EJPE
european journal of physics education

European J of Physics Education Volume 2 Issue 3 2011 1309-7202 Mesic

Modeling the Discrimination Power of Physics Items

Vanes Mesic

Faculty of Science, University of Sarajevo, Zmaja od Bosne 35, 71000 Sarajevo, Bosnia and Herzegovina

(Received 22.06.2011 Accepted 05.08.2011)

Abstract
For the purposes of tailoring physics instruction in accordance with the needs and abilities of the students it is useful to explore the knowledge structure of students of different ability levels. In order to precisely differentiate the successive, characteristic states of student achievement it is necessary to use test items that possess appropriate discriminatory power. By identifying the cognitive factors, which account for differences or similarities between high achievers and low achievers, we can evaluate the efficacy of developing various aspects of physics competence within the physics instruction. Further, knowing the predictors of physics item discrimination power makes it possible to systematically modify physics items with the purpose of improving their psychometric characteristics. In this study, we conducted a secondary analysis of the data that came from two large-scale assessments of student physics achievement at the end of compulsory education in Bosnia and Herzegovina. Foremost, we performed a content analysis of 123 physics items that were included within abovementioned assessments. Thereafter, an item database was created. Items were mainly described by variables, which were supposed to reflect some basic cognitive domain characteristics of high and low achievers. For each of the items, we calculated the item discrimination power. Finally, a regression model of physics item discrimination power was created. It has been shown that 43,6 % of item discrimination power variance can be explained by factors which reflect the automaticity, complexity and modality of the knowledge structure that is relevant for generating the most probable correct solution, as well as by the constructs of cognitive load and retention. Interference effects between intuitive and formal physics knowledge structures proved to influence the item discrimination power, too.
Keywords: discrimination power, physics, test construction, linear regression

Introduction

Today, it is widely accepted that instruction has to be tailored in accordance with the needs and abilities of students for whom it is intended. Differences in students' foreknowledge lead to differences in the attainment of knowledge. In order to maximize the learning of all students, we should be aware of the nature of such differences and develop strategies to meet them. Achievement tests represent the most practical way to get insight into the structure of students' physics knowledge. In order to faithfully reflect the specific characteristics of cognitive achievement for students of different ability levels, it is important to design test items with acceptable discriminatory power. The inclusion of such items within a test makes it possible to distinguish the specific, successive achievement states, which constitute the physics-learning pathway. Further, by analyzing the test results, we can identify the factors that account for differences between subgroups of students associated with the mentioned achievement states.

Taking into account that these factors reflect some basic aspects of the physics-learning pathway, we could design our lessons in a more systematic way, with the purpose of helping students from low achievement groups to pass over to the higher achievement groups. In order to get reliable feedback on achievement differences it is necessary to ensure that the assessments include representative student samples. For the purposes of physics education quality management, it is especially important to conduct large-scale assessments, as well as to analyze

and use the results of those assessments. So far, students from Bosnia and Herzegovina have participated in two large-scale assessments of cognitive achievement in physics.

In 2006, the local Standards and Assessment Agency (SAA) conducted a large-scale study in order to assess students' achievement at the end of compulsory education (eighth/ninth grade students, depending on region) in Bosnia and Herzegovina (Petrovic, 2006). One year later, students from Bosnia and Herzegovina participated in the Trends in International Mathematics and Science Study (TIMSS). TIMSS has been conducted in 4-year cycles. It incorporates assessments of student mathematics and science achievement at the end of fourth and eighth grade, as well as collecting data about teaching and learning contexts in each participating country (Olson, Martin & Mullis, 2008).

Results of primary analyses for the data obtained within abovementioned assessments provided only a general overview of students' physics achievement (Petrovic, 2006; Martin, Mullis & Foy, 2008). The practical meaning of the quantitative test results remained rather unclear. Furthermore, it has been reported that a large number of physics items had to be discarded because of their low discrimination power (Petrovic, 2006).

In order to receive useful feedback for all the participants of the physics education process at the level of compulsory education in Bosnia and Herzegovina, we attempted to identify the factors which influenced the achievement differences between groups of high and low achievers, as well as to rank them with respect to their importance. These results, along with item difficulties associated with identified factors, can be used to point out possible shortcomings of the physics instruction in Bosnia and Herzegovina.

The practical importance of this study is also reflected in the potential improvement of the test design process. Results of earlier research revealed that even experienced test developers and subject matter experts may have difficulty in predicting the psychometric characteristics of test items (Bejar, 1983). The item difficulty can be only known after piloting the test whereby, items with poor psychometric features are often automatically discarded. Therefore, "in actual test development practice, the number of test items that must be developed and pretested is typically greater, and sometimes much greater, than the number that is eventually judged suitable for use in operational test forms."(Chalifour & Powers, 1989)

By identifying the predictors of physics item discrimination power, we could achieve better control of physics item discriminatory behavior. Instead of automatically discarding items with originally poor psychometric characteristics after piloting the test, we could systematically modify them in order to improve their discrimination. The theoretical significance of this study is reflected in gaining additional insight into the nature of physics competence and its characteristic structure for subgroups of high and low achievers.

Review of the literature

Within the literature qualified as relevant for this study, we can sort out two types of research:

- Research primarily directed to comparisons of high and low achievers,
- Psychometric research directed at improvement of test design.

With regard to research examined the differences between high and low achievers, most existing articles are related to mathematics education. Taking into account the fact that mathematical ability is very important for the physics cognitive domain, we will consider such

papers as relevant for this study. Alexander (1960) explored the characteristic differences between high and low achievers in seventh grade mathematics problem solving. He identified cognitive categories for which high achievers proved to be more advanced in comparison to low achievers at a statistically significant level (see Table 1).

Table 1. Statistically significant differences in favor of high achievers (at the $p<0.05$ level)

Specific mental abilities	**Quantitative skills**	**General reading skills**	**Problem solving reading skills**	**Interpretation of quantitative materials**
General reasoning ability	Understanding mathematical terms and concepts	Comprehension of reading materials	Comprehension of statements in problems	Finding data from graphs, tables, charts and maps
Ability to understand verbal concepts	Skill in computation	Understanding words in context	Selection of relevant details in problems	Perception of relationships involving comparison of data
			Selection of correct procedures to solve problems	Recognition of limitations of given data

Surprisingly, the achievement differences regarding the ability to visualize objects in two or three dimensions were non-significant. Regarding psychometric research, relatively little attention has been devoted to the issues of helping test developers to better control the statistical characteristics of test items. For most of this research, the variable of primary interest has been item difficulty. According to Chalifour & Powers (1989), predictors of item difficulty also predict item discrimination power, but to a lesser degree. Thus, we can also refer to better-explored research field related to item difficulty predictors to single out potential item discrimination power predictors.

Rosca (2004) conducted a study with the purpose of identifying factors that made the TIMSS 2003 science items difficult for the students from United States of America. She sorted out 17 potential predictors of item difficulty and tested their statistical significance by creating a regression model of item difficulty. The obtained model made it possible to account for 29.8 % of item difficulty variance by means of Flesch reading ease score, ratio of the number of words in the solution and average number of words in distractors, cognitive level according to Bloom, average number of words in distractors and the presence of graphics in the item stem. Thereby, the "cognitive level according to Bloom" proved to be the strongest predictor of science item difficulty.

Mesic & Muratovic (2011) performed a similar study with the purpose of explaining sources of TIMSS 2007 and SAA 2006 physics item difficulties for students from Bosnia and Herzegovina. Their model accounted for 61.2% of item difficulty variance. It consisted of predictors which reflect the automaticity, complexity, and modality of the knowledge structure that is relevant for generating the most probable correct solution, as well as by the divergence of required thinking and interference effects between intuitive and formal physics knowledge structures.

Kauertz conducted another study related to physics item difficulty. According to Kauertz (2007) physics competence can be modeled based on combinations of cognitive activities, content complexity and guiding ideas (see Table 2).

Table 2. Kauertz's model of physics competence

Cognitive activities	***Content complexity***	***Guiding ideas***
Knowing	One fact	Concept of energy
Structuring	Several facts	Concept of matter
Exploring	One relationship	Concept of interaction
	Several unrelated relationships	Concept of systems
	Several related relationships	Mathematical formalism
	Basic concept	

Within the study by Kauertz only "content complexity" and "guiding idea" proved to be statistically significant predictors of physics item difficulty, whereas a much bigger effect was reported for the "content complexity" than for the "guiding idea" factor. We can conclude that the competency differences between high achievers and low achievers in physics are not sufficiently explored. Further, the factors that influence the value of physics item discrimination power are largely unknown. Therefore, it is useful to take into account Chalifour and Powers' conclusion according to which predictors of item difficulty often predict item discrimination power, too. Also, it is important to note that, generally, cognitive factors prove to be stronger predictors of items' psychometric properties than formal item features.

Methods and Procedures

Student sample

In 2006, SAA conducted an assessment of student achievement in physics at the end of compulsory education in Bosnia and Herzegovina. 1377 students participated in that study. One year later, 4220 students of same age as in the previous study (mostly 14 year old) participated in TIMSS.

In both studies, the student sample was generated by using a two-stage stratified cluster design (Petrovic,2006;Olson et al,2008). At the first stage, schools were sampled, and at the second stage a sample of students – mostly from one class – from the target grade in the sampled schools was drawn. The student samples were representative (Petrovic, 2006 ; Schütz, 2006).

For purposes of item discrimination power analyses, we defined the subsamples of high achievers and low achievers for both studies in line with Kelley's "upper-lower group size" recommendations (Kelley, 1939). Firstly, by using the achievement databases from the SAA 2006 and TIMSS 2007 studies (Ref.12-13), we ranked students with respect to their achievement score for each study separately. The subsample of high achievers consisted of 27% of students who were on the top of each rank-list and the subsample of low achievers included the 27% of students from the bottom of each list. When speaking of low and high achievers in this paper, we will refer to subsamples of students, as here defined.

Because of similar student sample characteristics (representativeness, sampling design, age of students, students from subsequent generations, etc.), we supposed the student samples

from both studies to be approximately equivalent. The shapes of the achievement score distributions proved to be similar, too.

Item sample

According to science item almanacs the TIMSS 2007 test booklets included 59 physics items, whereas the SAA 2006 test booklets included 64 physics items (Ref.12-13). Within the whole sample of 123 physics items, there were 66 multiple-choice items and 57 constructed response items.

The TIMSS 2007 physics items were created along the lines of TIMSS assessment frameworks and the SAA assessment of physics achievement was based on the local curricula that were current in 2006. The content coverage of the TIMSS 2007 and SAA 2006 test forms was approximately the same. Within the procedure of virtual equating of these test forms it has been shown that the slope of the best fit line within the cross plot of item difficulties doesn't deviate too much from the identity slope line which is an indicator of discrimination power comparability between these two test forms (Mesic&Muratovic,2011).

Design and procedures

Taking into account that the physics item discrimination power depends on differences and similarities between high and low achievers related to certain cognitive aspects of students' physics competencies, we studied the relevant literature with the purpose of identifying constructs that define the cognitive dimension of physics competence. Also, we largely referred to the set of item difficulty predictors, which had been already identified for the TIMSS 2007 and SAA 2006 assessments (Mesic & Muratovic, 2011). Several potential predictors have been sorted out based on experience. An item content analysis with respect to the identified cognitive constructs as variables has been performed. Mostly, these cognitive constructs were characterized by a hierarchical structure, so we had to describe items by multiple level variables. Each item was associated with only one level of each variable. When we were classifying items with respect to the allocated types of knowledge or cognitive processes, we assigned the item to the highest allocated level of the correspondent variable within the most probable solution (Teodorescu, Bennhold & Feldman, 2008).

In order to perform quantitative item analysis, we created an item database by using the SPSS software. The database contained information regarding the 123 physics items from the conducted large-scale assessments. We described items only by those variables (see **Table 3**) whose levels could be associated with at least 10 items. Therefore, some categories of the original Kauertz's content complexity construct had to be collapsed. Thus, we obtained the "Modified Kauertz's content complexity" variable. Its baseline category (declarative knowledge) can be used to describe items, which require static knowledge, whereby the other two levels (relationships and related relationships) can be used to describe the complexity of schematic knowledge required by some items. Thereby, the ''schematic knowledge'' construct represents ''knowledge which combines procedural and declarative knowledge.''(Marshall,1988)

Table 3. List of variables created for purposes of quantitative item analysis

Variable name	Levels of the variable	References

Modified Kauertz's content complexity	0 – declarative knowledge 1 – relationships (including rules for their use) 2 – related relationships (including the rules for their use)	**Ref. 8,9**
Analytic representation	0 – doesn't require the use of analytic representation 1 – requires the use of analytic representation	**Ref. 8**
Knowledge of experimental method	0 – doesn't require knowledge of experimental method 1 – requires knowledge of experimental method	**Ref.8**
Interferential effects of intuitive and formal physics	0 – negligible interferential effects 1 – high probability of positive intuitive physics or p-prim influence 2 – high probability of negative intuitive physics or p-prim influence	**Ref. 8,16,17,18**
Divergent thinking	0 – doesn't require divergent thinking 1 – requires divergent thinking	**Ref. 19**
Item openness	0 – multiple-choice items (4 options) 1 – constructed response items	**Ref. 8,20**
Correspondent grade	0 – seventh grade contents 1 – eighth grade contents	Personal experience
Number of words in item stem	Variable measured at the ratio level	**Ref. 7**
Number of depictors	The number of relevant entities (objects) which have to be taken into consideration within the process of item solving. Mostly this variable is related to the number of physical objects which are part of a physical phenomena. Variable measured at the ratio level	**Ref. 21**

Most of the variables from Table 3 were shown to be predictors of physics item difficulty (Mesic & Muratovic, 2011). For these variables the inter-coder agreement proved to be substantial (Mesic & Muratovic, 2011). In comparison with the list of variables created for the purposes of identifying predictors of physics item difficulty only two new variables have been added to the list within the study of item discrimination power:

- grade level at which the item contents are predominantly taught,
- the "Number of depicters" which are relevant for item solving.

The inter-coder agreement for coding the items with respect to the "Correspondent grade" variable proved to be perfect, and the Krippendorf's alpha for the "Number of predictors" variable amounted to 0.77, which is a satisfying value. With the purpose of evaluating the importance and statistical significance of singled out potential predictors, we had to establish a relationship between the presented theoretical item descriptors and an empirical measure of item discrimination power. Out of the large set of possible empirical item discrimination measures we chose to use Kelley's model for calculation of the discrimination index because of its compatibility with our research objectives. This discrimination index could be calculated based on the student achievement databases (Ref.12-13) by using the formula:

$$D = p_t - p_b,$$

where:

p_t – proportion of students in the high achievement group (top 27%) who got the item correct,
p_b – proportion of students in the low achievement group (bottom 27%) who got the item correct.

Thus, we could assign to all items within the item database empirical discrimination power measures. Now, it was possible to quantify the statistical significance and relative importance of the singled out potential item discrimination power predictors. For this purpose, we decided to create a linear regression model of physics item discrimination power.

Firstly, we had to check if the size of our item sample was big enough for regression analysis purposes. According to Miles and Shelvin (2001), if we expect to obtain a large effect, it is sufficiently to have 80 items of analysis. Clearly, this condition has been met. Furthermore, for categorical variables with more than two levels a dummy coding procedure had to be implemented. Thereby, the variable levels encoded with zero (see Table 3) had been chosen to represent baseline categories.

We were also interested in exploring interaction effects between some of the created variables. Thus the following list of potential predictors has been finally used within the SPSS linear regression procedure: "Relationships", "Related relationships", "Analytic representation", "Experimental method", "Positive influence of intuitive physics", "Negative influence of intuitive physics", "Divergent thinking", "Item openness", "Grade", "Number of words in item stem", "Number of depicters", "Item openness*Analytic representation", "Item openness * Relationships", "Item openness*Related relationships", "Grade*Relationships", "Grade*Related relationships" and "Grade*Analytic representation". Thereby the backward method was selected, because we were not in common with the relative importance of the singled out potential predictors of item discrimination power.

Results

Basic features of the obtained item discrimination power model

The implementation of procedures, which were described in the previous section, gave rise to a model whose basic features are given in **Table 4**.

Table 4. Item discrimination power model summary

Model	R	R - Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
	0.661	0.436	0.380	0.131	1.746

Predictors: (Constant), Related relationships, Item openness*Analytic representation, Item openness, Grade * Relationships, Analytic representation, Grade * Related relationships, Correspondent grade, Item openness * Relationships, Relationships, Number of depictors, Positive influence of intuitive physics.
Dependent variable: Item discrimination power

The obtained model makes it possible to explain 43.6 % of item discrimination power variance. Based on the difference between $R^2$and adjusted R^2, we can conclude that the use of a different item sample from the same population would probably cause an approximately 5% percent drop of the item discrimination power variance that could be accounted for by the

generated model. Only item discrimination power predictors that proved to be statistically significant at the $p<0.05$ level remained in the model (see Table 5).

Table 5. Item discrimination power predictor statistics; Percent of correct answers are given for the variable levels encoded by one

Predictor	B	Std. error	Beta	Tolerance	Percent correct (upper group)	Percent correct (lower group)
(Constant)	0.412	0.032				
Item openness (CR items)	0.144	0.035	0.430*	0.478	43%	11%
Related relationships	-0.223	0.042	-0.594*	0.414	34%	13%
Relationships	-0.088	0.042	-0.259*	0.333	48%	17%
Positive influence of intuitive physics	0.071	0.033	0.157*	0.939	63%	24%
Analytic representation	0.170	0.062	0.378*	0.270	35%	8%
Number of depictors	-0.035	0.013	-0.228*	0.763	-	-
Correspondent grade (8-th grade)	-0.092	0.041	-0.273*	0.341	51%	19%
Grade * Relationships	0.186	0.058	0.422*	0.298	46%	11%
Grade * Related relationships	0.157	0.064	0.282*	0.391	37%	15%
Item openness * Relationships	-0.113	0.052	-0.269*	0.333	39%	8%
Item openness * Analytic representation	-0.293	0.074	-0.579*	0.238	24%	4%

*Note: $R^2=0.436$; * $p<0.05$.*

Based on standardized β-coefficients, we can rank statistically significant predictors with respect to the size of their influence on item discrimination power. The predictor "Related relationships" exerts the largest influence on item discrimination power followed by "Item openness*Analytic representation", "Item openness", "Grade * Relationships", "Analytic representation", "Grade * Related relationships", "Correspondent grade", "Item openness * Relationships", "Relationships", "Number of depicters", and "Positive influence of intuitive physics".

The largest achievement difference has been obtained for the category of items which tap declarative knowledge – on average 69% of high achievers solved those items correctly, compared to 29% low achievers.

Identification of potential outliers and influential items

By performing case wise diagnostics, we identified four outliers (see Table 6).

Table 6. Case-wise diagnostics

Case Number	Std. Residual	Discrimination power	Predicted Value	Residual
55	2.006	0.59	0.327	0.264
77	-2.111	0.14	0.412	-0.278
79	-2.705	0.00	0.358	-0.356
113	-2.116	0.04	0.319	-0.278

So, the proportion of items whose standardized residuals are above 2 is below 5 %, and the proportion of those items whose standardized residuals are above 2.5 is less than 1 %. These values are tolerable (Field, 2005). By calculating Cook's distance values, we checked if there were any items that had exerted large influence on the model as a whole. For all items these values were considerably below 1 (see Figure 1). Thus, we can conclude that there were no influential items and that the model is stable.

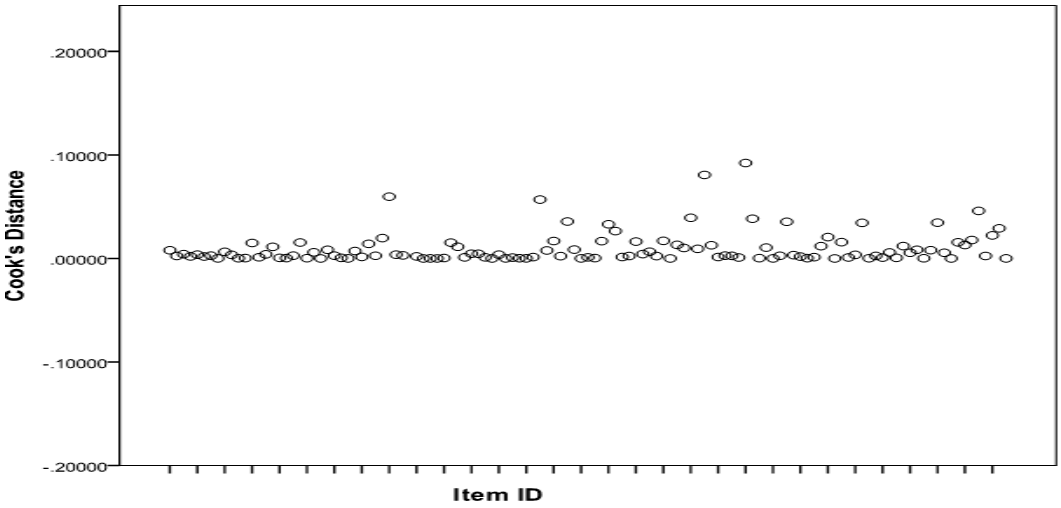

Figure 1. Cook's distances for used items

For the purpose of measuring the influence of each item on the individual predictors, DFBeta values (differences between β-coefficients when one item is included and not included, respectively) for each predictor were calculated. All the DFBeta values for the obtained model were below 0.55. It is supposed that the standardized DFBeta should not be above 1 (Field,2005). This condition is obviously met.

Testing assumptions of regression

In order to check the assumption of independent residuals, we calculated the Durbin-Watson statistics. Values above 3 or below 1 indicate that this assumption is not met, and the value 2 is ideal (Field, 2005). For our model the value of Durbin-Watson statistics (see Table 4) is 1.746. This is close to the ideal value, so we can claim that the assumption of independent residuals has been met. Based on the fact that the values of tolerance statistics (see Table 5) are above 0.2 for all the item discrimination power predictors, we can conclude that there is no significant multi-collinearity between them. In order to check the assumption of normally distributed residuals we calculated the Kolmogornov-Smirnov and Shapiro-Wilk statistics for standardized residuals (see Table 7).

Table 7. Tests of normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	0.057	123	0.200	0.989	123	0.429

Both tests are not statistically significant. Thus, we can conclude that the distribution of standardized residuals does not significantly deviate from the normal distribution. In order to check the assumptions of linearity and homoscedasticity, we have created a "standardized residual vs. standardized predicted value" plot (see Figure 2). There is a deviation from linearity if the spots within the scatterplot form a curve-like shape (Field, 2005). Further, we can suspect heteroscedasticity if the distances between extreme spots, with respect to the y-axis, are strongly varying when we move across the x-axis.

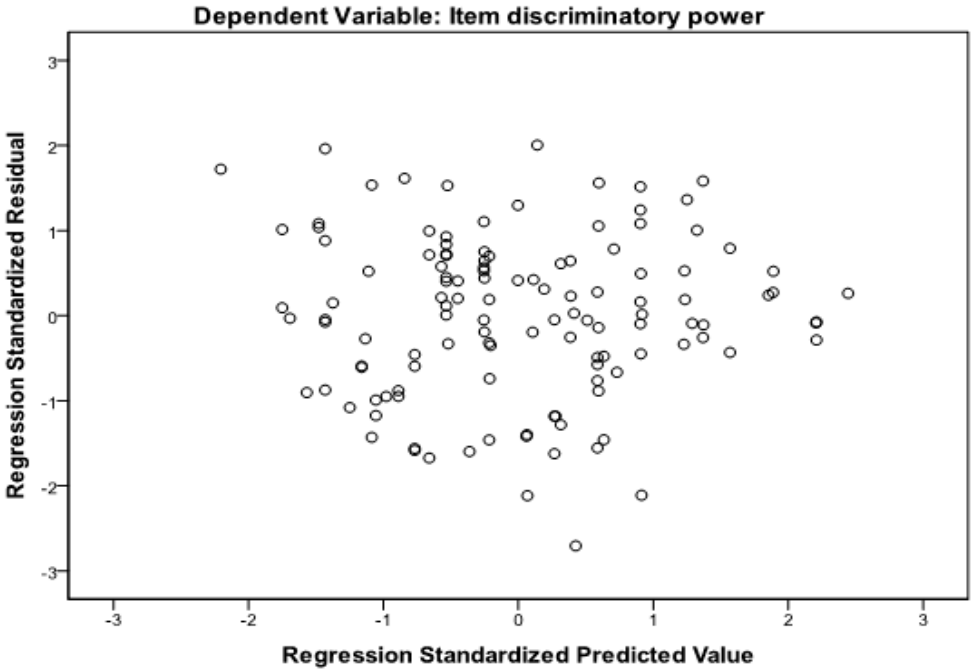

Figure 2. Checking the assumption of homoscedasticity

By analyzing the scatterplot we can conclude that the assumption of linearity has been met. Data are mostly homoscedastic.

Discussion

Based on the comparison of standardized β-coefficients for the predictors "Relationships" and "Related relationships" and on their interaction with the "Correspondent grade" and "Item openness" variables, we can conclude that the discrimination power of seventh grade multiple-choice items decreases if we increase the complexity of the knowledge structure which is most probably used for item solving, provided that all the other predictors are held constant. Taking into account the statistical significance of these predictors, we also can conclude that seventh grade multiple-choice items, which tap schematic knowledge, are significantly less discriminative than correspondent items that tap declarative knowledge.

Taking into account the reported achievement differences for the category of items that tap declarative knowledge we can conclude that the ability to remember facts is the feature of physics cognitive achievement, which largely differentiates high achievers from low achievers at the end of compulsory education in Bosnia and Herzegovina. It follows that the primary school physics instruction fails to foster lasting, higher level knowledge structures even in the case of high achieving students. It is important to note that the final influence of knowledge complexity on item discriminatory behavior depends also on the values of the "Correspondent grade" and "Item openness" variables. The nature of these interactions will be additionally discussed later on.

It has been shown that item discrimination power increases with an one unit change of the "Positive influence of intuitive physics" predictor, provided that all other predictors are held

constant. We can conclude that high achievers more effectively utilize their intuitive physics knowledge in comparison with low achievers. Generally, we should be aware of the fact that students develop their intuitive physics knowledge structures even before entering the formal physics education. Physics teachers should more often utilize the positive aspects of intuitive physics for effectively introducing and building of formal physics concepts (Clement, 1994). This practice could help low achievers to efficiently build links between their intuitive and formal physics knowledge structures.

"The number of relevant depicters" is the only ratio variable proved to be a statistically significant predictor of item discrimination power. It has been shown that it influences the item discrimination power by lowering it, provided that all other predictors are held constant. This result can be partly explained by the cognitive load theory (Sweller, van Merriënboer & Paas,1998). In fact, the human short-term memory is very limited with respect to the number of elements (chunks), which can be hold in the memory, at the same time. Cognitive operations on these elements occupy additional space (Sweller et al, 1998). Thus, clearly the cognitive demand increases with the number of physical objects, which have to be held and eventually manipulated within the short-term memory. In other words, the increasing cognitive demand associated with the increasing number of relevant depicters makes the correspondent item more difficult and this effect is more distinctive for high achievers than for low achievers. This indicates, one more time, that the largest differences between Bosnia-Herzegovinian high and low achievers were obtained on items not overly cognitive demanding.

"Analytic representation" proved to be a statistically significant predictor of item discrimination power, too. The need for usage of analytical representation makes the discrimination power of multiple-choice items to increase, provided that all the other variables are held constant. It is important to emphasize that the final influence of this factor on item discrimination behavior depends strongly on values of the "Item openness" variable. Regarding formal predictors, the variables "Item openness" and "Correspondent grade" showed up to be statistically significant, as well as the interactions "Item openness*Analytic representation", "Item openness * Relationships", "Grade * Relationships" and "Grade * Related relationships".

The item discrimination power difference between constructed-response and multiple-choice items is different for the two levels of the "Analytic representation" variable. If it is necessary to use the analytic representation in order to solve the item, constructed response items are less discriminative than multiple choice items and this difference is significantly different from the difference between opened and closed items which don't require the use of analytic representation. In other words, the need for analytic representation usage makes the discrimination power of constructed response items decrease, whereby the discrimination power of multiple-choice items increases, at the same time. Similarly, we can come to the conclusion that for constructed response items the discrimination power decreases if one has to use knowledge of relationships in order to solve the item (in comparison to items which assess declarative knowledge), whereas for multiple choice items the effect of increasing knowledge complexity is significantly different.

On the one hand, for multiple choice items there is a larger probability for the item to be solved correctly only by chance and this could explain the main effect of the "Item openness" variable reflected in the larger discrimination power of constructed response items. Further, the nature of interaction effects suggests that high achievers often better utilize the potentials of multiple-choice items (e.g. the possibility to check if their solution is among given options, thought guiding features etc.), which require the use of analytic representation or the use of

physics relationships. It seems as if even high achieving students from Bosnia and Herzegovina do not possess the effective conditional (situational) knowledge of physics often necessary for solving constructed response items.

The influence of the "Correspondent grade" predictor depends on the complexity of the knowledge structure relevant for item solving. Eighth grade items that require the use of relationships are more discriminative than correspondent seventh grade items and this difference is significantly different from the difference of eighth and seventh grade items, which require the use of declarative knowledge. For the interaction term "Grade * Related relationships" we obtained a similar effect. In other words, if we increase the grade by one, the discrimination power of items, which tap higher knowledge structures, increases, whereas items that tap declarative knowledge become less discriminative. This result indicates that high achievers could be superior in comparison to low achievers with respect to some characteristics of long term memory related to (factual) physics knowledge.

Conclusions

Based on the evaluation of the obtained results and on the categorization of the discussed predictors, it is possible to single out the following cognitive categories that influence the physics item discrimination power:

- complexity and automaticity of knowledge structures which are relevant for generating the most probable correct solution,
- retention of physics factual knowledge,
- cognitive load,
- the predominantly used type of knowledge representation,
- nature of interference effects of relevant formal physics knowledge structures and correspondent intuitive physics knowledge structures (including p-prims).

By taking into account qualities of complexity, automaticity and modality of knowledge, the model obtained within this study is in accordance with the model of types and qualities of knowledge by deJong & Fergusson-Haessler (1996). Further, the model includes factors, which reflect specificities of the physics cognitive domain (e.g. interference effects of intuitive and formal physics). Apart from the described cognitive categories, the item discrimination power depends on some statistical measures, as well – it is influenced by the probability of guessing the correct solution only by chance.

Within this study we have also showed that even high achieving students from Bosnia and Herzegovina often do not develop competencies related to combining physics relationships, especially if these actions need to be performed within the analytical representation. They mostly lack conditional (situational) knowledge, which is necessary for solving items presented in new contexts. We can conclude that even the knowledge of high achievers from Bosnia and Herzegovina is often inert. Furthermore, we noted that it should be paid more attention to creating links between the intuitive and formal physics knowledge structures, with the purpose of helping low achievers to improve their learning.

Generally, in order to improve the quality of physics education in Bosnia and Herzegovina, we have to reconsider the existing culture of setting and solving physics questions

and problems in our schools. It is of particular interest to introduce new types of physics problems that possess a greater potential of eliciting higher cognitive processes.

Besides providing feedback for physics education at the primary school level in Bosnia and Herzegovina, the obtained model of physics item discrimination power could improve the test design process for purposes of future large-scale assessments. Thereby, we could use the knowledge on item features that influence its discriminatory behavior, with the purpose of modifying items with originally poor psychometric characteristics.

Finally, we should take into account that the functionality of the obtained model is limited to samples of students at the end of compulsory education in Bosnia and Herzegovina, whereat the physics items should not elicit competencies that are fundamentally different to competencies we took into account within the process of model development.

References

Alexander, V. E. (1960). Seventh graders' ability to solve problems. *School Science and Mathematics, 60*, 603–606.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303-310.

Chalifour, C., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26*, 120-132.

Clement, J. (1994). Use of Physical Intuition and Imagistic Simulation in Expert Problem Solving. In D. Tirosh (Ed.), *Implicit and explicit knowledge*. Hillsdale, NJ: Ablex Publishing Corp.

de Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*, 105-113.

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction*, 10, 105 225.

Draxler, D. (2005). *Aufgabendesign und basismodellorientierter Physikunterricht*. Ph.D. thesis, Universitaet Duisburg-Essen.

Field, A. (2005). *Discovering statistics using SPSS*. London: SAGE.

Halloun, I.A. (2006). *Modeling Theory in Science Education*. Dordrecht: Springer.

Miles, J., & Shelvin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. London: SAGE.

Henderson and L. Hsu. (Eds.). *Proceedings of Physics education Research Conference 2008*. NY: 2008.

Kauertz, A. (2007). *Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben*. Ph.D. thesis, Universitaet Duisburg-Essen.

Kelley T.L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*, 17-24.

Martin, M.O., Mullis, I.V.S., Foy, P. (2008). *TIMSS 2007 International Science Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Marshall, S. (1988). Assessing problem solving: A short term remedy and a long term solution.

In R. I. Charles & E. A. Silver, *The teaching and assessing of problem solving* (pp.159–177). Reston, VA: The National Council of Teachers of Mathematics.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 248*, 122-130.

Mesic, V., & Muratovic, H. (2011). Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics Physics Education Research 7*.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

Olson, J.F., Martin, M.O., & Mullis, I.V.S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: Boston College.

Petrovic, L. (2006). *External assessment of student achievement at primary school level*: *An expert's report*. Sarajevo: Standards and Assessment Agency for Federation of BiH and RS.

Rosca, C. V. (2004). *What makes a science item difficult? A study of TIMSS-R items using regression and the linear logistic test model*. Ph.D. thesis, Boston College, Boston.

SAA 2006 Database available at the Sarajevo office of the Agency for Pre-school, Primary and Secondary Education in BiH (2006).

Schütz, G. (2006). School Size and Student Achievement in TIMSS 2003. In T. Loveless (Ed.), *Lessons Learned: What International. Assessments Tell Us about Mathematics Achievement*. Washington: Brookings. Institution Press.

Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Teodorescu, R., Bennhold, C., & Feldman, G. (2008). Enhancing Cognitive Development through Physics Problem Solving: A Taxonomy of Introductory Physics Problems. In M. S. C.

TIMSS 2007 International Database available at http:// timss.bc.edu/timss2007/idb_ug.html