DEVELOPMENT OF A FOUR-TIER INSTRUMENT OF ACID-BASE PROPERTIES OF SALT SOLUTION

Habiddin[1]*, Devita Novi Ameliana[1], Muhammad Su'aidy[1]

[1] Department of Chemistry, Universitas Negeri Malang, Jl. Semarang No. 5 Malang, Indonesia

**Corresponding author: habiddin_wuni@um.ac.id*

Abstract. *Many chemistry teachers claimed that identifying students' understanding genuinely is challenging due to the limited time available, and the overload works faced. Therefore, constructing an efficient and valid instrument to overcome this difficulty is a must. This study is a part of the big project to produce a robust four-tier instrument to identify students' conception of acid-base properties of salt solution (FTI-ABPS2). The project involved six groups of researchers, and each group carried out the study in 6 different participant groups and schools. This paper outlines the development and validation of a four-tier instrument of Acid-Base properties of salt Solution involving public secondary school students in Malang Regency. The instrument was constructed using the procedure as carried out by Habiddin & Page (2019). Twenty-eight questions were constructed for the instrument. The results show that the instrument is valid and reliable to be implemented further. In-depth analysis regarding content validity, as well as the empirical validity of the instrument in light with the characteristic of chemical concepts (acid-base properties of salt solution in particular), is also discussed.*

Keywords: *four-tier instrument, salt hydrolysis, acid-base properties of salt solution, students' understanding*

INTRODUCTION

Acid-base properties of the salt solution have been an issue for many students (Febriani et al., 2018; Nimmermark et al., 2016; Nyachwaya, 2016; Romine et al., 2016) that can lead to the obstacle of teaching and learning in this topic. Knowledge regarding these difficulties is timely in order to design proper and effective teaching and learning. Therefore, students' misunderstanding should be identified before they embark to the chemistry teaching and learning in the relevant topics. However, uncovering students' difficulty, particularly misconceptions using a common instrument such as multiple-choice questions, is a tricky exercise (Habiddin & Page, 2019). Recently, multi-tier instruments including three-tier (Caleon & Subramaniam, 2010a; Kirbulut, 2014; Milenković et al., 2016; Pesman & Eryilmaz, 2010) and four-tier instruments (Caleon & Subramaniam, 2010b; Habiddin & Page, 2019; Sreenivasulu & Subramaniam, 2014; Yang & Lin, 2015) have been applied in science education disciplines. This study aimed to develop and validate a four-tier instrument to identify students' conception in the topic of acid-base properties of salt solution.

METHOD

Sample and procedures of the study

It has been previously stated that this is part of a single future involving six research groups, including Husniah, Habiddin, Sua'idy, & Nuryono (2019), **Journal of Disruptive Learning Innovation, Vol. 1 No. 1**. Each of the groups carried out the same study in 6 different schools. The results of these studies will be later combined and are supporting each other to do a single further study. This study involved two groups of students (one the first data collection using the preliminary instrument and another on the second data collection using the four-tier instrument), which are 28 and 29 students of two different classes at a public secondary school in Malang Regency. The school is located in the countryside of Malang and categorized as a rural area school.

The procedure of this study adopted the one developed by Habiddin & Page (2019), as explained below.

1. Mapping concept
   In this step, the chemistry curriculum, particularly in the topic of acid-base for the secondary school in Indonesia, was analyzed to extract the significant concepts to be involved in questions. The concepts identified were a basis in constructing the preliminary instrument.
2. Testing
   This step was initiated by the development of multiple-choice questions with a free-response as the preliminary instrument by considering the identified concept in the previous step. The multiple-choice questions with the free-response is a common multiple-choice question test accompanied by a requirement for participants to explain the reason behind their choice among the options provided in the multiple-choice question. The preliminary instrument presented in the Indonesian language was applied to the first group.
3. Defining students' unscientific ideas
   the responses of the first group participant to the preliminary instrument was analyzed. Several students' unscientific ideas were uncovered.
4. Developing the prototype four-tier
   students' unscientific ideas found in the previous step were used to create the prototype four-tier instrument. Twenty-eight questions (items) in the form of four-tier questions were produced in this step. All the questions were in the Indonesian language. The four-tier instrument consists of multiple choice question (1^{st} tier/ Answer tier/ A-tier), the confidence rating of A tier (2^{nd} tier), the reason for students' answer to the 1st tier (3^{rd} tier/Reason tier/R tier) and the confidence rating of R tier (4^{th} tier). The example of a question is attached in Appendix 1.
5. Validating the prototype four-tier.
   The prototype of the four-tier instrument produce in the previous step was validated both in terms of content validity and empirical validity. Feedbacks from a chemistry faculty wasobtained as a content validity procedure. The empirical validity involved the second group of the participant to analyze the quality of the instrument in term of validity, reliability, difficulty index, discriminatory index, and the effectiveness of distractor.
6. Refining the final four-tier.
   In this step, the four-tier instrument was revised based on the result of empirical validity in the previous step. In this step, the Four-Tier Instrument of Acid-Base Properties of Salt Solution (FTI-ABPS2) was produced.

Data Analysis

Students' responses to the preliminary instrument (first data collection) were analyzed descriptively in light of the scientific concepts of acid-base properties of salt solution. Students' responses were coded and categorized into three categories, including scientific understanding, misunderstanding and random errors. Those scientific understanding responses are students' answers which are scientifically correct as approved by the scientific community. The misunderstanding responses are students' answers, which are inconsistent with the concepts approved by the scientific community. These misunderstandings are the basis for constructing the R tier. Random errors are students' answers which are difficult to be interpreted and can be attributed as the result of guessing.

Students' responses to the four-tier instrument (second data collection) were graded according to the following procedure. Score '1' was attributed to students' correct answers both in the A and R tiers. Score '0' was attributed to either or both students' responses to the A and R tiers are incorrect. Students' scores obtained from this grading procedure are the basis for measuring the quality or for validating the quality of the prototype FTI-ABPS2 in terms of validity, reliability, difficulty level, discriminatory index and distractor effectiveness. The procedure to measure these parameters are described below.

Difficulty Level (DL)

The difficulty level (DL) represents the number of students answering the question correctly (Allen & Yen, 2002). The formula to calculate this parameter is:

$$\text{Dif iculty level} = \frac{N_c}{N}$$

where, N_c = the number of students who answer correctly;
N = the total number of students who participated in this study

Equation 1. Equation for DL

The results of a calculation using this formula are interpreted according to the following criteria.

Table 1. The interpretation criteria of DL

Difficulty level	Criteria
0.00 – 0.30	Hard
0.31 – 0.70	Moderate
0.71 – 1.00	Easy

(Arikunto, 1993)

Discriminatory Index (DI)

The discriminatory index (DI) explains an item's ability to differentiate between students who understand the concept being tested and those who do not understand it. The formula to calculate this parameter is presented below.

$$\text{Discriminatory index} = \frac{N_u - N_l}{\frac{1}{2}N}$$

Where, N_u = the number of the upper group who answer correctly
N_l= the number of the lower group who answer correctly
N = the total number of students who participated

Equation 2. The equation for DI (Allen & Yen, 2002).

The results of a calculation using this formula are interpreted according to the following criteria.

Table 2. The interpretation criteria of DI

DI index	Category
0.00 - < 0.10	poor
0.10 - <0.30	fair
0.30 - < 0.75	good
0.75 - 1.00	excellent
Discrimination index negative	unsuitable item

(anonymous, n.d.)

Distractor Effectiveness

A distractor, an incorrect answer provided in a multiple-choice question, should be selected by at least one testee/students so that it can be claimed as an effective distractor (DiBattista & Kurzawa, 2011).

Validity and Reliability

The product-moment correlation is employed to calculate the validity of items in this study. The degree of validity increase with the increase in the r_{xy}-calculation value. Internal-Consistency reliability is applied in this study. The results of a calculation using this formula are interpreted according to the following criteria.

Table 3. The interpretation criteria of reliability

Reliability	Interpretation
0.90 and above	Excellent reliability
0.80 - 0.90	Very good
0.70 - 0.80	Good
0.50 - 0.60	Fair, revision is needed
< 0.5	Poor

RESULT AND DISCUSSION

Content Validity of The Prototype FTI-ABPS2

Data obtained in this procedure are in the form of validators' feedbacks regarding the content (acid-base properties of salt solution), the pedagogical aspect and also the language. This feedbacks are the basis for revising and improving the quality of the prototype before it is proceed to the testing of empirical validity. A chemistry faculty was the validator of this product. Table 4 below summarizes the quality of the prototype FTI-ABPS2 referring to the validator's feedback.

Table 4. Validator's Feedbacks of The Prototype FTI-ABPS2 items

Item	Criteria	Note	Item	Criteria	Note
1	Suitable	Reason option should be replaced/removed	15	Suitable	
2	Suitable	Reason option should be replaced/removed	16	Suitable	
3	Suitable		17	Suitable	
4	Suitable	Reason option should be replaced/removed	18	Suitable	
5	Suitable		19	Suitable	
6	Suitable		20	Suitable	
7	Suitable		21	Suitable	
8	Suitable		22	Suitable	
9	Suitable		23	Suitable	
10	Suitable		24	Suitable	
11	Suitable		25	Suitable	
12	Suitable		26	Suitable	
13	Suitable		27	Suitable	
14	Suitable		28	Suitable	

Table 1 shows that in term of the scope and the sequence of acid-base properties of

salt solution content and its relevance to the educational level and pedagogical aspect as well as the language aspect, the questions in the FTI-ABPS2 are mostly suitable except for three questions including Q1 (question number 1), Q2 and Q4. The three questions have been revised by taking validator's feedback into account. The suggested revisions were for the reason options.

Empirical Validity of The Prototype FTI-ABPS2

The quality of the prototype FTI-ABPS2 is reflected based on the result of empirical validity, which is measured in terms of the following parameter: validity, reliability, difficulty level (DL), discriminatory index (DI) and distractor effectiveness. The results of each parameter are presented below.

Table 5. The Validity of FTI-ABPS2 items

Tier	Item	1	2	3	4	5	6	7
A	r_{xy}	0,377	0,151	0,330	0,343	0,640	0,150	0,606
	Category	Valid	invalid	Valid	valid	Valid	invalid	Valid
R	r_{xy}	0,605	0,478	0,191	0,298	0,040	0,149	0,402
	Category	Valid	valid	invalid	valid	Valid	invalid	Valid
B	r_{xy}	0,549	0,265	0,282	0,321	0,447	0,326	0,498
	Category	Valid	valid	Valid	valid	Valid	Valid	Valid
Tier	**Item**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A	r_{xy}	0,676	0,310	0,673	0,456	0,287	0,418	0,692
	Category	valid	valid	valid	valid	Valid	Valid	Valid
R	r_{xy}	0,421	0,352	0,329	0,225	0,05	0,108	0,364
	Category	valid	valid	valid	invalid	invalid	invalid	Valid
B	r_{xy}	0,627	0,522	0,676	0,439	0,212	0,425	0,655
	Category	valid	valid	valid	valid	invalid	Valid	Valid
Tier	**Item**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
A	r_{xy}	0,419	0,495	0,016	0,432	0,787	0,391	0,308
	Category	valid	valid	invalid	valid	Valid	Valid	valid
R	r_{xy}	0,603	0,249	0,446	0,600	0,459	0,421	0,270
	Category	valid	invalid	valid	valid	Valid	Valid	Valid
B	r_{xy}	0,689	0,326	0,268	0,566	0,784	0,429	0,361
	Category	valid	valid	valid	valid	Valid	Valid	Valid
Tier	**Item**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
A	r_{xy}	0,296	0,492	0,400	0,271	0,328	0,370	0,413
	Category	valid	valid	valid	valid	Valid	Valid	Valid
R	r_{xy}	0,174	0,565	0,645	0,430	0,593	0,382	0,345
	Category	invalid	valid	valid	valid	Valid	Valid	Valid
B	r_{xy}	0,231	0,555	0,529	0,297	0,448	0,502	0,431
	Category	invalid	valid	valid	valid	Valid	Valid	Valid

A = A tier (students' answer/first tiert)

R = R Tier (students' reason/ 3rd tier)

B = Both tier (both students' answer and reason simultaneously)

Table 5 shows that the valid category dominated all the questions in all tiers (A, R and B tiers). A small fraction of invalid questions also existed in all tiers. Those questions falling in the invalid categories were the subject to be revised. However, the revisions were carried out by considering the results of all the five parameters altogether.

In detail, the A tier for 25 questions (89.3%) fell in the "valid" category and only 3 (10.7%) invalid questions which are Q2, Q6 and Q17. The DL indices of these invalid questions fell in the "easy" category as well as poor DI indices. The low DL indices explain that all testees/participants answered the questions correctly. It is rooted that both the higher achievement group and the low achievement group provided the correct answer meaning the question can not differentiate the understanding of the two groups well.

For the R tier, 21 items (75%) fell in the "valid" category and 7 items (25%) were invalid. For B tier representing students' answers to the A and R tiers simultaneously, 26 items (92,8%) fell in the "valid" category and only 2 invalid items, which were Q12 and Q22. The root of the invalid category of Q12 is due to the fact that the distractors were not functional. This also confirmed the poor category of its DI index. For Q22, the distractors were functional, but the DL and DI indices were extremely low.

Table 6. The Reliability of FTI-ABPS2 instrument

Reliability	**A**	**R**	**B**
	0,833	0,801	0,806
Category	Very good	Very good	Very good

Table 6 shows that the reliability of the FTI-ABPS2 instrument in all tiers falls in the "very good" category. This implies that the instrument is suitable to be used to identify students' conception in the topic of Acid-base properties of salt solution.

Table 7. The DL indices of FTI-ABPS2 items

	Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	**DL index**	0,75	0,70	0,68	0,61	0,61	0,75	0,53
	Category	easy	moderate	moderate	moderate	moderate	easy	moderate
R	**DL index**	0,70	0,67	0,75	0,68	0,82	0,65	0,56
	Category	moderate	moderate	easy	moderate	easy	moderate	moderate
B	**DL index**	0,67	0,63	0,67	0,60	0,56	0,56	0,53
	Category	moderate	moderate	moderate	moderate	moderate	moderate	moderate
	Item	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A	**DL index**	0,56	0,68	0,60	0,33	0,65	0,65	0,46
	Category	moderate	moderate	moderate	moderate	moderate	moderate	moderate
R	**DL index**	0,61	0,19	0,79	0,68	0,88	0,88	0,72
	Category	moderate	hard	easy	moderate	easy	easy	easy
B	**DL index**	0,47	0,11	0,54	0,30	0,63	0,58	0,42
	Category	moderate	hard	moderate	hard	moderate	moderate	moderate
	Item	**15**	**16**	**17**	**18**	**19**	**20**	**21**
A	**DL index**	0,63	0,58	0,81	0,54	0,39	0,67	0,63
	Category	moderate	moderate	easy	moderate	moderate	moderate	moderate
R	**DL index**	0,51	0,72	0,67	0,53	0,70	0,60	0,84
	Category	moderate	easy	moderate	moderate	moderate	moderate	easy
B	**DL index**	0,42	0,53	0,58	0,39	0,39	0,51	0,58
	Category	moderate	moderate	moderate	moderate	moderate	moderate	moderate
	Item	**22**	**23**	**24**	**25**	**26**	**27**	**28**
A	**DL index**	0,70	0,56	0,58	0,67	0,51	0,44	0,54
	Category	moderate	moderate	moderate	moderate	moderate	moderate	moderate
R	**DL index**	0,74	0,58	0,53	0,67	0,53	0,63	0,68
	Category	easy	moderate	moderate	moderate	moderate	moderate	moderate
B	**DL index**	0,65	0,53	0,44	0,58	0,49	0,40	0,44
	Category	moderate	moderate	moderate	moderate	moderate	moderate	moderate

Table 7 shows that most questions were considered "moderate' and only small portion considered "easy" and "hard". In the A tier, 2 questions (7,2%) fell in the "easy" category and 26 questions (92,8%) in the "moderate" category. In the R tier, 9 questions (32,1%) fell in the "easy" category and 18 questions (64,3%) in the "moderate" category and only 1 question in the "hard" category. When both tiers combined and considered simultaneously (B tier), 26 questions (92,8%) fell in the "moderate" category, 2 questions (7,2%) in the "hard" category and none in the "easy' category. This implies that getting the correct answer in the B tier requires a solid scientific understanding.

Table 8. The DI Indices of FTI-ABPS2 items

	Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	DI index	0,34	0,09	0,34	0,34	0,69	0,06	0,65
	Category	good	poor	good	fair	good	poor	good
R	DI index	0,45	0,37	0,06	0,20	0,06	-0,01	0,37
	Category	good	good	poor	fair	poor	unsuitable	good
B	DI index	0,51	0,30	0,23	0,30	0,51	0,16	0,58
	Category	good	good	fair	good	good	fair	good
	Item	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A	DI index	0,65	0,20	0,58	0,33	0,20	0,41	0,58
	Category	good	fair	good	good	fair	good	good
R	DI index	0,34	0,25	0,20	0,27	0,03	0,03	0,41
	Category	good	fair	fair	fair	poor	poor	good
B	DI index	0,61	0,21	0,55	0,33	0,23	0,48	0,65
	Category	good	good	good	good	fair	good	good
	Item	**15**	**16**	**17**	**18**	**19**	**20**	**21**
A	DI index	0,37	0,48	0,03	0,41	0,65	0,37	0,23

	Category	good	good	poor	good	good	good	fair
R	DI index	0,54	0,13	0,37	0,51	0,45	0,23	0,17
	Category	good	fair	good	good	good	fair	fair
B	DI index	0,51	0,30	0,06	0,58	0,65	0,40	0,27
	Category	good	good	poor	good	good	good	fair
	Item	**22**	**23**	**24**	**25**	**26**	**27**	**28**
A	DI index	0,09	0,37	0,27	0,23	0,19	0,26	0,26
	Category	poor	good	fair	fair	fair	fair	fair
R	DI index	0,17	0,55	0,44	0,44	0,37	0,16	0,20
	Category	fair	good	good	good	good	fair	fair
B	DI index	0,06	0,44	0,26	0,27	0,23	0,33	0,26
	Category	poor	good	fair	fair	fair	good	fair

Table 8 shows that DI indices of the questions were dominated by the "good" and "fair" categories. A small portion of questions fell in the "poor" category. An unsuitable question was found in the R tier of a question, implying the reason should be removed and replaced. The dominant number of DI indices with good and fair categories confirms that the FTI-ABPS2 instrument was able to differentiate among students who understand the relevant concepts and those who don't understand the relevant concept.

Table 9. The Distractor Effectiveness of FTI-ABPS2 items

Item→	**1**		**2**		**3**		**4**		**5**		**6**		**7**	
Option↓	A	R	A	R	A	R	A	R	A	R	A	R	A	R
A	5	70	14	14	16	77	61	9	33	82	74	74	25	16
B	14	9	70	70	7	7	9	7	2	12	11	11	7	14
C	75	4	5	5	72	5	28	68	61	2	4	4	16	12
D	21	18	11	11	5	11	2	16	4	4	12	12	53	58
Item	**8**		**9**		**10**		**11**		**12**		**13**		**14**	
Option	A	R	A	R	A	R	A	R	A	R	A	R	A	R
A	56	4	9	18	18	2	18	23	23	2	65	5	4	5
B	16	58	7	47	47	82	33	4	4	2	4	2	47	4
C	23	25	67	12	12	12	25	68	68	89	30	7	46	19
D	5	14	18	23	23	4	25	5	5	7	2	86	4	72
Item	**15**		**16**		**17**		**18**		**29**		**20**		**21**	
Option	A	R	A	R	A	R	A	R	A	R	A	R	A	R
A	4	35	56	2	82	18	54	9	35	5	18	12	4	12
B	60	49	33	72	9	68	4	53	2	21	65	14	19	82
C	23	7	2	21	5	2	12	19	35	67	12	60	18	2
D	14	9	9	5	4	12	30	19	28	7	5	14	60	4
Item	**22**		**23**		**24**		**25**		**26**		**27**		**28**	
Option	A	R	A	R	A	R	A	R	A	R	A	R	A	R
A	67	72	32	30	2	2	9	2	39	54	37	28	19	19
B	14	14	7	7	35	40	67	12	5	7	47	65	54	54
C	12	4	5	5	60	53	19	68	5	32	11	5	11	11
D	7	11	56	58	4	5	5	18	51	7	5	2	16	16

Table 9 shows that all the options, either A tier or R tier in each question, are selected by more than one participant (testee). This implies that all the options are effective and attractive to be chosen, particularly for those who don't understand the relevant concept scientifically.

CONCLUSION

This study found that the FTI-ABPS2 is a robust instrument to be applied in investigating students' conception of both scientific understanding as well as misunderstanding. The good validity and reliability indices confirmed this firm finding. Reliability and validity are the two parameters representing the quality of an instrument as a whole (Kimberlin & Winterstein, 2008). Also, the fact that the questions were considered to be more difficult when both tiers are considered simultaneously confirm the essential of this the four-tier instrument in uncovering students' scientific understanding. B tier could only be answered correctly by those who hold a firm scientific understanding.

REFERENCES

[1] Allen, M. J., &Yen, W. M. 2002. *Introduction to Measurement Theory*. Waveland Press, Inc.

[2] Anonymous. n.d.. *Item Analysis of Classroom Tests: Aims and Simplified Procedures*. http://www.udel.edu/educ/gottfredson/451/unit9-guidance.htm

[3] Arikunto, S. 1993. *Dasar-Dasar Evaluasi Pendidikan*. PT Bumi Aksara.

[4] Caleon, I., & Subramaniam, R. 2010a. Development and Application of a Three-Tier Diagnostic Test to Assess Secondary Students' Understanding of Waves. *International Journal of Science Education*, *32*(7), 939–961.

[5[Caleon, I., & Subramaniam, R. 2010b. Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions. *Research in Science Education*, *40*(3), 313–337.

[6] DiBattista, D., & Kurzawa, L. 2011. Examination of the Quality of Multiple-choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning: CJSoTL*, *2*(2).

[7] Febriani, G., Marfu'ah, S., & Joharmawan, R. 2018. Identifikasi Konsep Sukar, Kesalahan Konsep, dan Faktor-Faktor Penyebab Kesulitan Belajar Hidrolisis Garam Siswa Salah Satu SMA Blitar. *J-PEK (Jurnal Pembelajaran Kimia)*, *3*(2), 35–43.

[8] Habiddin, H., & Page, E. M. 2019. Development and Validation of a Four-Tier Diagnostic Instrument for Chemical Kinetics (FTDICK). *Indonesian Journal of Chemistry*, *19*(3), 720. https://doi.org/10.22146/ijc.39218

[9] Husniah, I., Habiddin, H., Sua'idy, M., & Nuryono, N. 2019. Validating an instrument to investigate students' conception of Salt hydrolysis. *Journal of Disruptive Learning Innovation (JODLI)*, *1*(1), 1–6.

[10] Kimberlin, C. L., & Winterstein, A. G. 2008. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm*, *65*(23), 2276–2284.

[11] Kirbulut, Z. D. 2014. Using Three-Tier Diagnostic Test to Assess Students' Misconceptions of States of Matter. *EURASIA Journal of Mathematics, Science & Technology Education*, *10*(5), 509–521.

[12] Milenković, D. D., Hrin, T. N., Segedinac, M. D., & Horvat, S. 2016. Development of a Three-Tier Test as a Valid Diagnostic Tool for Identification of Misconceptions Related to Carbohydrates. *Journal of Chemical Education*, *93*(9), 1514–1520.

[13] Nimmermark, A., Ohrstrom, L., Martensson, J., & Davidowitz, B. 2016. Teaching of chemical bonding: a study of Swedish and South African students' conceptions of bonding. *Chemistry Education Research and Practice*.

[14] Nyachwaya, J. M. 2016. General chemistry students' conceptual understanding and language fluency: acid–base neutralization and conductometry. *Chemistry Education Research and Practice*, *17*(3), 509–522.

[15] Pesman, H., & Eryilmaz, A. 2010. Development of a Three-Tier Test to Assess Misconceptions About Simple Electric Circuits. *Journal of Educational Research*, *103*(3), 208–222.

[16] Romine, W. L., Todd, A. N., & Clark, T. B. 2016. How Do Undergraduate Students Conceptualize Acid–Base Chemistry? Measurement of a Concept Progression. *Science Education*, *100*(6), 1150–1183.

[17] Sreenivasulu, B., & Subramaniam, R. 2014. Exploring Undergraduates' Understanding of Transition Metals Chemistry with the use of Cognitive and Confidence Measures. *Research in Science Education*, *44*(6), 801–828.

[18] Yang, D. C., & Lin, Y. C. 2015. Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test. *Educational Research*.