*Interdisciplinary Description of Complex Systems* 10(3), 270-283, 2012

# USING OF STRUCTURAL EQUATION MODELING TECHNIQUES IN COGNITIVE LEVELS VALIDATION

**Natalija Ćurković***

Zagreb, Croatia

DOI: 10.7906/indecs.10.3.5
Regular article

*Received:* 24 August 2012.
*Accepted*: 18 October 2012.

## ABSTRACT

When constructing knowledge tests, cognitive level is usually one of the dimensions comprising the test specifications with each item assigned to measure a particular level. Recently used taxonomies of the cognitive levels most often represent some modification of the original Bloom's taxonomy. There are many concerns in current literature about existence of predefined cognitive levels. The aim of this article is to investigate can structural equation modeling techniques confirm existence of different cognitive levels. For the purpose of the research, a Croatian final high-school Mathematics exam was used ($N$ = 9626). Confirmatory factor analysis and structural regression modeling were used to test three different models. Structural equation modeling techniques did not support existence of different cognitive levels in this case. There is more than one possible explanation for that finding. Some other techniques that take into account nonlinear behaviour of the items as well as qualitative techniques might be more useful for the purpose of the cognitive levels validation. Furthermore, it seems that cognitive levels were not efficient descriptors of the items and so improvements are needed in describing the cognitive skills measured by items.

## KEY WORDS

cognitive levels, structural equation modeling, validity

## CLASSIFICATION

APA: 2820
JEL: I21
PACS: 89.90.+n

*Corresponding author, $\eta$: natalija.curkovic@gmail.com; +385 1 5577215;
Lopašićeva 14, HR-10000 Zagreb, Croatia

# INTRODUCTION

One of the most influential taxonomies of educational outcomes based on the levels of cognitive processes is for sure the one proposed by Bloom in 1956. The Bloom's taxonomy (BT) represents a classification of six cognitive processes: knowledge, comprehension, application, analysis, synthesis and evaluation [1]. All categories, except the knowledge, form together the "abilities and skills". The basic assumption is that the categories lie on the continuum which represents a cumulative hierarchical structure [2]. The existence of this framework has several important consequences for the whole system in which the educational process takes place, from the experts who make decisions on the curriculum, to the teachers and students.

The authors of the BT believed that they had created a common framework for the classification of the educational outcomes that can contribute in positive change and development of item writing for the large scale assessments. Hence, they believed that the taxonomy can affect the testing procedures and create new ideas and paradigms within the field of testing. One of the primary goals of the BT was to emphasize that the simple information recognition and recalling are not the only aims of the education. The BT underlined the importance of many different cognitive processes testing rather than asking only for factual knowledge.

Very soon after its origination the BT became very popular and has been highly accepted both among educational scientists and practitioners. Thus, numerous research works aimed to determine all the procedures in which taxonomy can be applied and how to incorporate the taxonomy in the specific educational fields. Researchers also became very interested in investigating whether the BT's proposed hierarchical structure is truly presented in the real tests and is it possible to construct a test that would measure six different levels of cognitive processing. Therefore, the problem of the BT's validity became a serious research issue.

One of the first and most important research among the researches on the BT's validity was definitely the one conducted by Kropp and Stoker in 1966. Many other authors replicated lately their study or used their data in their own studies in which many of them intended to develop new approaches in the BT's validation. Kropp and Stoker [3] constructed four knowledge tests consisted of the six subtests. Each of the subtests was constructed in order to measure one cognitive level proposed by the BT. The tests were administered to the 15 to 17 year old students. In addition to the knowledge tests, a battery of the intelligence tests was delivered. Kropp and Stoker confirmed the cumulative hierarchical structure only for the first four levels. They also found that the correlation between each subtest and the general intelligence factor increases when the cognitive level measured by the subtest is becoming higher.

Madaus, Woods and Nuttal [4] taken over the data gathered by the Kropp and Stoker [3] and used causal modeling methodology in examining the relationship between the results of knowledge tests and the general intelligence factor. They found a letter Y-shaped model where the base is consisted from the first three levels (knowledge, understanding and application). Then, one branch goes from the application to the analysis and the other one from the application to the synthesis and evaluation. They concluded that the obtained Y-shaped model can be identified as the Cattell's model of the fluid and crystallized intelligence where the first branch goes from knowledge to the analysis approximates the crystallized intelligence. The second branch consisted of the synthesis and evaluation could be identified with the fluid intelligence. These results were confirmed in the study conducted by Miller, Snowman and O'Hara in 1979 [5].

Similarly to the results of the previous research, Smith [6] showed in his study that the BT can be divided into two parts. The first four cognitive levels (knowledge, comprehension,

application, analysis) go to the first part and the last two categories (synthesis and evaluation) go to the second part. His interpretation was different from those given by Madaus, Woods and Nuttal [4]. Smith linked the two obtained parts of the BT to the intelligence and creativity. He found that the first part correlates only with the intelligence and the second part both with the intelligence and creativity. It is similar to the modern construct of the divergent and convergent production.

The functioning of the cognitive taxonomy within the test specification of an allied health certification examination was studied in the Webb, Kalohn and Cizek's [7] research. The taxonomy used was a simplified BT in which items were classified as comprehension, application and analysis. A factor analysis of responses did not support expected cumulative hierarchical model of the cognitive complexity. The results of the factor analysis suggested unidimensional solution.

Gierl [8] in 1997 conducted an examination to determine whether the BT can provide an accurate model to guide item writers for anticipating the cognitive processes used by students on a large-scale achievement test in Mathematics. Thirty seventh graders were asked to think aloud as they solved items on the Mathematics test. Their responses were classified with a coding system based on the BT. The overall match between the cognitive processes expected by the item writers and those observed from the students was about 54 %. The author concluded that the BT does not provide an accurate model for guiding item writers to anticipate the cognitive processes used by students.

Lipscomb [9] in 2001 compared in his study a six level semantic differential scale using the bipolar terms "simple" and "complex" to the BT for classifying eighteen test questions. The participants were junior college faculty who had participated in one of the two instructional sessions: a BT session or a semantic differential session. The proportion of responses of each group was compared on each of the eighteen questions using the chi square statistic. The result showed that there was no difference between classifying the items according to the BT and to the six level complexity scale. Hence, the study showed that the BT does not represent an improvement over the scale "simple-complex".

Most of the conducted studies pointed out that the construct validity of the Bloom's taxonomy is questionable and that the dimensions could be replaced by simpler concepts such as the complexity of the items [9] or with some model of intelligence such as Cattell's [4]. Despite the lack of evidences that would confirm existence of different cognitive levels in knowledge tests, original or some form of the revised versions of Bloom's taxonomy are used in modern educational systems round the world [10]. Accordingly, abbreviated version of the Bloom's taxonomy consisted of the first three levels became the central model used in development of the state-level tests in Croatian educational system.

Although many theorists suggested discarding the BT as well as the other classification schemes arised from it [11], contemporary researchers [12-13] believe that the taxonomy itself is not so problematic but its use. They strongly recommend using the taxonomies when constructing the items rather than using them as a part of post-hoc item specification procedures which has become customary. Furthermore, they claim that the more complex methodological procedures, can classify the items into cognitive categories more accurately than the traditional single-method approaches.

## STRUCTURAL EQUATION MODELING

Since the researches on the validity of the BT were mainly conducted to the mid 90's, methodology of the structural equation modeling (hereinafter: SEM) generally was not used.

Although this methodology dates back to the seventies, the existence of fast and powerful computers was necessary to make it more popular among scientists. As it was described in the introductory part, researchers as Madaus, Woods and Nuttal [4] as well as Miller, Snowman and O'Hara [5] used approaches similar to SEM to test BT validity. Validity researches conducted in the last twenty years have proven usefulness and necessity of SEM for that type of research [14]. Hence, this methodology will be used in this study.

SEM represents a group of methods that are regularly used for representing dependency (arguably "causal") relations in multivariate data in the behavioral and social sciences [15]. Generally, a structural equation model is a complex composite statistical hypothesis. It consists of two main parts: the measurement model represents a set of *p* observable variables as multiple indicators of a smaller set of *m* latent variables, which are usually common factors. The path model describes relations of dependency—usually accepted to be in some sense causal—between the latent variables. The term *structural* model is reserved here for the composite SEM, the combined measurement and path models. SEM provides: testing multivariate hypotheses; testing causal relationships even in the correlation studies (only with a proper research design); testing alternative hypotheses. It allows reducing of number of variables to a simpler model and determination of the mutual effects size between latent and observed variables [14]. Thus, it can be said that SEM represents an analytical framework that unifies several multivariate methods which purpose is to provide meaningful and parsimonious explanations of relationships between a set of variables.

SEM consists of the following six steps: model specification, model identification, construct operationalization, parameter estimation, hypotheses evaluation, and, model respecification [14]. The first step is model specification, which means the representation of the research hypotheses in the form of structural models. This involves drawing a model diagram using a set of more or less standard graphical symbols or writing of series of equations. These equations define the model's parameters, which correspond to presumed relations among observed and/or latent variables that will be estimated with sample data. Regardless of the representation form (graphical model or set of equations), the model must contain clearly defined parameters that indicate the relationships between observed and latent variables. This model serves as a framework for testing the sample data. Model specification is probably the most important step. This is because results from later steps assume that the model is basically correct.

The next step is to identify the model and parameters. The model is identified if there is a theoretical possibility to derive a unique estimate for each parameter of the model. The term parameter refers to a numerical value that describes some aspect of the model in the population [16]. One of the main goals of the SEM is to assess as precise as possible the values of model parameters.

The third step is a selection of measures which will be represented by variables in the model. This step also includes data collection and their screening. This is followed by the next step: analysing the model. The fourth step involves assessing the parameters values and their fit to the collected data. There are many methods used to estimate parameters, but the most usual one is maximum likelihood method (ML). ML estimates all parameters simultaneously. The name "maximum likelihood" describes the statistical principle that takes place during the process of parameters derivation: the estimates are the ones that maximize the likelihood (the continuous generalization) that the data (the observed covariances) were drawn from this population [17]. This method has iterative nature which means that the computer program derives the initial solution and then tries to improve the estimation. The improvement means that the covariance model in each of the next iteration is more similar to the observed covariance.

The penultimate step is the evaluation of the proposed model with the observed data. If it turns out that the fitting is satisfactory, the following actions are interpretation of the parameters and consideration of equivalent of near-equivalent models. Equivalent model explains the data just as well as the proposed model but does so with a different configuration of hypothesized relations among the same variables.

If the proposed model does not fit well to the observed data, respecification and evaluation of the revised model should be done on the same data. As an initial specification, this new research should be guided by hypotheses.

There is a large number of goodness-of-fit indices that can be used when evaluating the model fit. Those indices can be divided into several groups: comparative, absolute, parsimonic, residual indices, and the proportion of variance explained by the model [18]. Therefore, selection of the specific index is not easy. However, high-quality models have desirable measures of fitting regardless of the choice of indicators. If different indicators give conflicting and inconsistent information, it is recommended to reconsider the model. Whatever combination of indices is selected, it should be taken into account several things. First, the values of these indices show an average fitting between models and data. Thus, it may happen that some parts of the model fit poorly with the data even though the overall indicators of fitting have the optimum values. In addition, the goodness-of-fit indices do not say anything about the theoretical meaning and significance of the results [19]. Therefore, even if the indices are satisfactory, theory driven interpretation of the parameters is critical for the model evaluation. And finally, values of goodness-of-fit indices do not say anything about the model predictive power. In conclusion, the model evaluation and testing of its fitting with the collected data is not a binary decision whether the model fits or not, but it is rather a process in which is more appropriate to describe model with the terms such as: reasonable, adequate, satisfactory, etc. with referring to a number of criteria.

## OBJECTIVE

When constructing knowledge tests, cognitive level is usually used in the test blueprint to help describe what each item is designed to measure. Recently used taxonomies of the cognitive levels most often represent some modification of the original Bloom's taxonomy. There are many concerns in current literature about existence of predefined cognitive levels. Accordingly, the aim of this article is to investigate do structural equation modeling techniques confirm the existence of different cognitive levels?

It is expected to get three predefined cognitive levels: knowledge, comprehension, and application. That would confirm functionality and purpose of using cognitive levels in the test construction. Otherwise, if it is not possible to find different cognitive levels even if they were used while constructing the items, last few decades of practice using cognitive taxonomies for test construction could be reconsidered.

## METHODOLOGY

For the purpose of this research, a Croatian final high-school Mathematics test was used as a central instrument. It was administrated in June 2010 to $N = 9626$ senior high-school students. The test consisted of 45 items: 15 multiple-choice and 30 open-ended. Some of the items were polytomously scored so the highest possible score on the test was 60 points. Subject-matter specialists who constructed the test classified the items into three categories based on the cognitive level that they supposed to measure: knowledge, comprehension, and application. These categories represent the abbreviated version of the Bloom's six-level

hierarchical taxonomy. Hence, the assumption is that the items that measure higher levels require mastering the items that measure lower cognitive levels.

To examine the existence of three-level hierarchical structure, Mathematics test data were analysed by using confirmatory factor analysis (CFA) and structural regression model (SR). Both of these methods can be considered as the SEM techniques. The technique of CFA analyses measurement models in which both the number of factors and their correspondence with the indicators are explicitly specified. Standard CFA models have the following characteristics: (i) each indicator is a continuous variable represented as having two causes - a single factor that the indicator is supposed to measure and all other unique sources of influence (omitted causes) represented by the error term; (ii) the measurement errors are independent of each other and of the factors; (iii) all associations between the factors are unanalysed (the factors are assumed to covary). The assumption under which CFA was conducted was the next: if the cognitive complexity differentiates the test items, then cognitive complexity levels should appear as different factors. Under a similar assumption as [7] conducted an exploratory factor analysis (EFA) in their study with the similar hypothesis. In this study, CFA was used rather than exploratory EFA for several reasons. As it was previously mentioned, analyses the measurement model in which the number of factors and their relationships with indicators are explicitly defined. Therefore, it can be implemented only if the theory that is being tested is well established. Since in this case there is a clear hypothesis of the existence of three latent variables that is theory driven, CFA is a logical choice.

The CFA tests only hypothesis about existence of different cognitive levels. To test hypothesis about relationships among latent variables (cognitive levels) and their directions, it is necessary to conduct SR model. It is probably the most general kind of core structural equation model. An SR model comprises both structural and measurement model which makes possible to test at the same time both structural and measurement relations. The aim of the SR model in this case was to determine the relationships among latent variables marked as cognitive levels.

Beforehand, items were parcelled in order to reduce number of indicators per first two latent variables (cognitive levels). Since there were only five items supposed to measure the third level, they were note parcelled. There are few reasons why the parcelling was done. Bandalos and Finney [20] offered two categories of argument in favour of parcels. The first category is oriented on the differing psychometric characteristics of items and parcels. When comparing the separate items with aggregate-level data or parcels, item-level data contain at least one of the following disadvantages: lower reliability, lower communality, a smaller ratio of common-to-unique factor variance, and a greater likelihood of distributional violations. Item-level data also have fewer, larger, and less equal intervals between scale points than do aggregate-level data [21-22]. The stated arguments are related to the basic psychometric theory. The second category of argument is focused on the factor-solution and model fit. The various indexes of model fit are expected to be more acceptable when parcels, rather than items, are used because of the psychometric and estimation advantages of parcels. Compared with the item-level data, models based on parcelled data have fewer estimated parameters in defining a construct as well as in representing the whole model. Hence, these models are more parsimonious. Furthermore, their residuals have smaller chances to be correlated, and they reduce various sources of sampling error [22].

There was one more important reason for parcelling. The test used in this study is mostly consisted from dichotomous items. Since the structural equation techniques are primarily developed to deal with the continuous variables [14], the grouping items into parcels was necessary in order to create suitable indicators for SEM analysis. An important assumption that has to be satisfied when creating parcels is the unidimensionality of the items to be

parceled. To determine the dimensionality of each sets of items to be parcelled (those supposed to measure first two cognitive levels), principal component analysis were conducted. The first component for both item sets have explained about 30 % of the variance and the ratio of the first two eigen-values were bigger than five. According to Hattie [23], obtained results indicate the unidimensionality. The first item set contained twelve items and the second set had twenty-eight items. Since the optimal number of indicators per latent variable is three to four [14], items from the first item set were aggregated into three parcels consisted of four items. The second item set was grouped into four parcels made of seven items. Parcelling was done using the "item-to-construct balance" method proposed by Little, Cunningham, Shahar, and Widaman [24]. The aim of that method is to create parcels that are equally balanced in terms of their difficulty and discrimination.

When the parcelling was done, three alternative models were proposed and tested. The analyses were done with LISREL 8.80 software. Covariance matrix used in the analyses is presented in the Appendix.

CFA was firstly conducted. Its aim was to check the existence of three latent variables which represent cognitive levels. As it can be observed from the graphical representation (Model A in Figure 1), the model allows covariations among latent variables which are expected based on the Bloom's taxonomy assumptions. Prior to the parameters estimation, model identification was checked. There are some straightforward rules for CFA models that concern minimum numbers of indicators per factor. For standard CFA it means that every indicator loads on just one factor and there are no measurement error correlations. More precisely, according to the two-indicator rule, a standard CFA model is identified if every factor has two or more indicators, each row of the $\Lambda$ matrix (matrix of factor loadings $\lambda$) has only one nonzero element; error covariance matrix ($\delta$) is diagonal, and has zero elements outside the diagonal; and variance/covariance matrix of latent variables ($\Phi$) has at least one nonzero element [14]. These requirements are all met in the CFA model presented in this article.

Parameter estimation was made using maximum likelihood estimation on covariance matrices. Covariance matrix for the observed data is given in the Table 1. When the parameters were estimated, four different goodness-of-fit indicators were considered: Chi-Square ($\chi^2$), comparative fit index (CFI), root mean square error of approximation (RMSEA), and Akaike information criterion (AIC).

Since the standard CFA is limited on testing whether latent variables exist or not rather than explaining their relationships, two alternative models were built and tested in order to provide a clearer insight into the nature of relationships among latent variables, as well as among observed and latent variables. Both of them are grounded in the BT assumptions and the findings of the previous studies reported in the introductory part of the article. The first proposed SEM model, model B (Figure 1), represents the expected hierarchical cumulative structure of the BT. The expected structure is operationalized through the latent-growth model in which each previous level "cause" the next one. Usually, that type of models can be found in the longitudinal studies where are they employed in order to control high covariations between

**Table 1.** Goodness-of-fit statistics for tested models.

| Model | Goodness-of-fit Statistics | | | |
|---|---|---|---|---|
| | $\chi^2$ | RMSEA | AIC | CFI |
| A | 1874,5 ($N$ = 9626, df = 51, $p$ < 0,001) | 0,061 | 1938,0 | 0,988 |
| B | 1875,8 ($N$ = 9626, df = 52, $p$ < 0,001) | 0,061 | 1944,1 | 0,988 |
| C | 1874,5 ($N$ = 9626, df = 52, $p$ < 0,001) | 0,061 | 1938,0 | 0,988 |

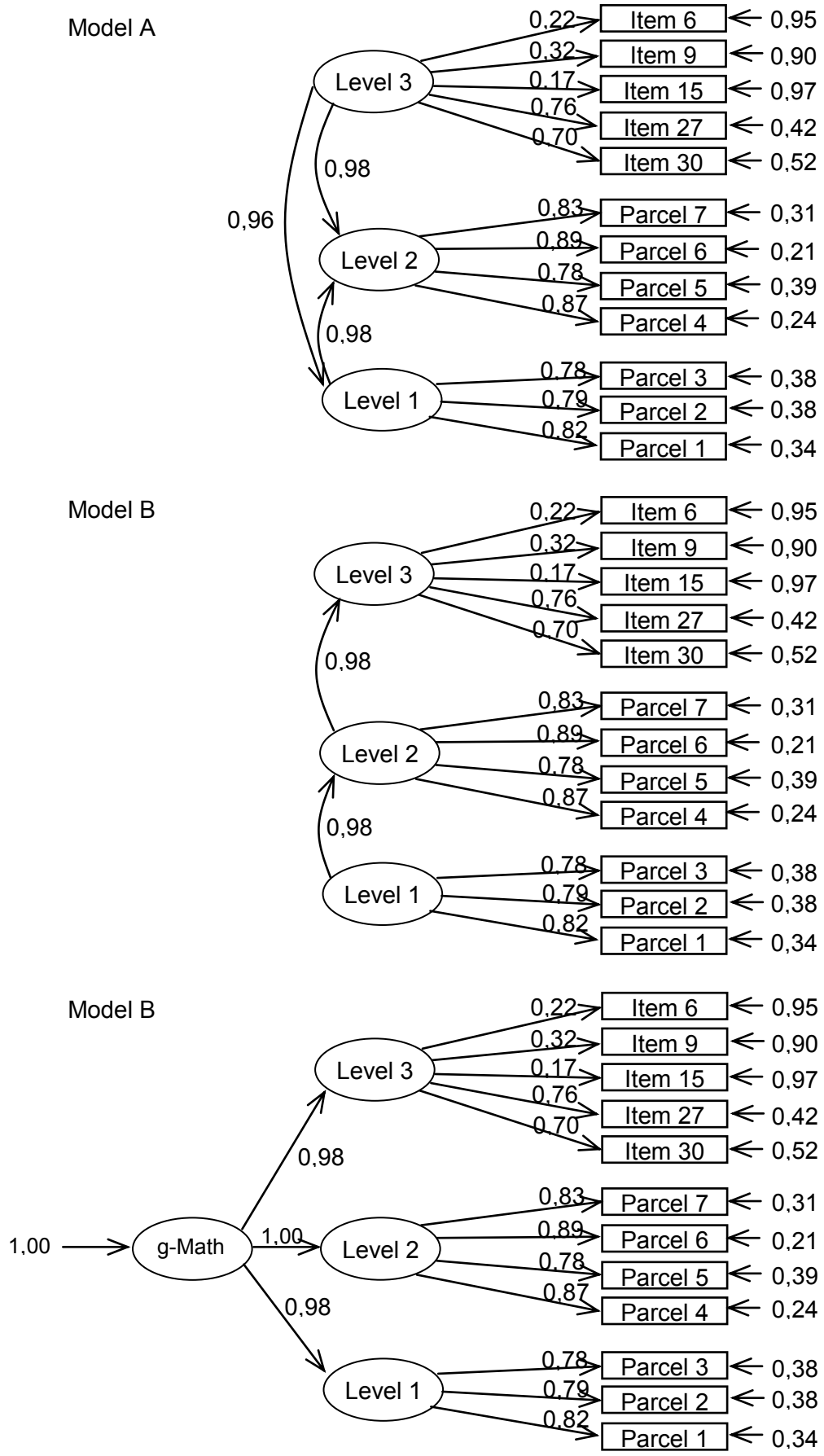


**Figure 1.** Tested models A, B and C of cognitive levels. The path coefficients are standardized.

measurements of the same construct in different points in time. It should be noted that the model is recursive which automatically means that it is identified [14].

In the second model, model C (Figure 1), higher order factor was introduced. In its nature it is an extension of the standard CFA model. The second-order factor was named "general knowledge of Mathematics". The model is mostly based on the results of earlier studies which dominantly did not confirm expected hierarchical cumulative structure. Furthermore, studies that included testing of the general intelligence factor showed its strong relationship with the cognitive levels. Even though the intelligence was not measured as a part of this study, it is expected that the general knowledge can also successfully explain the most of variance of the test achievement [25]. Therefore, it seems plausible to predict that the general knowledge of Mathematics would be a factor responsible for the covariations between the first-order latent variables.

For both models parameters were estimated by the maximum likelihood method and the goodness-of-fit statistics were checked in the same manner as for the first CFA model.

## RESULTS AND DISCUSSION

The set of three latent models was tested. Their purpose was to examine the existence of three latent variables responsible for covariations among the items that measure specific cognitive level. The tested models with obtained standardized parameters are shown in Figure 1. Goodness-of-fit indices related to each model are presented in Table 1.

The model A was the simplest and it should have tested only the existence of three predefined factors or cognitive levels. Comparative and parsimonic indices of goodness-of-fit (RMSE, CFI, and ACI) for this model indicate excellent agreement between estimated and observed parameters. However, the estimated standardized parameters between latent variables were almost equal to one which indicates strong unidimensionality rather than existence of three separate factors. Therefore, two additional models were introduced. Their purpose was a reduction of the variance among latent variables and provision of a clearer picture of the relationships between latent variables and their relations to observed variables.

The first of these two models, model B, was a hierarchical factor model. Hierarchical confirmatory factor analysis models depict at least one construct as a second-order factor that is not directly measured by any indicator. This second-order factor is also presumed to have direct effects on the first-order factors, which have indicators. These first-order factors do not have unanalysed associations with each other. Instead, their common direct cause, the second-order factor, is presumed to explain the covariances among the first-order factors [14]. In this case the model was introduced with the expectation that it would "pick up" the variance among the lower order factors (cognitive levels). Such a solution agrees with the theory because the second-order factor can be explained as a general factor of mathematical abilities, skills and knowledge. Some earlier attempts of BT validation also used the general factor in explaining covariations between cognitive levels [3]. However, introduction of general factor did not change an impression that the data are strongly unidimensional. Relationships between each first-order factor and second-order factor are almost equal and extremely high.

Model C was built under the assumption of the cumulative-hierarchical structure of the BT. It represents an example of the latent-growth models that take into account high covariations among latent variables. However, this model revealed similar results as the previous two models. In other words, covariations among latent variables remained extremly high which indicate the existence of unidimensional construct. These findings are consistent with some previous research findings that used SR models. Thus, Hill [26] and Hancock [27] generally

failed to confirm the cumulative-hierarchical structure of the BT. Furthermore, similar results to these, but obtained by using exploratory factor analysis were reported by Webb, Kalohn and Cizek [7]. They also confirmed existence of only one factor instead of three.

There are several possible explanations for the findings of this study. The most likely one is related to the construction of the test. The aim of the construction is to select the best possible items on which results could be made valid conclusions about student achievement and knowledge. One of the most important indicators of the quality of items is their discriminative power. Hence, the items for the Mathematics test were mostly selected based on their discrimination parameters. Although the tests within the Croatian state level testing program are not pretested, content method specialists who build the tests use the results of previously delivered tests. They are advised to retain item types or to clone the items that proved to be very discriminative in the previous applications. Item cloning is a practice that is well established in the modern testing centres in order to reduce expenses related to the construction of new items [28]. Glas, and van der Linden [29] showed in an extensive study that cloned items remain very similar psychometric characteristics as "items-parents". According to fact that the Mathematics test included only those items which "parents" showed high discrimination in the past five years, it was expected to have items with very similar psychometric characteristics. In other words, it was expected to have items with high discrimination parameters (above 0,3 in the classical test theory terms). The average value of the discrimination parameters in this test was 0,45 which indicates that the tasks were carefully chosen; such high values could not be achieved by random selection, or only based on the item content. Thus, a common feature of all items was their high discrimination. It is likely that the very high discriminations were the common factor or the variable that created high covariations among the latent variables. This situation is not uncommon in the test preparation and there are numerous examples where the psychometric characteristics of items, such as difficulty or discrimination, strongly affect the dimensionality of a test [30-31].

The next finding to which attention should be paid is goodness-of-fit of the tested models. The parsimonic and comparative indices showed excellent agreement between model and the data while the absolute index ($\chi^2$) and standardized residuals indicated complete disagreement between the models and data. Chi-square in this case is not so problematic because of the large number of participants. The chi-square statistic is a sample size sensitive. When using the SEM techniques, it is recommended to work with the large samples because of the more accurate estimates. At the same time, with the large samples null hypothesis that the model fits the data is with chi-square almost always rejected. That is why chi-square in this case is not suspicious. However, high values of standardized residuals are not expected. It is surprising that the differences between observed and estimated covariance matrices are so big. Another unexpected findings are extremely high regression coefficients obtained between the latent variables. Such high regression coefficients together with the high standardised residuals indicate the existence of certain problems with the data. Since there are no clear guidelines for the interpretation of the residuals [14, 32], it is necessary to consider all the characteristics of the results obtained by the different analyses. One of the possible explanations for such high values of standardised residuals is a fact that the various analyses of the observed variables showed up a number of overlaps between the subtests that are supposed to represent the factors, which indicate nonlinear relationships. On the other hand, SEM techniques are often referred as linear structural modeling which emphasizes that those techniques assume linear relationships between variables. It is possible that the nonlinearity makes impossible to find a structural model that would fit the data and also confirm the predicted structure.

Based on the stated arguments, it could be concluded that the structural evidences of the BT validity are not unambiguous. The analyses of latent variables indicate strong

unidimensionality which can be reasonably explained by the principles used in construction of the test, or extremely high discrimination indices.

## CONCLUSION

Cognitive levels are commonly used in item writing and test construction. An extensive SEM methods study was conducted to investigate whether the cognitive levels used in item and test construction really exist. Results obtained by confirmatory factor analysis and structural regression model indicated the existence of strong unidimensionality. Even though the goodness-of-fit statistics for the three proposed models suggest good fit of each of the models, the relationships between cognitive levels are so high that it is not realistic to keep them as separate latent variables. Such strong relationships between levels are possibly the consequence of the item selection procedure. The psychometric criterion was their discrimination parameters. Since they all have psychometrically similar characteristics, it is not surprising that they revealed unidimensionality. The results of this research suggest retaining the cognitive levels as an important part of the test construction procedures but also their reconsideration directed towards making them more operationalizable.

Studies results of the SEM methodological approaches used in this study should be compared in the future to some of the cognitive-psychometric models which allow for modeling the relationship between the item responses and student proficiency in various cognitive processes [12, 13, 33]. Also, new studies emphasize the role of qualitative research methodology in examination of the cognitive levels used to solve the items [34].

## APPENDIX

**Table 2.** Covariances among observed variables ($N = 9626$).

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parcel 1 | 1,858 | | | | | | | | | | | |
| Parcel 2 | 1,219 | 1,910 | | | | | | | | | | |
| Parcel 3 | 1,296 | 1,304 | 2,288 | | | | | | | | | |
| Parcel 4 | 1,214 | 1,147 | 1,272 | 1,561 | | | | | | | | |
| Parcel 5 | 1,913 | 1,893 | 2,139 | 1,965 | 5,475 | | | | | | | |
| Parcel 6 | 1,424 | 1,365 | 1,513 | 1,394 | 2,327 | 2,101 | | | | | | |
| Parcel 7 | 1,342 | 1,354 | 1,460 | 1,387 | 2,371 | 1,621 | 2,301 | | | | | |
| Item 6 | 0,279 | 0,194 | 0,294 | 0,233 | 0,463 | 0,289 | 0,221 | 1,000 | | | | |
| Item 9 | 0,340 | 0,348 | 0,388 | 0,328 | 0,602 | 0,385 | 0,451 | 0,207 | 1,000 | | | |
| Item 15 | 0,092 | 0,070 | 0,057 | 0,159 | 0,425 | 0,191 | 0,205 | 0,016 | 0,053 | 1,000 | | |
| Item 27 | 1,334 | 1,421 | 1,574 | 1,352 | 2,390 | 1,603 | 1,614 | 0,140 | 0,371 | 0,416 | 2,816 | |
| Item 30 | 1,742 | 1,574 | 1,800 | 1,648 | 2,926 | 2,042 | 1,890 | 0,386 | 0,568 | 0,389 | 1,980 | 5,154 |

## ACKNOWLEDGMENTS

This work is financed by the National Foundation for Science of the Republic of Croatia.

## REFERENCES

[1] Bloom, B.S., ed.: *Taxonomy of educational objectives. Handbook 1: Cognitive domain.* Longman, New York, 1956,

[2] Anderson, L. and Krathwohl, D.R., eds.: *A taxonomy for learning, teaching and assessing. A revision of Bloom's taxonomy of educational objectives.* Longman, New York, 2001,

[3] Kropp, R.P. and Stoker, H.W.: *The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives*.
Florida State University, Tallahassee, 1966,
[4] Madaus, G.F.; Woods, E.M. and Nuttall, R.L.: *A causal model analysis of Bloom's taxonomy.*
American Educational Research Journal **10**(4), 253-262, 1973,
http://dx.doi.org/10.3102/00028312010004253,
[5] Miller, W.G.; Snowman, J. and O'Hara, T.: *Application of alternative statistical techniques to examine the hierarchical ordering in Bloom's taxonomy*.
American Educational Research Journal **16**(3), 241-248, 1979,
http://dx.doi.org/10.3102/00028312016003241,
[6] Leon Smith, I.: *Validity of test of the cognitive processes.*
Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972,
[7] Cizek, G.J.; Webb, L.C. and Kalohn, J.C.: *The use of cognitive taxonomies in licensure and certification test development: reasonable or customary?*
Evaluation & the Health Professions **18**(1), 77-91, 1995,
http://dx.doi.org/10.1177/016327879501800106,
[8] Gierl, M. J.: *Comapring cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy.*
The Journal of Educational Research **91**(1), 26-32, 1997,
http://dx.doi.org/10.1080/00220679709597517,
[9] Lipscomb Jr., J.W.: *Is Bloom's taxonomy better than intuitive judging for classifying test questions?*
Education **106**(1), 102-107, 2001,
[10] Booker, M.J.: *A roof without walls: Benjamin Bloom's taxonomy and the misdirection of American education*.
Academic Questions **20**(4), 347-355, 2008,
http://dx.doi.org/10.1007/s12129-007-9031-9,
[11] Wineburg, S. and Schneider, J.: *Was Bloom's Taxonomy pointed in the wrong direction?*
Kappan **91**(4), 56-61, 2009,
[12] Gorin, J. S.: *Test design with cognition in mind*.
Educational Measurement: Issues and Practice **25**(4), 21-36, 2006,
http://dx.doi.org/10.1111/j.1745-3992.2006.00076.x,
[13] Rupp, A.: *Unique characteristics of cognitive diagnosis models.*
Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 2007,
[14] Kline, R. B.: *Principles and practice of structural equation modeling*. 3rd edition.
The Guilford Press, New York, 2011,
[15] McDonald, R. P. and Ringo Ho, M.-H.: *Principles and practice in reporting structural equation analyses.*
Psychological Methods **7**(1), 64-82, 2002,
http://dx.doi.org/10.1037/1082-989X.7.1.64,
[16] Gonzalez, R. and Griffin, D.: *Testing parameters in structural equation modeling: Every"one" matters.*
Psychological Methods **6**(3), 258-269, 2001,
http://dx.doi.org/10.1037/1082-989X.6.3.258,
[17] MacCallum, R. C. and Austin, J. T.: *Applications of structural equation modeling in psychological research.*
Annual Review of Psychology **51**, 201-236, 2000,
http://dx.doi.org/10.1146/annurev.psych.51.1.201,

[18] Ullman, J.: *Structural Equation Modeling*.
In Tabachnik, B.G. and Fidell, L.S.,eds.: *Using Multivariate Statistics*. Allyn and Bacon, Boston, 2001,
[19] Mulaik, S., et al.: *Evaluation of goodness-of-fit indices for structural equation models*.
Psihological Bulletin **105**(3), 430-445, 1989,
http://dx.doi.org/10.1037/0033-2909.105.3.430,
[20] Bandalos, D.L. and Finney,S.J.: *Item parceling issues in structural equating modeling.*
In Marcoulides, G.A. and Schumacker, R.E., eds.: *New developments and techniques in structural equation modeling*. Lawrence Erlbaum Associates, Inc., Mahwah, 2001,
[21] Bagozzi, R. P. and Heatherton, T. F.: *A general approach to representing multifaceted personality constructs: Application to state self-esteem*.
Structural Equation Modeling: A Multidisciplinary Journal **1**(1), 35-67, 1994,
http://dx.doi.org/10.1080/10705519409539961,
[22] MacCallum, R.C.; Widaman, K.F.; Zhang, S. and Hong, S:. *Sample size in factor analysis.*
Psychological Methods **4**(1), 192-211, 1999,
http://dx.doi.org/10.1037/1082-989X.4.1.84,
[23] Hattie, J.A.: *Methodology review: Assessing unidimensionality of a set of test items.*
Applied Psychological Measurement **9**(2), 139-164, 1985,
http://dx.doi.org/10.1177/014662168500900204,
[24] Little, T.D.; Cunningham,W.A.; Shahar, G. and Widaman,K.F.: *To parcel or not to parcel: exploring the question, weighing the merits?*
Structural Equation Modeling: A Multidisciplinary Journal **9**(2), 151-173, 2002,
http://dx.doi.org/10.1207/S15328007SEM0902_1,
[25] Abedi, J.: *NAEP TRP Task 3e: Achievement Dimensionality, Section A.*
National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, 1994,
[26] Hill, P.W.: *Testing hierarchy in educational taxonomies: A theoretical and empirical investigation.*
Evaluation in Education **8**(3), 179-278, 1984,
http://dx.doi.org/10.1016/0191-765X(84)90004-1,
[27] Hancock, G.R.: *Cognitive complexity and the comparability of multiple choice and constructed-response test formats.*
The Journal of Experimental Education **62**(2), 143-157, 1994,
http://dx.doi.org/10.1080/00220973.1994.9943836,
[28] van der Linden, W.J.: *Linking response-time parameters onto a common scale.*
Journal of Educational Measurement **47**(1), 92-114, 2010,
http://dx.doi.org/10.1111/j.1745-3984.2009.00101.x,
[29] Glas, C.A.W. and van der Linden, W.J.: *Computerized adaptive testing with item cloning.*
Applied Psychological Measurement **27**(4), 247-261, 2003,
http://dx.doi.org/10.1177/0146621603027004001,
[30] Cizek, G.J.; O'Day, D.M. and Robinson, K.L.: *Nonfunctioning options: a closer look.*
Educational and Psychological Measurement **58**(4), 605-611, 1998,
http://dx.doi.org/10.1177/0013164498058004004,
[31] Trevisan, M. S., Sax, G. and Michael, W. B.: *The effects of the number of options per item and student ability on test validity and reliability*.
Educational and Psychological Measurement **51**(4), 829-837, 1991,
http://dx.doi.org/10.1177/001316449105100404,
[32] Hayduk, L.A.: *Structural equation modeling with LISREL. Essentials and advances.*
The Johns Hopkins University Press, Baltimore, 1987,
[33] Leighton, J. P. and Gierl, M. J.:(2007*). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes.*
Educational Measurement: Issues and Practice **26**(2), 3-16, 2007,
http://dx.doi.org/10.1111/j.1745-3992.2007.00090.x,

[34] Daniel, R.C. and Embretson, S.E.: *Designing cognitive complexity in mathematical problem-solving items.*
Applied Psychological Measurement **34**(5), 348-364, 2010,
http://dx.doi.org/10.1177/0146621609349801.

# UPORABA STRUKTURALNOG MODELIRANJA U ISPITIVANJU VALJANOSTI KOGNITIVNIH RAZINA

N. Ćurković

Zagreb, Hrvatska

## SAŽETAK

Pri konstrukciji tekstova znanja kognitivne razine uobičajeno se koriste kao dio specifikacije testa gdje se navodi koji zadatak mjeri koju kognitivnu razinu. Najčešće korištene taksonomije kognitivnih razina predstavljaju neku od modifikacija originalne Bloomove taksonomije. U literaturi se navode brojne poteškoće vezane uz korištenje te taksonomije. Stoga je cilj ovoga rada istražiti može li se tehnikama strukturalnog modeliranja potvrditi postojanje različitih kognitivnih razina.

U svrhu istraživanja, korišten je ispit iz matematike primijenjen na hrvatskoj državnoj maturi u ljeto 2010. ($N = 9626$). Tri različita modela testirana su pomoću konfirmatorne faktorske analize te strukturalno-regresijskih modela. Rezultati dobiveni primjenom ovih tehnika ne podržavaju postojanje različitih kognitivnih razina te se takvi nalazi mogu interpretirati na više načina. Neke druge statističke metode koje uzimaju u obzir nelinearno „ponašanje“ zadataka mogle bi možda biti učinkovitije u postupku validacije kognitivnih razina. Nadalje, čini se da kognitivne razine definirane modificiranom Bloomovom taksonomijom nisu osobito učinkovit deskriptor zadataka te su očito potrebne promjene i poboljšanja u korištenju taksonomija kognitivnih razina pri opisivanju zadataka.

## KLJUČNE RIJEČI

kognitivne razine, strukturalno modeliranje, ispitivanje valjanosti