ACTA DIDACTICA NAPOCENSIA
Volume 11, Number 2, 2018

UTILISATION OF RASCH MODEL FOR THE ANALYSIS OF AN INSTRUMENT DEVELOPED BY MAPPING ITEMS TO COGNITIVE LEVELS OF MARZANO TAXONOMY

Roxana S. Timofte, Laura Siminiciuc

Abstract: The scope this article was to develop an instrument to measure Chemistry students' ability regarding 'physical bonding' and to validate it. A number of 24 items were developed by mapping items to cognitive levels described by the Marzano taxonomy. A number of N=73 students were evaluated. Four items exhibited a MNSQ >1.3 and were eliminated from the final data analysis. At the final data analysis, the item difficulty measures were in normal ranges, as well as item separation and item reliability values. Person separation and person reliability values were showing that the number of items must be increased, since the instrument may not be sensitive enough to differentiate between low and high performers. Nevertheless, it was proven that the utilisation of Marzano's taxonomy in the development of items was successful in the sense that the items had difficulty measures in normal ranges and that the items had different levels of difficulty.

Key words: Rasch model, Marzano taxonomy, assessment, physical bonding, Chemistry students

1. Introduction

Item Response Theory (IRT) is used for the validation of instruments and to provide data regarding the item difficulty and person ability. Many studies focused on the comparison between Classical Test Theory and Item Response Theory (for example, Hambleton & Jones, 1993; Wiberg, 2004).

The two most used IRT models are the 1- and 2-parameter logistic (1PL & 2 PL). The probability for student *i* to respond correct at question / item *j* is calculated after the following equation (Equation 1):

$$P_j(\theta i) = \frac{\exp\{\theta i - bj\}}{1+\exp\{\theta i - bj\}}$$

Equation 1. *Probability for student i to respond correct to to item j - The 1PL Model*

Where θi is student's ability and b_j is the difficulty of task *j* (Rasch, 1960, Boone *et al*, 2014). High values for θ indicate a high ability level and high values for b indicate difficult items. The values of item difficulty or person ability are normally in the range [-3, 3] logit. Values outside this range show that there is a problem with the measured items.

Information regarding how well a particular item discriminates between students with different abilities can be also obtained. The discrimination parameter, a_j, could be added, forming the 2PL model (Equation 2):

$$P_j(\theta i) = \frac{\exp\{aj(\theta i - bj)\}}{1+\exp\{aj(\theta i - bj)\}}$$

Equation 2. *Probability for student i to respond correct to to item j- The 2PL Model*

Received: 2 May 2018, accepted 8 June 2018.

Cite as: Timofte, R. S.; & Siminiciuc, L. (2018). Utilisation of Rasch model for the analysis of an instrument developed by mapping items to cognitive levels of Marzano taxonomy. *Acta Didactica Napocensia*, 11(2), 71-78, DOI: 10.24193/adn.11.2.6.

where θi and b_j have the same meaning like in the 1PL model, and a_j represents discrimination of item j. The values for aj are normally in the range [0, 2], and the higher values indicate items which differentiate better students' ability (Sudol & Struder, 2010). Altrough the 1PL model is sometimes associated with the Rasch model, there are differences between the Rasch model and the 1PL model (Linacre, 2005).

The Rasch model was elaborated by the Danish mathematician George Rasch in 1960. This model was developed in order to overcome the problems which appear when using classical test theory in analysis of instruments (Boone, 2016, Jackson *et al*, 2002). More exactly, because the items have different difficulty levels, the raw scores obtained by summing up the correct answers can not be used to compare students 'ability. Furthermore, Rasch technique can be used to transform the non-linear raw data in linear scales, which can then be evaluated by the utilization of statistical parametric tests.

Some of the parameters considered when analyzing an instrument using the Rasch model are: item fit, item separation and reliability, person separation and reliability, Wright map, discrimination (Linacre & Wright, 2000). Assessment of item fit is reported in two ways: INFIT and OUTFIT (Linacre & Wright, 2000, Jackson *et al*, 2002). INFIT shows unexpected behavior observed at the responses near the person's ability level. OUTFIT shows if unexpected responses or outliers are found, taken in consideration person's ability. Both INFIT and OUTFIT are reported as unstandardized as mean square (MNSQ) and standardised MNSQ (Zstd). The MNSQ shows the amount of randomness and its values is taken in account to measure item fit or misfit. Item misfit is indicated by MNSQ >1.3 (Linacre & Wright, 2000). The item separation and reliability, person separation and reliability have different applications and implications. Person separation value is useful for person classification. When the number of participants in the study is large enough, a small value for separation (<3) when the value for reliability is also small (<0.8) shows that the instrument is not sensitive enough to differentiate between students with different abilities. In this case more items may be needed. (Linacre & Wright, 2000). Item separation verifies items hierarchy. A small value for items separation (<3) and reliability <0.9 implies that the number of study participants is not large enough to confirm items' difficulty hierarchy (Linacre & Wright, 2000). A high value of reliability (of persons or items) calculated by using the Rasch techniques implies that there is a high probability that the persons or items estimated with high values for measures have in fact higher measures that the persons or items who were estimated with low measures. Hence, reliability in this method estimates the replicability of item difficulty on a difficulty scale across students having different abilities (Linacre & Wright, 2000, Jackson *et al*, 2002). The Wright map shows items' difficulty hierarchy and persons' ability hierarchy, measured on the same logit scale.

Rasch method is a very good method to analyze the validity of an instrument. However, it is important to take in consideration the number of participants of the study versus the number of parameters measured for each item. Trying to estimate too many parameters with small amount of data may induce errors.

Utilization of Rasch model to assess instrument quality is a frequent practice among Science Education researchers (for example: Ziepprecht *et al*, 2017, Neumann *et al*, 2011) and psychometricins (for example: Boone, 2016, Wilson *et al*, 2006). Most often, the instruments are developed by using a competence model (for example: Ziepprecht *et al*, 2017, Walpuski *et al*, 2011) or a taxonomy of learning domains (for example*:* Kim *et al*, 2012).

In 2001 Marzano proposed a new taxonomy of learning, as an answer to the shortcomings of Bloom taxonomy (Irvine, 2017). Marzano's new taxonomy (Marzano & Kendall, 2007) includes three systems (self system thinking, metacognition and cognitive domain) and a knowledge domain (information, mental procedures, physical procedures). The different levels for cognitive domain are: retrieval, comprehension, analysis, knowledge utilization. The mental processes associated with each level of difficulty for cognitive domain are depicted below (Table 1):

Table 1. *Levels of difficulty for cognitive domain (Marzano & Kendall taxonomy)*

Level	Mental process
Level 1 – Retrieval	Recognizing, Recalling, Executing
Level 2 – Comprehension	Integrating, Symbolising
Level 3 – Analysis	Matching, Classifying, Analysig, Generalising, Specifying
Level 4 – Knowledge Utilisation	Decision-making, Problem-Solving, Experimenting, Investigating

2. Scope of this study

The scope of this study was to develop an instrument to measure Chemistry students' ability regarding 'physical bonding' and to validate it.

3. Design of the study

A number of N=73 Chemistry and Chemistry Engineering students participated at this study: 29 students (40%) in the third year of study and 45 students (60%) in the second year of study. 83.6% of participants were female, 16.4% were male.

A number of 24 items were developed by using the different cognitive levels described by Marzano taxonomy. The ratio between items with low difficulty, items with medium difficulty and items with high difficulty was 1:1:1. Examples of items are presented in Annex.

Data was analyzed with Winsteps version four. Information regarding interpretation of Winsteps outputs could be found at http://www.winsteps.com/index.htm.

4. Results and Discussion

Data analysis was started with item misfit analysis. It is recommended that items with MNSQ value > 1.3, as these items may induce errors in measuring. Four items exhibited values >1.3 for Outfit MNSQ (Figure 1) and were eliminated. A number of 13 participants in this study exhibited MNSQ >1.3. This shows that there are some issues with these participants; however, they could not be eliminated from the study. Final data analysis was undertaken with 20 items and 73 people.

Item difficulty and people ability

The measures for item difficulty were in the range [-1.83, 1.98] logit, M=0.00, SD=1.25. These values are in the [-3, 3] normal range. The measures for people ability were in the range [-1.47, 4.85] logit, M=0.83, SD=1.38. The measure for the ability of three persons was 4.85 logit. The rest of values were < 3. Hence, it can be considered that those three people whose ability measure was 4.85 logit had a higher ability level than the difficulty level of the tested items. The Wright map in which items ability and persons ability are presented on the same logit scale in presented in Figure 2. In Table 2 are depicted the values for measured difficulty of items by comparison with difficulty levels envisaged by Marzano taxonomy. As it can be observed, there is not a perfect alignment between the estimated levels of items and the measured values (for example, it was envisaged that item 12 has a difficulty level 4 after Marzano taxonomy, and the measured value was -1.41 logit, when the range of measures was [-1.83, 1.89]). However, utilisation of Marzano taxonomy enabled us to develop items of different difficulty levels and with measures in normal ranges.

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
21	18	73	1.90	.30	1.01	.1	1.31	1.3	.36	.41	81.9	78.0	Item LS21
18	21	73	1.64	.28	.83	-1.3	.77	-1.2	.55	.41	83.3	75.0	Item LS18
2	24	73	1.41	.27	1.04	.3	1.10	.6	.37	.41	70.8	72.3	Item LS2
8	25	73	1.34	.27	1.07	.7	1.02	.2	.36	.41	69.4	71.5	Item LS8
15	25	73	1.34	.27	1.33	2.7	1.64	3.2	.08	.41	61.1	71.5	Item LS15
22	25	73	1.34	.27	.91	-.7	.96	-.2	.47	.41	77.8	71.5	Item LS22
11	31	73	.92	.26	.99	.0	.97	-.1	.41	.40	66.7	67.6	Item LS11
23	32	73	.85	.26	.73	-3.2	.72	-2.4	.63	.40	86.1	67.2	Item LS23
17	36	73	.59	.26	.99	-.1	.97	-.2	.41	.40	63.9	66.1	Item LS17
24	36	73	.59	.26	.96	-.5	.91	-.7	.44	.40	66.7	66.1	Item LS24
5	38	73	.46	.26	1.36	3.7	1.50	3.4	.06	.39	50.0	66.0	Item LS5
7	43	73	.13	.26	.98	-.2	.97	-.1	.40	.38	70.8	67.4	Item LS7
9	44	73	.06	.26	.81	-2.1	.75	-1.8	.54	.37	77.8	67.8	Item LS9
16	48	73	-.22	.27	1.17	1.5	1.10	.6	.23	.36	58.3	70.5	Item LS16
10	54	73	-.67	.29	.81	-1.4	.71	-1.3	.49	.33	81.9	75.4	Item LS10
4	55	73	-.76	.29	1.37	2.4	2.08	3.4	-.10	.32	69.4	76.3	Item LS4
20	58	73	-1.02	.31	.91	-.5	.73	-.9	.40	.30	80.6	79.4	Item LS20
13	59	73	-1.12	.31	.99	.0	.98	.0	.30	.29	79.2	80.6	Item LS13
19	59	73	-1.12	.31	.84	-.9	.74	-.8	.44	.29	84.7	80.6	Item LS19
14	60	73	-1.22	.32	.96	-.1	.86	-.3	.32	.29	81.9	81.9	Item LS14
12	62	73	-1.44	.34	.90	-.4	.66	-.9	.39	.27	84.7	84.7	Item LS12
3	63	73	-1.57	.36	.97	-.1	.81	-.4	.30	.26	86.1	86.1	Item LS3
6	63	73	-1.57	.36	.94	-.2	1.18	.6	.28	.26	86.1	86.1	Item LS6
1	65	73	-1.84	.39	.89	-.3	.55	-1.0	.38	.23	88.9	88.9	Item LS1
MEAN	43.5	73.0	.00	.29	.99	.0	1.00	.0			75.3	74.9	
P.SD	15.5	.0	1.17	.04	.17	1.5	.34	1.5			10.2	7.1	

Figure 1. *Item statistics (for 24 items)*

Separation and reliability

The values for item separation and reliability (Item separation: 3.80, item reliability: 0.94, Figure 3) show that the number of persons who participated to this study was large enough to confirm the hierarchy of items with regard to their difficulty level.

The values for person separation and reliability (for 70 non-extreme persons: separation: 1.55, reliability: 0.71; for 73 extreme and non-extreme persons: separation: 1.68, reliability: 0.74, Figure 3) show that the instrument is not sensitive enough to differentiate between students with different ability levels.

Table 2. *Item difficulty measures*

Item no	Level of item difficulty estimated using Marzano Taxonomy	Item's difficulty measure (Rasch model) (logit)
1	1	-1.83
3	3	-1.54
6	4	-1.54
12	4	-1.41
14	2	-1.18
13	1	-1.07
19	1	-1.07
20	2	-0.96
10	3	-0.59
16	3	-0.1
9	3	0.21
7	1	0.28
17	4	0.79
24	4	0.79
23	4	1.08
11	4	1.16
8	2	1.63
22	3	1.63
2	2	1.72
18	4	1.98

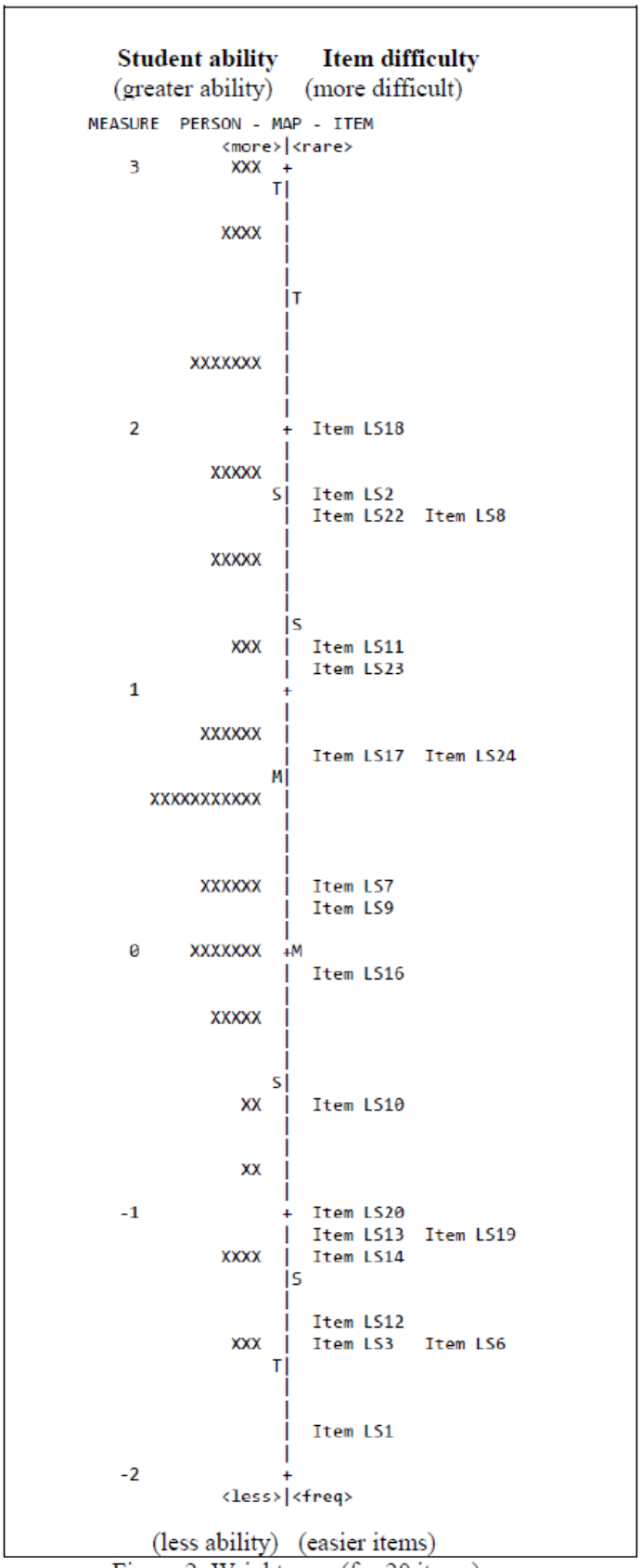

Figure 2: Wright map (for 20 items)

SUMMARY OF 73 MEASURED (EXTREME AND NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	12.4	20.0	.83	.63				
P.SD	3.9	.0	1.38	.26				
S.SD	3.9	.0	1.39	.26				
MAX.	20.0	20.0	4.85	1.85				
MIN.	5.0	20.0	-1.47	.53				

REAL RMSE .71 TRUE SD 1.19 SEPARATION 1.68 PERSON RELIABILITY .74
MODEL RMSE .68 TRUE SD 1.20 SEPARATION 1.77 PERSON RELIABILITY .76
S.E. OF PERSON MEAN = .16

PERSON RAW SCORE-TO-MEASURE CORRELATION = .97
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .78 SEM = 1.81

SUMMARY OF 20 MEASURED (NON-EXTREME) ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	45.4	73.0	.00	.31	1.00	.0	1.00	.0
P.SD	15.1	.0	1.25	.04	.14	1.2	.30	1.0
S.SD	15.5	.0	1.28	.04	.14	1.2	.30	1.0
MAX.	65.0	73.0	1.98	.40	1.33	2.6	1.86	1.7
MIN.	21.0	73.0	-1.83	.27	.72	-2.6	.55	-1.8

REAL RMSE .32 TRUE SD 1.21 SEPARATION 3.80 ITEM RELIABILITY .94
MODEL RMSE .31 TRUE SD 1.21 SEPARATION 3.89 ITEM RELIABILITY .94
S.E. OF ITEM MEAN = .29

ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00

Figure 3. *Item and person separation and reliability*

5. Conclusion

The item difficulties are in normal ranges, as well as item separation and item reliability values. It was proven that the instrument containing items developed by incorporating the different cognitive levels of Marzano taxonomy into items is an instrument containing items exhibiting difficulty measures in normal ranges. Furthermore, the developed items have different levels of difficulty. However, person separation and person reliability values are showing that the number of items must be increased, since the instrument is not sensitive enough to differentiate between low and high performers. The goal for a further study is to increase the number of items with medium difficulty, in order to have the following ratio of item difficulty: items with low difficulty: items with medium difficulty: items with high difficulty: 25%:50%:25%, and the final instrument to be tested and the results analyzed.

References

[1] Boone W.J., Staver J.R., Yale M.S. (2014). *Rasch Analysis in the Human Sciences*, Dordrecht, Netherlands: Springer.

[2] Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?, *CBE Life Sci Educ*, *15*:*rm4*, 1-7.

[3] Hambleton, RK, Jones, RW (1993). Comparison of classical test theory and item response theory and their applications to test development, *Education Measurement: Issues and Practices*, *12*, 38–47.

[4] Irvine, J. (2017). A comparison of revised Bloom and Marzano's New Taxonomy of Learning, *Research in Higher Education*, 33, 1-16.

[5] Jackson, T. R., Draugalis, J. R., Slack, M. K., Zachry, W. M., D'Agostino, J. (2002). Validation of authentic performance assessment: a process suited for Rasch Modeling, *Am. J, Pharm. Educ.*, *66*, 233-243.

[6] Kim, M.-K., Patel, R. A., Uchizoni, J. A., Beck, L. (2012). Incorporation of Bloom's Taxonomy into Multiple-Choice Examination Questions for a Pharmacotherapeutics Course, *Am. J, Pharm. Educ.*, *76*(6), 1-8.

[7] Linacre J.M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model, *Rasch Measurement Transactions*, 19 (3), 1032.

[8] Linacre, J. M., & Wright, B. D. (2000). *Winsteps*, Chicago, IL: MESA Press

[9] Neumann, I., Neumann, K., Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science test, *International Journal of Science Education*, *33*(10), 1373-1405.

[10] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Chicago: University of Chicago Press.

[11] Robert J. Marzano John S. Kendall (2007). *The New Taxonomy of Educational Objectives*, Hawker Brownlow Education.

[12] Sudol, L. A., Studer, C. (2010). Analyzing test items: using item response theory to validate assessments, *Proceedings of the 41st ACM technical symposium on Computer science education*, 436-440.

[13] Walpuski, M., Ropohl, M., Sumfleth, E. (2011). Students' knowledge about chemical reactions – development and analysis of standard-based test items, *Chemistry Education: Research and Practice*, *12*, 174-183.

[14] Wiberg, M. (2004). Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test (No. 50). Umeå, Sweden: Umeå University.

[15] Wilson, M., Allen, D., & Li, J. Corser. (2006). Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach, *Health Education Research*, *21*, i19–i32.

[16] http://www.winsteps.com/index.htm, retrieved on 02 May 2018.

[17] Ziepprecht, K.; Schwanewedel, J.; Heitmann, P.; Jansen, M.; Fischer, H. E.; Kauertz, A.; Kobow, I.; Mayer, J.; Sumfleth, E.; Walpuski, M. (2017). Modellierung naturwissenschaftlicher Kommunikationskompetenz: ein fächerübergreifendes Modell zur Evaluation der Bildungsstandards, *Zeitschrift für Didaktik der Naturwissenschaften*, 1-13.

Authors

Dr Roxana S. Timofte is lecturer in Chemistry Education field at Babeş-Bolyai University, Cluj-Napoca, Romania. Email address: roxana.timofte@ubbcluj.ro

Laura Siminiciuc is MSc student enrolled at the Chemistry Teacher Training Program at Babeş-Bolyai University, Cluj-Napoca, Romania. Email address: laura.siminiciuc@yahoo.com

Acknowledgement

RST is grateful to Prof Maik Walpuski (Duisburg-Essen University, Germany) for providing the Winsteps program.

Annex

Test – PHYSICAL BONDING

Examples of items, for different levels of Marzano taxonomy:

1. Level 1

Choose the correct afirmation:
a. van der Waals forces are stronger than hydrogen bonds;
b. the strongest intermolecular interaction is the hydrogen bond;
c. the intermolecular bonds are as strong as the covalent ones;
d. dispersion forces are stronger than hydrogen bonds.

2. Level 2

In the strong Hydrogen bonds:
a. the proton is placed asymmetrically in relation to the atoms it makes bonds with;
b. a tri-centric asymmetrical bond is formed;
c. the proton in bonded in a bi-centric way to the closest atom;
d. the proton, tri-centric bonded, is found at the same distance by all atoms it bonds.

3. Level 3

The hydrogen bonds take place between a Hydrogen atom in a polar molecule and a _____ atom in another polar molecule.
a. high electronegative;
b. low electronegative;
c. low ionization energy;
d. electropozitive.

4. Level 4

Which of the following pairs of compounds will form hydrogen bonds?
a. CCl_4 and CH_3OH
b. H_2O and NH_3
c. NO_2 and HF
d. CH_4 and H_2O