International Journal of Instruction *January 2026 ● Vol.19, No.1*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 401-418

Article submission code: 20250519115953
Received: 19/05/2025 Revision: 00/00/2025
Accepted: 00/00/2025 OnlineFirst: 03/10/2025

Identification of Problematic Items and Analysis of Distractors of the van Hiele Geometry Test

Veronika Bočková
Constantinte the Philosopher University in Nitra, Slovakia, *vbockova@ukf.sk*

Gabriela Pavlovičová
Constantinte the Philosopher University in Nitra, Slovakia, *gpavlovicova@ukf.sk*

Ľubomír Rybanský
Constantinte the Philosopher University in Nitra, Slovakia, *lrybansky@ukf.sk*

A suitable tool for determining the level of geometric thinking is the van Hiele Geometry Test. In our research we quantitatively evaluate the success rate in solving individual items of the van Hiele Geometry Test and identify the reasons behind the difficulty of certain test items. In addition, we focus on modeling the nominal data of the test and analyzing the distractors in the test. Test was administered to a sample of 781 9th graders in Slovakia. We modeled the nominal data of the test with a nested 4PNL model and analyzed the distractors using characteristic curves. The research results prove that the van Hiele Test items are hierarchical and reflect the basic properties of van Hiele's theory. At each level of geometric thinking, test items 7, 9, 13 were found to be the most straightforward and test items 5, 10, 14, 19 were found to be the most challenging. Test items 14 and 19, which focused on connecting logic and geometry, had the lowest success rate of correct solutions, and their characteristic curves differed from those of the other items in the test. These items are also problematic in studies conducted in other countries and are considered the most challenging test items.

Keywords: geometry education, geometry thinking, van Hiele Geometry Test, elementary school, problematic items, distractors

INTRODUCTION

Geometric thinking plays a crucial role in the development of mathematical thinking. As stated by the National Council of Teachers of Mathematics (2000), geometry provides an aspect of mathematical thinking that is different but still connected to numbers. Through geometric thinking, we understand the ability of pupils to use geometric concepts not only in mathematics teaching but also in various areas of everyday life (Hardianti et al., 2017). According to Fisher (2015), geometric thinking depends on how the human mind can use the properties of geometric shapes and spatial relationships.

Citation: Bočková, V., Pavlovičová, G., & Rybanský, Ľ. (2026). Identification of problematic items and analysis of distractors of the van Hiele Geometry Test. *International Journal of Instruction*, *19*(1), 401-418.

There are various studies dealing with geometric thinking from different perspectives. One of the most important studies dealing with geometric thinking has been van Hiele's theory for several decades. Similarly, separate levels of geometric thinking were also determined by Stoljar (Gábor et al., 1989). Both authors' thinking levels are similar, but van Hiele and Stoljar differ on the age assignment to individual levels.

The most recognized theory on cognitive development in geometry was created by Dutch educators Pierre and Dina van Hiele. Their theory describes how students learn geometry, the progressive levels of geometric thinking, and potential challenges in the learning process. Additionally, it provides recommendations for fostering geometric thinking.

Van Hiele's theory consists of five levels of geometric thinking. The individual levels of geometric thinking can be briefly characterized as follows:

- Level 1 – Visualization: Students identify geometric shapes based on their complex visual perception or similarity to a known shape, the orientation of the shapes is dominant.
- Level 2 – Analysis: Students already know geometric shapes' properties and can create classes of geometric shapes based on their common properties. They define geometric shapes by listing all their properties, even those that are not necessary.
- Level 3 – Informal Deduction: Students are aware of the relationships between the properties of individual geometric shapes. They also know that the properties of the shape are arranged and interconnected. They can formulate correct abstract definitions, which are characterized by their economy.
- Level 4 – Formal Deduction: Students understand the logical system of geometry and deduction; they know why the axioms, sentences, and definitions are essential. They can prove the claims at the secondary school level.
- Level 5 – Rigor: Students can compare different axiomatic systems and they understand non Euclidean geometry. The students can use all types of proof. (Usisnkin, 1982)

Pupils generally have difficulty understanding the geometry curriculum and have only formal knowledge. This problem is also encountered by pupils in Slovakia, which is also evident from the results of national testing in the ninth year of primary school (Testing 9), where pupils typically achieve the lowest or second-lowest percentage success rate in geometry problems. Problems with success in solving geometric problems may be due to a low level of geometric thinking and mathematical competence. That was the reason why we decided to focus on researching the level of geometric thinking of pupils in Slovakia at the end of primary school. We focused not only on determining the level of geometric thinking of pupils, but also on identifying problematic test items and exploring possible causes in the context of the geometry curriculum in Slovakia. We used the van Hiele Geometry Test (hereinafter referred to as VHGT that is an instrument developed by Usiskin (1980) to measure geometric

thinking levels based on van Hiele theory. The analysis of distractors in the test was also important for us, so that we could assess the specifics of individual test items in the context of geometry teaching in Slovakia. The connection between geometric competencies and van Hiele's theory of geometric thinking is close and significant, as van Hiele's model provides a framework for understanding how pupils' geometric competencies develop over the course of education. Slovakia is currently undergoing a reform of the primary school curriculum, which opens up space for adjusting the content and methods of teaching mathematics. We consider identifying weaknesses in students' knowledge to be key in this process. We therefore set the following research objectives:

- to quantitatively evaluate the success of solving individual items of the van Hiele Geometry Test,
- to determine the reasons for some items' difficulty,
- to model the nominal data of the van Hiele Geometry Test and analyzed the distractors in the test.

EMPIRICAL RESEARCHES DEAL WITH THE VAN HIELE GEOMETRY TEST

The most important research dealing with VHGT is that of Usiskin (1982), who presents the results of a study conducted in the 1980s as part of the CDASSG project. The publication provides a detailed description of van Hiele's theory of geometric thinking, the individual levels of geometric thinking, and a proposal for the van Hiele Geometry Test and the determination of the level of geometric thinking of 2700 high school pupils. As stated by Senk et al. (2022), VHGT currently has three different uses in international research:

- to assess the van Hiele levels of samples of interest,
- to evaluate the effectiveness of some educational innovations,
- to select pupils with particular characteristics for further studies.

Research in determining the level of geometric thinking provides teachers with an overview of how pupils think in geometry at individual levels of education - primary, secondary, and higher education. For example, Levenson et al. (2011) and Clements &Sarama (2014) address geometric thinking in the preschool age. The geometric thinking of pupils in elementary schools is the subject of studies by various researchers: Halat (2006), Ma et al. (2015), MdYunus et al. (2019), Andini et al. (2018), and Hardianti et al. (2017). These studies consistently indicate that primary school students are typically at the levels of visualization and analysis. VHGT has been used in research in secondary schools by Haviger &Vojkůvková (2015), Alex &Mammen (2015), Naufal et al. (2021) and Kundu & Ghose (2016). Research here commonly shows that high school students typically achieve the levels of visualization, analysis, and informal deduction. Research also focuses on future mathematics teachers for secondary level in universities with studies such as Knight (2006) determining the level of geometric thinking of students at the University of Maine or Jupri (2018) elementary school teachers in Indonesia, Yilmaz &Koporan (2016) and Halat (2008) in Turkey, Armah et

al. (2017) in Ghana, Patkin &Burkai (2014) in Israel and Pavlovičová et al. (2022), Pavlovičová & Bočková (2021) in Slovakia. Many future and primary school mathematics teachers often do not reach the required higher levels of geometric thinking, specifically informal deduction, formal deduction, or rigour. VHGT is also a suitable tool of validating new teaching methods. As stated by Senk et al. (2022), VHGT is commonly employed in both pre-tests and post-tests to evaluate the effectiveness of educational interventions. MdYunus et al. (2019) confirmed that pupils who were taught van Hiele's theory using Google SketchUp achieved better results in geometry. Research by Adelabu et al. (2019) confirms that dynamic mathematical software significantly improves geometric thinking. Students who used GeoGebra software in their lessons were better able to use analysis and deduction when working with geometric shapes. Based on their study, Hardianti et al. (2017) proposed a POGIL model that increases pupils' geometric thinking levels. Hassan et al. (2020) proposed effective teaching strategies for secondary school pupils based on van Hieles' learning phases.

The selection of pupils based on the VHGT result was carried out, for example, by Astuti et al. (2018), who selected only 6 pupils from a sample of 38 respondents based on the test result; those were subsequently administered the geometry skill test and conducted personal interviews. Sofiyati (2022), in his research on critical thinking, selected pupils for the control and experimental samples based on the VHGT results. According to the level of geometric thinking, he selected 14 pupils, administered the critical thinking test and interviewed them to obtain more accurate and in-depth information. As stated by Senk et al. (2022), the particular method of using the VHGT is a concept that is not addressed in the test implementation system. Still, it is a suitable tool that helps select pupils for further in-depth research and qualitative evaluations.

In addition to identifying the level of geometric thinking, various studies have analyzed the VHGT items and their suitability in the test. The study by Senk et al. (2022) confirms that the VHGT considers the discreteness and hierarchy of the levels of geometric thinking, but some research reports overlapping item difficulties at similar levels. Chen et al. (2019) argue that the difficulty of individual items is related to changes in the curriculum since the creation of the VHGT. As per their findings, some items testing the Level 2 - analysis are too easy (item 7) or too complex (item 10) for a particular level. Similar findings are reported by Stols et al. (2015), who identified specific test items at the Level 2 - analysis (item 10), abstraction (item 14), and deduction (item 19) that were disproportionately difficult for the level they were testing. On the contrary, the items at the Level 2 - analysis (items 7 and 9) are too easy for the given level because they are similar in difficulty to the items at the Level 1 - visualization. However, an essential result of the studies is that the results confirm the sequential nature of the individual test items at the first three levels of geometric thinking (Haviger & Vojkuvková, 2015). Also, the first four levels (Chen at al., 2023), i.e. the sets of test items correspond to the sequence of van Hiele's theory.

It is important to note that the aforementioned research on VHGT item analysis has primarily been conducted in an international context, rather than within the Slovak educational system or based on the Slovak curriculum. We decided to quantitatively

evaluate the success of solving individual items of the van Hiele Geometry Test and determine the reasons for some items' difficulty because we had identified a research gap in Slovakia.

METHOD

Our research focused on analyzing the items of globally used van Hiele Geometry Test. In accordance with the first research objective, it was necessary to translate the VHGT into Slovak and verify its reliability and validity. Subsequently, we distributed the test to schools and quantitatively evaluated the students' solutions, identifying the easiest and most challenging items in the test. Based on these findings, we linked the problematic items with the content and methodology of teaching geometry in primary schools in Slovakia, thereby fulfilling the second research objective. We also statistically evaluated the test with regard to the analysis of distractors, and it was important to choose an appropriate statistical model for their evaluation. Through this analysis, we identified weaknesses in the test for Slovak students, which we compared with other foreign studies, thereby fulfilling the third objective of our research.

Research sample

The research sample consisted of 9th grade (15 years old) elementary school pupils in Slovakia. Their knowledge represents the output knowledge when graduating from elementary school, i.e., lower secondary education. Geometry and measurement are topics included in the curricula of each grade of elementary school. Within elementary school, pupils gradually expand their knowledge of geometry. In the ninth grade, all pupils in Slovakia take the national mathematics test, Testovanie 9, and the secondary school admission exams. As part of their preparation, they will review the elementary school geometry curriculum and consolidate their acquired knowledge. The research sample consisted of 781 9th grade pupils from 29 different elementary schools in 23 cities across Slovakia.

Research tool

The van Hiele Geometry Test, translated into Slovak, was used as a research tool. The test was used with the permission of its authors. The test was created by Usiskin (1980) from the University of Chicago based on the van Hiele model of geometric thinking and created as part of the CDASSG (Cognitive Development and Achievement in Secondary School Geometry) project to verify van Hiele's theory. The test contains a total of 25 test questions with a choice of five answers (A-E), and five questions for each of the five levels of geometric thinking. Items VH1 - VH5 are intended to determine the first level of geometric thinking, items VH6 -VH10 for the second level, and so on. The time to complete the test is 35 minutes. The test items are created so that the group corresponds to pupils' knowledge and skills at individual levels according to the van Hiele model (Knight, 2006). Due to the research sample, we tested only the first 20 items, as the last five tasks are at the university level.

Statistical methods

Each pupil possesses a certain level of geometric thinking, which, however, cannot be measured directly and is therefore considered a latent variable. The VGHT is a tool used to estimate the level of this latent variable. Modeling a continuous latent variable measured through binary (generally discrete) manifest variables employs methods derived from Item Response Theory (IRT). IRT is an umbrella term for a group of statistical techniques that identify and estimate the characteristics of items as well as respondents. For the case of dichotomous manifest variables or items, several unidimensional IRT models have been developed, such as the Rasch model, the 1PL (one-parameter logistic) model, the 2PL (two-parameter logistic) model, the 3PL model, and the 4PL model. Each of these models estimates the probability of selecting the correct response to an item, which depends on the respondent's latent trait level and the item parameters. For items with more than two response options, multiple polytomous unidimensional models have been formulated, including the Rating Scale Model (RSM), the Partial Credit Model (PCM), and the Generalized Partial Credit Model (GPCM). In IRT, item parameters and the latent variable are measured on the same scale, known as logits. A key advantage of the IRT approach is that item parameters and the respondent's latent trait level are estimated independently. Consequently, the estimation of the latent trait level is unaffected by the specific items, and the estimation of item parameters is not influenced by the respondents' latent trait levels.

The traditional approach to extracting information from distractors is to model nominal data with the Bock Nominal Response Model (hereafter referred to as NRM) (Bock, 1972), which is a multinomial adaptation of the 2PL IRT (Two Parameter Logistic Item Response Theory) model, where $P(X_{ij} = v|\theta_j)$ is the probability that respondent j will choose response category v (which can be the correct answer or a distractor) among m_i possible responses for item *i*.

This probability is modeled as a function of the respondent's latent trait level θ_j, the attractiveness parameter of response category *v* (high positive values correspond to attractive items) ξ_{iv}, and the slope parameter λ_{iv} of response category v for item *i*.

$$P(X_{ij} = v|\theta_j) = \frac{e^{\xi_{iv}+\lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik}+\lambda_{ik}\theta_j}}.$$

The disadvantage of this approach is that all response categories are viewed as mutually equivalent (nominal), which is also a source of the model name. However, when solving tasks with multiple choice, two cases may occur:

- the pupil understands the task correctly, or is able to solve the task and chooses the correct answer from the options - in this case, distractors are not feasible,
- the pupil cannot solve or understand the task correctly, and then the guessing is present (assuming that we do not consider possible penalties for an incorrect answer).

This sequential response selection procedure in an item was the motivation to create a new class of IRT models for items with the option of choosing the correct answer. For situations where the response selection is carried out sequentially, nested logit models (NLM) were created. The goal of NLM is to approximate the probability of the response in a sequential decision-making process by combining the best IRT model for each decision step into a single model. NLM has two levels dividing the set of responses into two nested groups. At a higher level (level 1), the model distinguishes between choosing the correct answer and choosing any incorrect answer (the incorrect answer has a code of 0 and the correct answer has a code of 1), which allows the use of binary logistic models (2PL, 3PL, 4PL). At a lower level (level 2), the model distinguishes the probability of selecting a particular distractor (versus another distractor) as the product of the probability of selecting any distractor (incorrect response) and the probability modeled using the propensities of each distractor, which is similar to NRM.

At the first level, items can be modeled with 2PL, 3PL or 4PL IRT models, which we briefly introduce. The probability of choosing the correct answer (category *u*) for the jth respondent in the i-th item is modeled in each. This probability depends on the latent trait level θj respondent and on the parameters of the *i*th item: α_i (discrimination parameter), β_i (difficulty parameter), γ_i (item guessing parameter, also called lower asymptote) and δ_i (upper asymptote) so that:

$$P(x_{ij} = u|\theta_j) = \gamma_i + \frac{\delta_i - \gamma_i}{1+e^{-(\beta_i+\alpha_i\theta_j)}} (4PL)$$

$$= \gamma_i + \frac{1-\gamma_i}{1+e^{-(\beta_i+\alpha_i\theta_j)}} (3PL)$$

$$= \frac{1}{1+e^{-(\beta_i+\alpha_i\theta_j)}}. \quad (2PL)$$

Although 4PL models are less commonly used than 2PL and 3PL models, Myszkowski and Storme (2017) found that 4PL models adequately correct for inattention errors and improve measurement efficiency (to avoid underestimating the latent trait level due to random error).

At the second level, where distractor modeling occurs, the probability $P(X_{ij} = v|\theta_j)$, that the *j*th respondent will choose distractor v from among m_i-*1* distractors in item *i* is modeled as the product of the probability of choosing the wrong answer $1 - P(x_{ij} = u|\theta_j)$ and the probability of choosing distractor *v*. This is where the NRM model is applied, where the tendency to choose a distractor is a function of the respondent's latent trait level θj, the attractiveness parameter ξiv, and the slope parameter λiv of the *i*-th item. The resulting model for the probability that the *j*-th respondent in the ith item chooses distractor *v* is

$$P(x_{ij} = v|\theta_j) = \left(1 - P(x_{ij} = u|\theta_j)\right) \cdot \frac{e^{\xi_{iv} + \lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik} + \lambda_{ik}\theta_j}} ..$$

If the 4PL model is used at the first level, then the model for distractors leads to the 4PNL model

$$P(x_{ij} = v|\theta_j) = \left(1 - \left(\gamma_i + \frac{\delta_i - \gamma_i}{1 + e^{-(\beta_i + \alpha_i\theta_j)}}\right)\right) \cdot \frac{e^{\xi_{iv} + \lambda_{iv}\theta_j}}{\sum_{k=1}^{m_i} e^{\xi_{ik} + \lambda_{ik}\theta_j}}.$$

Compared to the NRM model, the NLM model's probability of selecting a distractor is conditional on the probability of selecting an incorrect answer. In contrast, in the NRM model, the correct and incorrect answers interact with each other, i.e., they are viewed only as different levels.

To assess the overall fit of the considered models, we employed the chi-square goodness-of-fit test, the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). The chi-square goodness-of-fit test has the limitation that the null hypothesis of model adequacy is sometimes rejected even when the model is appropriate, particularly in the case of large sample sizes (Barrett, 2007). Both the CFI and TLI range from 0 to 1, with values greater than or equal to 0.95 indicating good model fit (Hu & Bentler, 1999). RMSEA values less than 0.08 suggest a good fit between the model and the data, while values less than 0.05 indicate an excellent fit (Browne & Cudeck, 1993). The Akaike Information Criterion (AIC) was used to select the best model among multiple candidate models. The advantage of AIC is that it considers not only the quality of the model's fit to the data but also the model's complexity, thereby helping to prevent overfitting. The model with the lowest AIC value is regarded as the best, as it optimally balances model simplicity and fit to the data.

All analyses were performed using the freely distributable program R (R Core Team, 2018) using the libraries ltm (Rizopoulos, 2006), mirt (Chalmers, 2012), KernSmoothIRT (Mazza et al., 2014), VIM (Kowarik & Templ, 2016), jrt (Myszkowski, 2021), ggplot2 (Wickham, 2016).

FINDINGS

As first, we evaluated the pupils´ success when solving the test. 558 (71%) pupils answered all 20 items. At least one missing answer occurred in 223 (29%) pupils, 40 (5%) pupils not answering two items, 24 (3%) pupils not answering three items, and five pupils (0.6%) not answering any of the items of the VHGT. The total number of missing answers was 268, which represents 1.7% of all answers.

Chart 1 presents the percentage success rate of solving individual VHGT items. The items are divided into four groups according to the level of geometric thinking they test. As visible in Chart 1, the success rate of solving items in groups gradually decreases with an increasing level of geometric thinking required by the item. In each group of items, one has a significantly lower success rate than the other items.

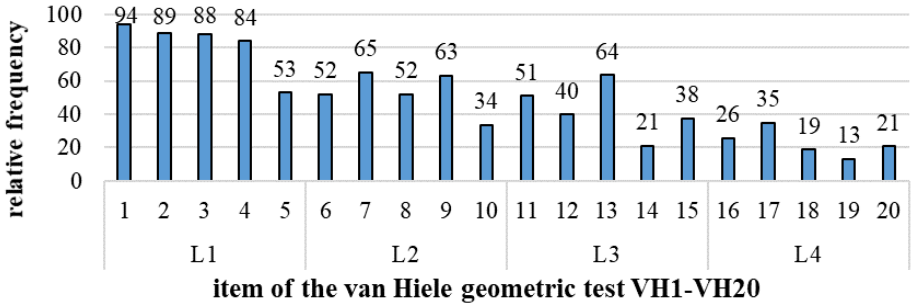

Chart 1
Percentage success rate of solving VHGT items

For pupils, the most challenging item in the first group of items at the Level 1 – visualization was item VH5. This is visible from the success rate of 53.1%. In the item, pupils had to choose from the figures those that are parallelograms. Pupils are familiar with square, rectangle and triangle concepts since preschool. Therefore, they had only slight issues with items VH1 - VH4. According to the State Educational Program in the Slovak Republic, a parallelogram is introduced only in the eighth grade of elementary school. The low success rate of solving the item VH5 indicates that pupils have not sufficiently mastered the concept of a parallelogram. Pupils had difficulty determining that the required figure, that was rotated by 45° in VH5 was a parallelogram.

At the Level 2 – Analysis, the most challenging item for pupils was item VH10, which was solved correctly by only 33.5% of pupils. The success of the items again reflects the acquired knowledge of the pupils. In items VH6 – VH9, they decided on the truth of statements resulting from the properties of a square (VH6), a rectangle (VH7), a triangle (VH8) and a parallelogram (VH9). In item VH10, the students had to decide on the truth of a statement resulting from the properties of a circle, its radius, a central angle and a chord. The curriculum about the circle is included in the 8th grade, which is later compared to other shapes. Therefore, the cause of the poor success of the item VH10 may be insufficient anchoring of knowledge about the circle.

In the third group of items at the Level 3 – Informal Deduction we can see that VH14 was the most challenging for pupils. Only 21.1% of pupils solved it correctly. In items VH11, VH12 and VH15, pupils had to decide on the truth of statements about the relationships between the properties of two geometric shapes. Pupils were able to imagine the individual properties of two geometric shapes and then connect them. In VH13, pupils had to determine which shape was a rectangle, which was a very easy task. Item VH14 dealt with the creation of subgroups of individual geometric shapes. In mathematics classes, students are often unaware of the hierarchical arrangement of individual geometric shapes, such as quadrilaterals, they acquire knowledge about individual geometric concepts in isolation without any connection to each other.

At the Level 4 – Formal Deduction, VH19 was the most challenging item. Only 13.2% of pupils solved this item correctly. They had to decide on the truth of statements in connection with defining terms and proving statements. Pupils in the second stage of

elementary school are not familiar with the required knowledge and could only intuitively deduce the correct solution. Items at the Level 4 were difficult and not appropriate to the pupils' knowledge. Therefore, we could observe low success rates in this group of items. It also confirms that items at Level 4 in the VHGT are suitable for testing students' secondary school mathematics knowledge.

Table 1 shows the frequencies and relative frequencies (data in parentheses) for the VHGT items for the entire research sample. Bold highlights the data regarding the correct answer in the respective item. As shown in the table, in items VH14, VH18, VH19 and VH20, the probability of marking a distractor is greater than marking the correct answer. Table 1 also provides information on all distractors that pupils most often marked in each item.

Table 1
Frequencies and relative frequencies for VHGT items for the entire respondent set

	Answer					
Item	A	B	C	D	E	NA
VH1	9 (1%)	**733 (94%)**	6 (1%)	20 (3%)	5 (1%)	8 (1%)
VH2	4 (1%)	14 (2%)	37 (5%)	**692 (89%)**	16 (2%)	18 (2%)
VH3	57 (7%)	1 (0%)	**689 (88%)**	14 (2%)	6 (1%)	14 (2%)
VH4	28 (4%)	**656 (84%)**	25 (3%)	26 (3%)	21 (3%)	25 (3%)
VH5	83 (11%)	35 (4%)	198 (25%)	35 (4%)	**415 (53%)**	15 (2%)
VH6	90 (12%)	**407 (52%)**	168 (22%)	68 (9%)	16 (2%)	32 (4%)
VH7	112 (14%)	28 (4%)	70 (9%)	45 (6%)	**510 (65%)**	16 (2%)
VH8	**406 (52%)**	70 (9%)	100 (13%)	88 (11%)	82 (10%)	35 (4%)
VH9	64 (8%)	47 (6%)	**493 (63%)**	51 (7%)	110 (14%)	16 (2%)
VH10	83 (11%)	141 (18%)	120 (15%)	**262 (34%)**	129 (17%)	46 (6%)
VH11	83 (11%)	126 (16%)	**398 (51%)**	74 (9%)	76 (10%)	24 (3%)
VH12	146 (19%)	**311 (40%)**	101 (13%)	83 (11%)	88 (11%)	52 (7%)
VH13	**496 (64%)**	39 (5%)	80 (10%)	106 (14%)	45 (6%)	15 (2%)
VH14	**165 (21%)**	153 (20%)	109 (14%)	111 (14%)	202 (26%)	41 (5%)
VH15	128 (16%)	**293 (38%)**	106 (14%)	123 (16%)	96 (12%)	35 (4%)
VH16	144 (18%)	122 (16%)	**202 (26%)**	141 (18%)	105 (13%)	67 (9%)
VH17	147 (19%)	131 (17%)	**274 (35%)**	104 (13%)	92 (12%)	33 (4%)
VH18	162 (21%)	190 (24%)	132 (17%)	**148 (19%)**	90 (12%)	59 (8%)
VH19	313 (40%)	132 (17%)	123 (16%)	**103 (13%)**	73 (9%)	37 (5%)
VH20	**163 (21%)**	105 (13%)	107 (14%)	253 (32%)	115 (15%)	38 (5%)

Table 1 in column NA illustrates the frequencies (relative frequencies) of pupils not providing the answer to a particular item. Most often, pupils gave answers to all items at the Level 1 –visualization (VH1 – VH5), with a maximum of 3% missing answers. Pupils most often did not provide an answer to items at the Level 3 and Level 4: VH16 (9%), VH18 (8%) and VH12 (7%).

In addition to the quantitative analysis, we were interested in what model is appropriate to model the nominal data of the VHGT test. Using the EM algorithm in the mirt program package, we estimated four models for nominal data – a model for nominal categories (NRM), a 2-parameter nested logit model (2PNL), a 3-parameter nested logit model (3PNL) and a 4-parameter nested logit model (4PNL). All models converged

successfully. However, the information matrix for the 4PNL model could not be converted, which made it impossible to calculate the standard errors of the parameter estimates.

The model fit indices of the four models are listed in Table 2. In terms of the CFI criterion, the minimum acceptable value of which is 0.95, we conclude that only the 4PNL model is sufficiently suitable. According to the TLI criterion (the minimum acceptable value is 0.95), none of the models is sufficiently suitable. All models achieve better than the maximum acceptable value of 0.06 for the RMSEA index. The Akaike Information Criterion (AIC) value is the lowest for the 4PNL model, which means it is the best among the models considered (respecting this criterion). The likelihood ratio test (LR test) shows that the 4PNL model is better than the 3PNL model ($\chi^2(20) = 39, 22$; $p = 0{,}006$), at the same time, the 3PNL model and the 2PNL model are not statistically significantly different ($\chi^2(20) = 31, 31$; $p = 0{,}051$), the 4PNL model is better than the 2PNL model ($\chi^2(40) = 71, 53$; $p = 0{,}016$) and the 4PNL model is also better than the NRM model ($\chi^2(8) = 73, 50$; $p = 0{,}001$). It is clear that among the considered models, the 4PNL is the best, and except for the TLI criterion, we state that it is good agreement with the data.

Table 2
Fit parameters of nested IRT models

Model	χ2	df	p	CFI	TLI	RMSEA	AIC
NRM	99.10	52	<0.001	0.864	0.658	0.039	26075.79
2PNL	96.13	52	<0.001	0.897	0.779	0.039	25748.89
3PNL	64.82	32	<0.001	0.923	0.733	0.042	25746.12
4PNL	25.60	12	.012	0.968	0.705	0.045	25739.02

Empirical reliability for all models reaches a satisfactory value (greater than 0.7), with the highest 0.82 for the 4PNL model, and for the remaining models as follows: 0.80 for the 3PNL model, 0.77 for the 2PNL model and 0.73 for the NRM model.

The suitability parameters of the nested IRT models show that the 4PNL model is suitable for modeling nominal data, which also corrects errors due to inattention and improves measurement efficiency compared to the other analyzed models. The 4PNL model was further analyzed by the nominal test data.

The interpretation of the model parameters, which are listed in Table 2, is best presented through the characteristic curves of the items listed in Figure 1. For example, the course of the characteristic curves can be explained in item VH1. In a particular item, regardless of the latent trait level (geometric thinking), the most likely option is to indicate the correct answer (option B, orange curve). In contrast, for respondents with a latent trait level greater than -1.0 logit, this probability is practically equal to 1. For respondents up to the -1.0 logit level, the most likely option is distractor D (red curve), the second most likely option is distractor A (blue curve), and the least likely are distractors C (green curve) and E (purple curve).

The characteristic curves for correct answers have an increasing character; with an increasing level of geometric thinking, the probability of a correct answer increases. Looking at the characteristic curves belonging to correct answers, we realized that all

items have an increasing character, except for item VH19. The characteristic function for item VH19 is distinctly decreasing. This finding signifies that with an increasing level of the latent trait, the probability of a correct answer (option D) decreases. This unequivocally indicates that item VH19 is fundamentally flawed, as it incorrectly rewards lower-ability students over those with higher abilities. Another significantly problematic item is VH14, whose characteristic function, though non-decreasing, presents a severe issue. For respondents with a high latent trait level, the correct answer (option A) is not the most probable. Instead, option E (a distractor) shows a higher probability, indicating a serious flaw either in the design of the item's distractors or in its overall wording and clarity."

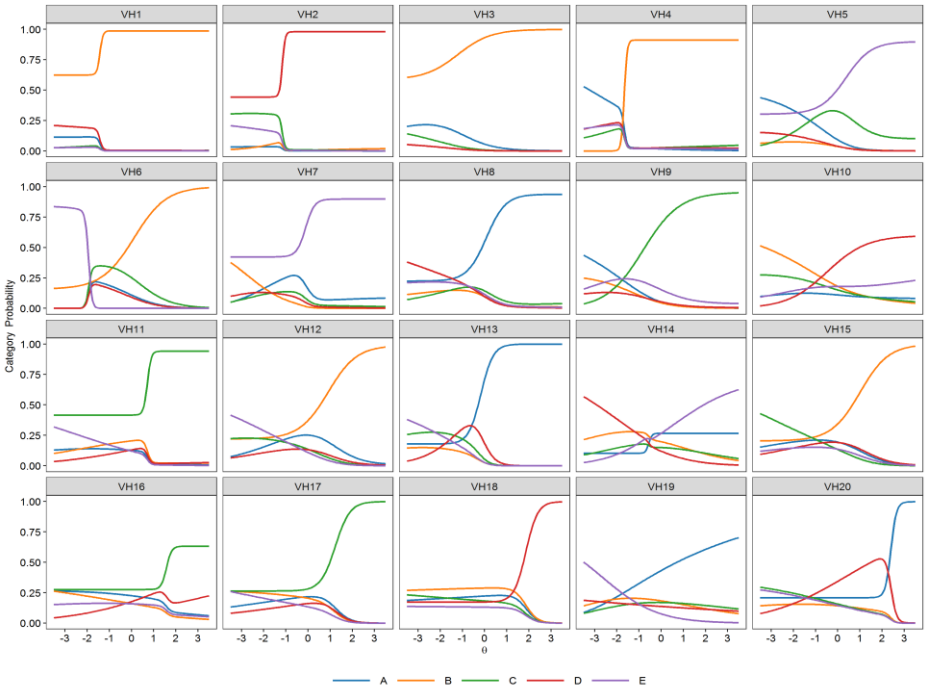

Figure 1
Item characteristic curves in a 4-parameter nested (4PNL) model. (Correct answers: 1b, 2d, 3c, 4b, 5e, 6b, 7e, 8a, 9c, 10d, 11c, 12b, 13a, 14a, 15b, 16c, 17c, 18d, 19d, 20d)

DISCUSSION AND CONCLUSION

The success of solving individual VHGT items shown in Chart 1 confirms the hierarchical classification of levels of geometric thinking, since the success of solving items at individual levels as a unit always has a decreasing character. With the gradually increasing difficulty required by individual levels of geometric thinking, the success of solving many items decreases. Except of the items VH7, VH9, VH13, which are less challenging for pupils. The reason can be found in the methodology of teaching

geometry in Slovakia. These items test knowledge from the standard geometry curriculum about the properties of geometric shapes in elementary school.

In each group of five items testing different levels of geometric thinking, the challenging items with significantly lower success rate were identified. The challenging items were VH5 at the Level 1 – Visualization and VH10 at the Level 2 – Analysis. The lowest success rate of these two items was related to the gradation of items within one level and the methodology of teaching geometry in Slovakia. Children in kindergarten can distinguish squares, triangles and rectangles based on visual perception (items VH1 - VH4). Still, pupils become familiar with the concept of a parallelogram (VH5) visually and through the properties of the shape only in the 8th grade. The pupil can solve items VH6 - VH9 directly based on the acquired knowledge about a square, rectangle, rhombus and isosceles triangle. Item VH10 combines several properties of the geometric shapes. The challenging items were item VH14 at the Level 3 – Informal Deduction and item VH19 at the Level 4 – Formal Deduction. The success rate of these items have to also be linked to the analysis of distractors, as the probability of marking the correct answer was not directly related to the level of geometric thinking.

We also modeled the nominal data of the VHGT results with an appropriate nested IRT model. The model fit indices show that the 4PNL model is suitable for modeling nominal data, which corrects pupils' errors due to inattention and improves measurement efficiency compared to the other analyzed models. The characteristic curves of the test items show how choosing the correct answer with increasing item number required a higher latent variable level, i.e., geometric thinking. This phenomenon can be observed in eighteen characteristic curves. The characteristic curve of item VH14 differs from the other curves in that, although it is non-decreasing, the correct answer is not the most probable. The characteristic curve of item VH19 is entirely different from the other characteristic curves. As the level of geometric thinking increases, the probability of the correct answer decreases.

There are relatively few tasks similar to VH14 in geometry curriculum in Slovak textbooks. Most tasks are focused on memorizing the curriculum, without deeper understanding, which can lead to formal pupils´ knowledge in geometry. This problem can be solved by using appropriate didactic activities that support thinking at higher levels of geometric thinking. Specific examples of didactic activities that address these geometric challenges are further explored in the article by Bočková et al. (2024)

Item VH19 is closely tied to the axiomatic construction of Euclidean geometry, an abstract and deductive method of teaching that requires a more advanced level of knowledge and skills than those developed in elementary school. In this item, pupils were asked to choose one of four possible answers, each formulated as a quantified statement involving the concepts of definition, assertion, proof, and truth—terms they had not previously worked with in elementary school. The Slovak curriculum places significant emphasis on the identification, properties and measurement of geometric shapes, encompassing the calculation of areas, surface areas and volumes of various geometric figures or the application of the Pythagorean theorem. The curriculum is oriented towards fostering visual and intuitive understanding. Since neither the

Euclidean construction of geometry nor propositional logic is included in the geometry curriculum at the elementary school level in Slovakia or many other countries, the outcome of this task was not unexpected.

The same problematic items are also reported in a study by Stols et al. (2015), who used the Rasch analysis in his research using the Rasch model. Stols et al. (2015), in their study, state that item VH14 is more challenging than other items for determining the Level 3 – Informal Deduction. The item is focused on class inclusion. However, VH14 does not explain the choice of definitions behind the class inclusion questions. Adding the definitions could improve the particular test item. According to Stols et al. (2015), item VH19 was the most challenging for pupils since the formal knowledge required to complete the item is not part of the pupils' geometry course. The pupils only guessed the answer to the particular item - the solution's success did not depend on the pupils' geometric thinking.

Chen et al. (2019) used standard test theory and cognitive diagnostic modeling to compare VHGT classification criteria. In both assessments, item VH19 was the most challenging question in the test and, together with other items at Level 4, had a high curve estimate. Chen et al. (2023) found through standardized factor loadings that items VH16 and VH19 of the VHGT may not be able to measure what they are supposed to measure, both had low proportions correct item VH19 was the most difficult even in this assessment method. The difficulty of items VH14 and VH19 of the VHGT is also confirmed by Wilson (1990). In the research, he used rough approximations - logit bias and standardized bias. The research does not provide more detailed reasons why the items are difficult for pupils.

According to van Hiele´s theory of geometric thinking, it is essential to remember that pupils' geometric thinking is influenced by their geometric experiences, the extent and depth of geometric education embedded in the curricula of different countries, which can vary significantly. In addition, the approach and methodology of teaching geometry, including the methods and forms of teaching used, play a significant role, which not only differs in individual countries but also depends on the approach of the mathematics teacher. However, our findings are consistent with findings in other countries (Stols et al.2015; Chen et al., 2019; Wilson, 1990).

Our study focused on analyzing pupils' responses to the Van Hiele Geometry Test (VHGT) in relation to the hierarchical model of geometric thinking levels as defined by the van Hiele theory. The results confirmed a decreasing success rate in solving items corresponding to higher levels of geometric thinking, consistent with the expected hierarchy. Exceptions were observed in items VH7, VH9, and VH13, which assess knowledge from the standard elementary geometry curriculum in Slovakia and were therefore less cognitively demanding for pupils. The most problematic items were VH5, VH10, VH14, and VH19. Their lower success rates are linked to the geometry teaching methodology in Slovakia and the limited presence of cognitively demanding tasks in mathematics textbooks. However, success was not determined solely by the level of geometric thinking, distractors within the test items also had a significant influence. To model the nominal response data, a nested Item Response Theory (IRT) model was

applied. The four-parameter nominal logistic model (4PNL) proved to be the most suitable, effectively accounting for inattentive errors and improving measurement precision. Special attention was given to items VH14 and VH19, which integrated logical and geometric reasoning. These items had the lowest success rates and showed atypical item characteristic curves: item VH14 had a non-decreasing curve where the correct answer was not the most probable, and in item VH19, the probability of a correct response decreased with increasing geometric thinking level. The findings highlight the complexity of assessing higher-order cognitive processes in geometry and point to the need to enrich geometry teaching with didactic activities supporting higher levels of thinking and systematic work with abstraction and logic.

Investigating the reasons for pupils' failure in items VH14 and VH19, which appeared in several countries' research, could be a topic for further research in this area. It could be interesting to compare the contents, methods and forms of geometry teaching in individual educational systems, as well as to compare curriculum, textbooks and methodologies for teaching quadrilaterals.

ACKNOWLEDGEMENTS

This work was supported by Scientific Grant Agency Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences VEGA 1/0532/23 Mathematical competences in the field of geometry of pupils at the end of lower middle education.

REFERENCES

Adelabu, F., Makgato, M., & Ramaligela, M.S. (2019). Enhancing Learners' Geometric Thinking Using Dynamic Geometry Computer Software. *Journal of Technical Education and Training*, *11*(1), 44 – 53.

Alex, J. K., & Mammen, K. J. (2014). Gender Differences amongst South African Senior Secondary School Learners' Geometric Thinking Levels. *Mediterranean Journal of Social Sciences, 5*(20), 1908 – 1915.

Andini, S., Fitriana, L., & Budiyono, B. (2018). Elementary School Students Visual Spatial Comprehension Based on van Hiele Theory: The Case in Madiun, East Java, Indonesia. *Journal of Physics: Conference Series, 983*(1), 1 – 6.

Armah, R. B., Cofie, P. O., & Okpoti, C. A. (2017). The geometric thinking levels of pre-service teachers in Ghana. *Higher Education Research*, *2*(3), 98-106.

Astuti, R., Suryadi, D., & Turmudi, T. (2018). Analysis on geometry skills of junior high school students on the concept congruence based on Van Hiele's geometric thinking level. *Journal of Physics: Conference Series*, *1132*(1), 1-5.

Barrett, P. (2007). Structural Equation Modeling: Adjudging Model Fit. *Personality and Individual Differences*, *42*, 815-824.

Bočková, V., Pavlovičová, G., & Rumanová, L. (2024). Support of Geometric Thinking by Using Ict in the Teaching of Elementary Mathematics, EDULEARN24 Proceedings, pp. 671-679.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K.A., & Long, J.S. [Eds.] Testing structural equation models. Newbury Park, CA: Sage, 136–162.

Chalmers, R. P. (2012). *mirt*: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1 – 29.

Chen, Y., Senk, S. L., Thompson, D. R., & Voogt, K. (2019). Examining Psychometric Properties and Level Classification of the van Hiele Geometry Test Using CTT and CDM Frameworks. *Journal of Educational Measurement*, *56*, 5.

Chen, Y.-H., Hsu, CH.-L., Wu, Y.-J., Yi, Z., Wang, Y., & Thomson, D. R. (2023). Exploring Attribute Hierarchies of the van Hiele Theory Using Diagnostic Classification Modeling and Structural Equation Modeling. University of Chicago School Mathematics Project.

Clements, D. H., & Sarama, J. (2014). *Learning and Teaching Early Math: The Learning Trajectories Approach*. New York: Routledge. 344 p.

Fisher, J. (2015). *Geometric Thinking Concept Map.* Assessment Resource Banks. [online].

Gábor, O., Kopanev, O., & Križalkovič, R. (1989). *Teória vyučovania matematiky pre študentov matematiky učiteľského štúdia na univerzitách a pedagogických fakultách*. Bratislava: SPN. 321 p.

Halat, E. (2006). Sex-Related Differences in the Acquisition of the van Hiele Levels and Motivation in Learning Geometry. *Asia Pacific Education Review, 7*(2), 173 – 183.

Hardianti, D., Priatna, N., & Priatna, A. (2017). Analysis of Geometric Thinking Students' and Process Guided Inquiry Learning Model. *Journal of Physics: Conference Series, 895*(1), 1 – 7.

Hassan, M. N., Abdullah, A. H., & Ismail, N. (2020). Effects of Integrative Interventions with van Hiele Phase on Students' Geometric Thinking: A Systematic Review. *Journal of Critical Reviews*, *7*(13), 1133 – 1140.

Haviger, J., & Vojkůvková, I. (2015). The van Hiele Levels at Czech Secondary Schools. *Procedia - social and behavioral sciences*, *171*, p. 912 – 918.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Jupri, A. (2018). Using the Van Hiele theory to Analyze Primary Schol Teachers' Written Work on Geometrical Proof Problems. In *4th International Seminar on Mathematics, Science, and Computer Science Education.* Bandung: Curran Associates, p. 735 – 740.

Knight, K. CH. (2006). *An Investigation into the Change in the van Hiele Level of Understanding Geometry of Pre-service Elementary and Secondary Mathematics Teachers*: diploma thesis. Maine: Univerzity of Maine.

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(7), 1 – 16.

Kundu, A., & Ghose, A. (2015). Van Hiele levels of geometry thinking among H.S. students from different streams of study. *Journal of Education, 9,* 119-124.

Levenson, E., Tirosh, D., & Tsamir, P. (2011). *Preschool Geometry. Theory, Research, and Practical Perspectives.* Rotterdam: Sense Publishers.134 p.

Ma, H. L., Lee, D. C., Lin, S. H., & Wu, D. B. (2015). A Study of Van Hiele of Geometric Thinking among 1st through 6th Graders. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(5), 1181 – 1196.

Mazza, A., Punzo, A., & Mcguire, B. (2014). *KernSmoothIRT*: An R Package for Kernel Smoothing in Item Response Theory. *Journal of Statistical Software*, *58*(6), 1 – 34.

Md Yunus, A. S., Ayub, A. F. M., & Hock, T.T. (2019). Geometric Thinking of Malaysian Elementary School Students. *International Journal of Instruction*, *12*(1), 1095 – 1112.

Myszkowski, N. (2021). Development of the R Library jrt: Automated Item Response Theory Procedures for Judgment Data and their Application with the Consensual Assessment Technique. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(3), 426 – 438.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston: National Council of Teachers of Mathematics. 419 p.

Naufal, M. A., Abdullah, A. H., Osman, S., Abu, M. S., & Ihsan, H. (2021). The effectiveness of infusion of metacognition in van Hiele model on secondary school students' geometry thinking level. *International Journal of Instruction*, *14*(3), 535-546.

Patkin, D., & Barkai, R. (2014). Geometric Thinking Levels of Pre- and In-Service Mathematics Teachers at Various Stages of Their Education. *Hong Kong Educational Research Association*, *29*(1), 30 – 49.

Pavlovičová, G. & Bočková, V. (2021). Geometric Thinking of Future Teachers for Primary Education—An Exploratory Study in Slovakia. *Mathematics*, *9*(23), 2992-3006.

Pavlovičová, G., Bočková, V., & Laššová, K. (2022). Spatial Ability and Geometric Thinking of the Students of Teacher Training for Primary Education. *TEM journal, 11*(1), 388-395.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: Austria.

Rizopoulos, D. (2006). *ltm*: An R Package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1 - 25.

Senk, S., Thompson, D. R., Chen, Y. H., Voogt, K., & Usiskin, Z. (2022). *The Van Hiele Geometry Test: History, Use, and Suggestions for Revisions*.

Sofiyati, E. (2022). Critical Thinking Process Analysis Based on Van Hiele's Theory Through the Discovery Learning Model. *Pasundan Journal of Mathematics Education: Jurnal Pendidikan Matematika*, *12*(1), 44-59.

Stols, G., Long, C., & Dunne, T. (2015). An Application of the Rasch Measurement Theory to an Assessment of Geometric Thinking Levels. *African Journal of Research in Mathematics, Science and Technology Education*, *19*(1), 69-81.

Usiskin, Z. (1982). *Van Hiele Levels and Achievement in Secondary School Geometry*. CDASSG Project.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. 276 p. ISBN 331924275X.

Wilson, M. (1990). Measuring a van Hiele Geometry Sequence: A Reanalysis. *Journal for Research in Mathematics Education*, *21*(3), 230-237.

Yilmaz, G. K., & Koparan, T. (2016). The Effect of Designed Geometry Teaching Lesson to the Candidate Teachers' Van Hiele Geometric Thinking Level. *Journal of Education and Training Studies*, *4*(1), 129-141.