistance programs to foster the development of the missing cognitive structures.

10. Limitations and Avenues of Further Research

An important characteristic of the methodology proposed in this article is that examinees are placed in a problematic situation that requires the use of the logical operations of interest. On the one hand,

such an approach allows interpreting the test results not in isolation, as in the "laboratory" conditions of classical Piagetian tasks, but within real-life contexts. On the other hand, diagnostic scenarios impose limitations on the extrapolation of test results. It cannot be guaranteed that the student will apply the same logical operations in a broad range of contexts.

There is empirical evidence that the same student deploys logical reasoning skills of different levels to solve problems presented in different contexts [Bart 1972; 1978; Cohen 1980; Twidle 2006]. This characteristic of logical thought development requires further examination of the phenomenon itself and the aspects of its measurement.

A promising avenue of further personalized learning research and practice is to examine more closely the relationships between levels of logical thinking and academic success in specific disciplines. Interpretation of the reasons for academic failure in school disciplines could be performed more thoroughly and associate learning outcomes not only with teaching quality or curriculum difficulty but also with the individual rates and characteristics of logical structure development.

The article was prepared in the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075–15–2020–928).

We are grateful to the team of the 4C's Project of the Laboratory for New Construct Measurement and Test Design at the Centre for Psychometrics and Measurements in Education (Institute of Education, National Research University Higher School of Economics), particularly to Anastasia Kamaeva, for their creative and professional approach to item design.

References

Almond R. G., Mislevy R. J., Steinberg L. S., Yan D., Williamson D. M. (2015) *Bayesian Networks in Educational Assessment*. New York: Springer. doi:10.1007/978–1–4939–2125–6

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (eds) (2014) *Standards for Educational and Psychological Testing*. Lanham, MD: American Educational Research Association.

Avila de E., Pulos S. (1979) Group Assessment of Cognitive Level by Pictorial Piagetian Tasks. *Journal of Educational Measurement*, vol. 16, no 3, pp. 167–175. doi:10.1111/j.1745–3984.1979.tb00098.x

Bakken L., Thompson J., Clark F. L., Johnson N., Dwyer K. (2001) Making Conservationists and Classifiers of Preoperational Fifth-Grade Children. *The Journal of Educational Research*, vol. 95, no 1, pp. 56–61. doi: 10.1080/00220670109598783

Baldina N. P. (1987) *Usvoenie logicheskikh priyomov pri raznykh tipakh ucheniya* [Mastering Logical Techniques by Different Types of Teaching] (PhD Thesis), Moscow: Lomonosov Moscow State University.

Bart W. M. (1978) Issues in Measuring Formal Operations. *The Genetic Epistemologist*, vol. 7, pp 3–4.

Bart W. M. (1972) Construction and Validation of Formal Reasoning Instruments. *Psychological Reports*, vol. 30, no 2, pp. 663–670. doi:10.2466/pr0.1972.30.2.663

Bergling B. M. (1998) Constructing Items Measuring Logical Operational Thinking: Facet Design-Based Item Construction Using Multiple Categories Scoring. *European*

Journal of Psychological Assessment, vol. 14, no 2, pp. 172–187. doi:10.1027/1015–5759.14.2.172

Bitner-Corvin B.L. (1988) Is the GALT a Reliable Instrument for Measuring the Logical Thinking Abilities of Students in Grades Six through Twelve? Paper presented at the *61st Annual Meeting of the National Association for Research in Science Teaching (Lake of the Ozarks, MO, April 10–13, 1988)*. Available at: https://files.eric.ed.gov/fulltext/ED293716.pdf (accessed 20 October 2021).

Bringuier J.-C. (2000) Besedy s Zhanom Piazhe [Conversations with Jean Piaget]. *Psikhologichesky zhurnal*, vol. 21, no 2, pp. 138–144.

Cohen H. G. (1980) Dilemma of the Objective Paper-and-Pencil Assessment within the Piagetian Framework. *Science Education*, vol. 64, no 5, pp. 741–745. doi:10.1002/sce.3730640521

Comenius J. A. (1939) Velikaya didaktika [The Great Didactic]. Moscow: State Educational and Pedagogical Publishing House of the People's Commissariat of Education of the RSFSR.

DeVries R. (1974) Relationships among Piagetian, IQ, and Achievement Assessments. *Child Development*, vol. 45, no 3, pp. 746–756.

Faust D. (1983) A Promising Approach to the Development of a Group Piagetian Measure. *Psychological Reports*, vol. 53, no 3, pp. 771–774. doi:10.2466/pr0.1983.53.3.771

Goldschmid M. L. (1967) Different Types of Conservation and Nonconservation and Their Relation to Age, Sex, IQ, MA, and Vocabulary. *Child Development*, vol. 38, no 4, pp. 1229–1246. doi: 10.1111/j.1467–8624.1967.tb04398.x

Graesser A., Kuo B.-C., Liao, C.-H. (2017) Complex Problem Solving in Assessments of Collaborative Problem Solving. *Journal of Intelligence*, vol. 5, no 2, Article no 10. doi:10.3390/jintelligence5020010

Griffin P., McGaw B., Care E. (2012) *Assessment and Teaching of 21st Century Skills*. Heidelberg: Springer.

Hathaway W. E., Hathaway-Theunissen A. (1975) The Unique Contributions of Piagetian Measurement to Diagnosis, Prognosis, and Research of Children's Mental Development. *Piagetian Theory and the Helping Professions* (ed. G. I. Lubin), Los Angeles: University of Southern California.

Hautamäki J., Thuneberg H. (2019) Koulutuksen Tasa-Arvotaseet [Equity Balances in Education]. *Vainikainen Perusopetus, Tasa-Arvo ja Oppimaan Oppiminen: Valtakunnallinen Arviointitutkimus Peruskoulun Päättövaiheesta* [Comprehensive Education, Equality, and Learning to Learn: Nationwide Evaluative Research on the Final Phase of Comprehensive Education] (eds J. Hautamäki, I. Rämä, M.-P.), Helsinki: University of Helsinki, pp. 77–96.

Hienonen N. (2020) Does Class Placement Matter? Students with Special Educational Needs in Regular and Special Classes. *Helsinki Studies in Education*, no 87. Available at: https://helda.helsinki.fi/bitstream/handle/10138/318683/Doesclas.pdf?sequence=1&isAllowed=y (accessed 20 October 2021).

Il'yasov I.I. (1986) *Struktura protsessa ucheniya* [Structure of the Teaching Process]. Moscow: Moscow University.

Inhelder B., Piaget J. (1958) *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures*. New York: Basic Books.

Jeon M., Draney K., Wilson, M., Sun Y. (2020) Investigation of Adolescents' Developmental Stages in Deductive Reasoning: An Application of a Specialized Confirmatory Mixture IRT Approach. *Behavior Research Methods*, vol. 52, no 1, pp. 224–235. doi:10.3758/s13428–019–01221–5

Kaufman R. A. (1972) *Educational System Planning*. New York: Prentice-Hall.

Klerk de S., Eggen T. J.H.M., Veldkamp B. P. (2016) A Methodology for Applying Students' Interactive Task Performance Scores from a Multimedia-Based Performance Assessment in a Bayesian Network. *Computers in Human Behavior*, vol. 60, July, pp. 264–279. doi:10.1016/j.chb.2016.02.071

Lawson A.E. (1978) The Development and Validation of a Classroom Test of Formal Reasoning. *Journal of Research in Science Teaching*, vol. 15, no 1, pp. 11–24. doi:10.1002/tea.3660150103

Lawson A. E., Blake A. J.D, Nordland F. H. (1974) Piagetian Tasks Clarified: The Use of Metal Cylinders. *The American Biology Teacher*, vol. 36, no 4, pp. 209–211. doi:10.2307/4444748

Lawson A. E., Renner J. W. (1975) Relationships of Science Subject Matter and Developmental Levels of Learners. *Journal of Research in Science Teaching*, vol. 12, no 4, pp. 347–358. doi:10.1002/tea.3660120405

Longeot F. (1965) Analyse Statistique de Trois Tests Genetiques Collectifs. *Bulletin de l'institut National D'Etude*, vol. 20, no 4, pp. 219–237.

Longeot F. (1962) An Essay of the Application of Genetic Psychology to Differential Psychology. *Bulletin de l'Institute d'Etude Du Travail et d'Orientation Professionelle*, no 18, pp. 153–162.

Lovell K. (1961) A Follow-Up Study of Inhelder and Piaget's The Growth of Logical Thinking. *British Journal of Psychology*, vol. 52, no 2, pp. 143–153. doi: 10.1111/j.2044–8295.1961.tb00776.x

Lovell K., Shields J. B. (1967) Some Aspects of a Study of the Gifted Child. *British Journal of Educational Psychology*, vol. 37, no 2, pp. 201–208. doi:10.1111/j.2044–8279.1967.tb01929.x

Malhotra S. (2020) Psychometric Intelligence and Academic Achievement: A Comparative Analysis of Elementary Schools. *EDUTEC: Journal of Education and Technology*, vol. 3, no 2, pp. 83–95. doi:10.29062/edu.v3i2.40

Messick S. (1994) *Alternative Modes of Assessment, Uniform Standards of Validity1*. Princeton, New Jersey: Educational Testing Service. Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333–8504.1994.tb01634.x (accessed 20 October 2021).

Messick S. (1992) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Paper presented at the *Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 1992)*. Available at: https://files.eric.ed.gov/fulltext/ED390891.pdf (accessed 20 October 2021).

Meyer C. E. (1972) Can Piaget's Theory Provide a Better Psychometry? *Piagetian Theory and the Helping Professions* (eds J. F. Magary, M. Poulsen, G. I. Lubin, G. Coplin), Los Angeles, CA: Children's Hospital, pp. 5–10. Available at: https://files.eric.ed.gov/fulltext/ED085612.pdf (accessed 20 October 2021).

Milakofsky L., Patterson H. O. (1979) Chemical Education and Piaget: A New Paper-Pencil Inventory to Assess Cognitive Functioning. *Journal of Chemical Education*, vol. 56, no 2, pp. 87–90. doi:10.1021/ed056p87

Piaget J. (2008) *Rech i myshlenie rebyonka* [The Language and Thought of the Child]. Moscow: Rimis.

Piaget J. (1994a) Genesis chisla u rebyonka [The Child's Conception of Number] *Izbrannye psikhologicheskie trudy* [Selected Psychological Works of J. Piaget], Moscow: International Pedagogical Academy, pp.237–582.

Piaget J. (1994b) Logika i psikhologiya [Logic and Psychology] *Izbrannye psikhologicheskie Trudy* [Selected Psychological Works of J. Piaget], Moscow: International Pedagogical Academy, pp. 583–628.

Piaget J. (1994c) Psikhologiya intellekta [The Psychology of Intelligence] *Izbrannye psikhologicheskie Trudy* [Selected Psychological Works of J. Piaget], Moscow: International Pedagogical Academy, pp. 51–236.

Piaget J., Inhelder B. (2003) *Psikhologiya rebyonka* [The Psychology of the Child]. Saint Petersburg: Piter.

Pogozhina I. N. (2006) Metodika diagnostiki formal'no-logicheskogo myshleniya: diagnostika sformirovannosti struktury INRC [Methods of Diagnostics of Formal-Logical Thinking: Diagnostics of the Formation of the INRC Structure]. *Shkol'ny psikholog*, no 9, pp. 40–43.

Raven R. J. (1973) The Development of a Test of Piaget's Logical Operations. *Science Education*, vol. 57, no 3, pp. 377–385. doi:10.1002/sce.3730570316

Razzouk R. (2011) *Using Evidence-Centered Design for Developing Valid Assessments of 21st Century Skills*. Bellevue, WA: Edvation.

Renner J. W., Sutherland J., Grant R., Lawson A. E. (1978) Displacement Volume, An Indicator of Early Formal Thought; Developing a Paper-and-Pencil Test. *School Science and Mathematics*, vol. 78, no 4, pp. 297–303. doi:10.1111/j.1949–8594.1978.tb09362.x

Roadrangka V. (1991) The Construction of a Group Assessment of Logical Thinking (GALT). *Kasetsart Journal of Social Sciences*, vol. 12, no 2, pp. 148–154.

Roberge J. J., Flexer B. K. (1982) The Formal Operational Reasoning Test. *Journal of General Psychology*, vol. 106, no 1, pp. 61–67. doi: 10.1080/00221309.1982.9710973

Rowell J. A., Hoffmann P. J. (1975) Group Tests for Distinguishing Formal from Concrete Thinkers. *Journal of Research in Science Teaching*, vol. 12, no 2, pp. 157–164. doi:10.1002/tea.3660120210

Shayer M. (1979) Has Piaget's Construct of Formal Operational Thinking Any Utility? *British Journal of Educational Psychology*, vol. 49, no 3, pp. 265–276. doi:10.1111/j.2044–8279.1979.tb02425.x

Shayer M. (1978) The Analysis of Science Curricula for Piagetian Level of Demand. *Studies in Science Education*, vol. 5, iss. 1, pp. 115–130. doi: 10.1080/03057267808559861

Shayer M., Küchemann D. E., Wylam H. (1976) The Distribution of Piagetian Stages of Thinking in British Middle and Secondary School Children. *British Journal of Educational Psychology*, vol. 46, no 2, pp. 164–173. doi:10.1111/j.2044–8279.1976.tb02308.x

Staver J. R., Gabel D. L. (1979) The Development and Construct Validation of a Group-Administered Test of Formal Thought. *Journal of Research in Science Teaching*, vol. 16, no 6, pp. 535–544. doi:10.1002/tea.3660160607

Sun C., Shute V. J., Stewart A., Yonehiro J., Duran N., D'Mello S. (2020) Towards a Generalized Competency Model of Collaborative Problem Solving. *Computers & Education*, vol. 143, no 1, Article no 103672. doi:10.1016/j.compedu.2019.103672

Talyzina N. F. (2018) *Deyatel'nostnaya teoriya ucheniya* [Activity Theory of Teaching]. Moscow: Moscow University.

Tisher R. P. (1971) A Piagetian Questionnaire Applied to Pupils in a Secondary School. *Child Development*, vol. 42, no 5, pp. 1633–1636. doi:10.2307/1127935

Tobin K. G., Capie W. (1981) The Development and Validation of a Group Test of Logical Thinking. *Educational and Psychological Measurement*, vol. 41, no 2, pp. 413–423. doi:10.1177/001316448104100220

Twidle J. (2006) Is the Concept of Conservation of Volume in Solids Really More Difficult than for Liquids, or Is the Way We Test Giving Us an Unfair Comparison? *Educational Research*, vol. 48, no 1, pp. 93–109. doi:10.1080/00131880500498511

Vainikainen M. P., Hautamäki J., Hotulainen R., Kupiainen S. (2015) General and Specific Thinking Skills and Schooling: Preparing the Mind to New Learning. *Thinking Skills and Creativity*, vol. 18, April, pp. 53–64. doi:10.1016/j.tsc.2015.04.006

Wang L., Shute V., Moore G. R. (2015) Lessons Learned and Best Practices of Stealth Assessment. *International Journal of Gaming and Computer-Mediated Simulations*, vol. 7, no 4, pp. 66–87. doi:10.4018/IJGCMS.2015100104

Watkins M. W., Lei P.-W., Canivez G. L. (2007) Psychometric Intelligence and Achievement: A Cross-Lagged Panel Analysis. *Intelligence*, vol. 35, no 1, pp. 59–68. doi:10.1016/j.intell.2006.04.005