ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 219-234

An Example of Empirical and Model Based Methods for Performance Descriptors: English Proficiency Test *

Serkan ARIKAN ** Sevilay KİLMEN *** Mehmet ABİ **** Eda ÜSTÜNEL *****

Abstract
Great emphasis is given to the development of high-stake tests all around the world and in Turkey. However, limited emphasis is given to adequate score reporting. Too much emphasis on rankings and almost no emphasis on performance level descriptors (meaning of the scores) have leaded a "ranking culture" in Turkey. There is an immense need to raise awareness about score reporting and performance level descriptions in Turkey. This study aims to raise awareness about the use of performance level descriptors in a high-stake exam in Turkey, an English proficiency exam. The study sample is consisted of 630 undergraduate students who took the 2016-2017 English proficiency exam of a public university in the southwest of the Turkey. In order to identify the potential exemplars, two types of item mapping methods (i.e. experimental based method and model-based method) were used in the present study. Item grouping for performance level descriptors provided hierarchical and interpretable structure. Using these performance level descriptors, it is possible to give criterion referenced feedback to each student about his/her reading abilities.

Key Words: Criterion referenced assessment, performance level descriptors, empirical method, model-based method, construct map.

INTRODUCTION

Every year many exams were prepared to evaluate student performances and to give pass or fail decisions all around the world. Generally, great emphasis is given to the development of these high-stake tests. However, limited emphasis is given to adequate score reporting (Goodman & Hambleton, 2004; Karantonis, 2017). Students get their scores, but they generally do not have any idea what these scores mean. Similarly, instructors give scores to their students, but could not use these scores adequately in their instructions as these scores do not make concrete sense to them, either. In the United States, effort is given to find effective ways to report results of high-stake tests by giving meaning to scores (Karantonis, 2017). The research on standard setting is focusing on which methods are more effective (Karantonis, 2017; Karantonis & Sireci, 2006). Karantonis (2017) stated that there is still a need to examine different item-mapping methods to identify exemplar items for performance level descriptors. However, in Turkey, although exams take a crucial role in every grade level even starting from primary education, very little emphasis is given to score reporting, standard setting procedures and performance level interpretations. Each component of education is strongly affected by high-stake exams; however, stakeholders of education could not interpret and use exam results as no performance level descriptors associated with the scores are given. Students and educators are mainly interested in the normative results such as the rank of students in an exam. Criterion referenced results are very rarely used. Too much emphasis on rankings and almost no emphasis on performance level descriptors

* A part of the study was presented at 2018 EDUCCON Education Conference, Ankara University, Ankara, Turkey.
** Assist. Prof. PhD., Boğaziçi University, Faculty of Education, İstanbul-Turkey, serkan.arikan1@boun.edu.tr, ORCID ID: 0000-0001-9610-5496
*** Assoc. Prof. PhD., Bolu Abant Izzet Baysal University, Faculty of Education, Bolu-Turkey, kaplansevilay@yahoo.com, ORCID ID: 0000-0002-5432-7338
**** Lect., Muğla Sıtkı Koçman University, College of Foreign Language, Muğla-Turkey, mehmetabi@mu.edu.tr, ORCID ID: 0000-0002-4976-5173
***** Prof. PhD., Muğla Sıtkı Koçman University, College of Foreign Language, Muğla-Turkey, eustunel@mu.edu.tr, ORCID ID: 0000-0003-2137-1671

To cite this article:
Arıkan, S., Kilmen, S., Abi, M., & Üstünel, E. (2019). An example of empirical and model based methods for performance descriptors: English proficiency test. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 219-234. doi: 10.21031/epod.477857

Received: 02.11.2018
Accepted: 30.06.2019

have led a ranking culture all over the country. Additionally, there is no public or academic demand to force private and national testing companies to report test results in clear and meaningful way. Turkish teachers reported they rarely use exam results to give feedback compared to European colleagues (Demirtaşlı, 2009). Therefore, there is an immense need to raise awareness about score reporting, standard setting procedures and performance level interpretations in Turkey. As Shulman (2009) stated "assessment is a powerful tool for raising the quality of teaching and learning. It should be used diagnostically and interactively, not as a form of autopsy" (p. 237). We need to use assessment more effectively and this study aims to raise awareness about the use of performance level descriptors in a high-stake exam in Turkey by describing and exemplifying the procedures of defining performance level descriptors. This study shows how a teacher group could get performance level descriptors by using empirical method to get performance level descriptors and also shows how experts could use ConstructMap to get performance level descriptors using model-based methods.

Performance Level Descriptor Methods

There are two major methods for defining performance level descriptors: the empirical method and the model-based method. These methods are described in this part.

The empirical method

The empirical method (Zwick, Senturk, Wang, & Loomis, 2001) corresponds to direct method, defined originally by Beaton and Allen (1992). According to this method, first a few carefully dispersed scale points are determined. These points are called *anchor points* or *anchor levels* and they are defined as judgmental. Then, the student groups at anchor points are determined. But since there may be a small number of students at these points or even no student may be present, a range of points near the anchor points is determined. The items correctly answered by the majority of the students in the range are determined. These items are called exemplars. Finally, the performance represented by these items is defined (Beaton & Allen, 1992).

For example, anchor points can be defined as 10, 20, 30, and 40 on a scale scored from 0 to 50. Regarding how close a point interval to anchor points is to be determined, Beaton and Allen (1992, p. 195) recommended that "this interval should be large enough so that there will be an adequate sample in group k and yet small enough so that the score values are clearly distinguishable from the adjacent anchor points". For the anchor points in the example, near the anchor point can be specified as anchor point ±2. In this case the first anchor point interval is determined as 8 to 12 points. Other anchor intervals are determined by adding and subtracting 2 points. After the near the anchor points are identified, the correct answers are determined by the majority of the students in that range. At this point, what is meant by the majority of students is needed to be operationally defined. Different correct response probabilities (e.g. 50%, 65%, and 80%) have been used in the literature (Beaton & Allen, 1992). One of these probabilities could be selected for this method. For example, if the probability of correct response is identified to be 65%, the items correctly answered by 65% of the individuals in each anchor interval are determined. For each anchor interval, the cognitive and content related properties measured by these items are determined and the performance for each anchor interval is defined.

The model-based method

In model-based method, as in the empirical method, exemplars are chosen based on the probability of correct answer of the item. The difference of the model-based method from the empirical method is that correct response probabilities are estimated based on the item response theory model (Zwick et al., 2001). According to item response theory, ability and item parameters can be placed on the same scale. At this scale, the difficulty parameter of an item is settled at the same time as individuals who are likely to respond to that item by 50%. By utilizing this property of item response theory, it is

possible to find items with 50% probability of responding in a certain proficiency score interval. These items are the items that are likely to be correctly answered 50% by the individuals in this point range (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). For example, the items that individuals in the range of 2.20 - 3.00 points can correctly answer with 50% probability are those with difficulty parameters ranging from 2.20 - 3.00.

As mentioned above, the difference between these two methods is the way in which the response probabilities are calculated. In the empirical method, the response probability is calculated based on the classical test theory, while in the model-based method, it is calculated based on the item response theory.

Purpose of the Study

This study aimed to illustrate how performance level descriptors could be defined using a dataset of an English proficiency test. There is a need to report educational test result more efficiently by developing adequate score reporting methods, especially in Turkey. Providing verbal descriptors for related score intervals, the exam results will be more meaningful and required feedback could be given to stakeholders. An example from a high-stake English proficiency exam was used to illustrate how empirical method and model-based method using ConstructMaps could be applied in practice. With this incentive, the research question of the study is set as How can we define performance level descriptors for an English proficiency exam?

METHOD

This study is a standard setting study that aims to give a meaning to test scores. This study expected to raise awareness about the use of performance level descriptors in a high-stake exam in Turkey. In order to achieve this goal, two item-mapping methods to identify exemplar items for performance level descriptors were used. The participants, instrument and data analysis procedures were described in this section.

Participants

Total of 630 undergraduate students took the 2016-2017 English proficiency exam of a public university in the southwest of the Turkey. Sixty two percent of the students were male, and thirty two percent were females. This public university mainly has programs in Turkish but there are some programs that have the medium instruction in English. The participants of this study were the students who were registered to preparatory class of foreign language school of this university. These students were required to get overall score of 60 out of 100 to start their undergraduate programs.

Data Collection Instrument

This study used English proficiency test to define performance level descriptors. The English proficiency test has four major dimensions: Reading, Listening, Writing and Speaking. This test was developed by test development team of foreign language school of the university. The proficiency test was developed based on the assessment framework of Common European Framework and aimed to be in B1 to B2 level. This study focuses on reading part of this test. Reading part included reading paragraphs and there were 19 items in the format of matching, short answer and multiple choice.

Data Analysis

Preliminary analysis

As a preliminary analysis, internal consistency of reading test was tested using The Cronbach's Alpha reliability coefficient. According to George and Mallery (2003) Cronbach's Alpha coefficient should be higher than .700. An instrument with Cronbach's Alpha coefficient higher than .800 is considered as a good instrument as and higher than .900 is considered as a marvelous instrument. Besides, descriptive statistics related to reading test results were reported. SPSS 22.0 was used to conduct internal consistency and descriptive statistics.

Reading test was developed to measure one main reading ability. Therefore, confirmatory factor analysis was conducted to test unidimensionality of the reading test. Confirmatory factor analysis requires an assessment to establish whether or not the proposed model is a good one. A good model is a model in which the difference between covariance matrix obtained from student data and covariance matrix implied by the hypothesized model is minimum (Ullman, 2001). This difference is evaluated by using several fit indices. Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA) are widely reported fit indices to assess goodness of fit of confirmatory factor analysis. In this study, CFI and TLI values higher than .900 was considered as acceptable fit and .950 and above was considered as good fit; and RMSEA values .080 or less was considered as an acceptable fit and .060 or less was considered as a good fit (Browne & Cudeck, 1993; Hu & Bentler, 1999). Confirmatory factor analysis was conducted by MPLUS 7.4 program (Muthen & Muthen, 2015).

Differential Item Functioning (DIF) analysis was conducted to evaluate the fairness and equality of tests on item level in investigating the comparability of gender performances. Having an instrument without DIF items is an indication of a well-prepared instrument in terms of group comparisons and fairness. In the study, logistic regression (LR) and Structural Equation Modeling (SEM) DIF methods were used. In the logistic regression procedure, as a first step, only total score (model1), then total score and grouping variable (model2), and finally total score, grouping variable and their interaction (model3) were used as predictors. Significance of country and their interaction, and the change in R^2 value were taken as evidence for uniform bias and non-uniform bias, respectively (Zumbo, 1999). Jodoin and Gierl (2001) proposed ΔR^2 higher than 0.035 indicates moderate DIF and higher than 0.070 indicates large DIF. SPSS 22.0 programs were used to conduct logistic regression analysis. In the SEM procedure, a Confirmatory Factor Analysis (unifactorial, with all items as indicators of the latent variable) is conducted to assess configural and scalar invariance. The difference between incremental types of model fit is evaluated as the factor loadings and intercepts are forced to be equal for comparison groups (van de Vijver, 2017). If the difference in comparative fit index (CFI) and Tucker Lewis index (TLI) between configural and the scalar invariance model is larger than .010 modification indices are investigated to identify DIF items (Cheung & Rensvold, 2002). Mplus 7.4 program was used for SEM DIF detection procedure (Muthen & Muthen, 2015).

Defining performance level descriptors

Determination of exemplars according to the empirical method: First, the exemplar items were determined. In order to determine the potential exemplars according to empirical method using 50%, 67%, and 80% response probability, first, raw scores were converted to zero to hundred grade scale. The scores were clustered into five categories (0 - 20; 21 - 40; 41 - 60; 61 - 80; 81 - 100). The students in each score category was identified and then the proportion of correct response of each item for each score category was calculated using IBM SPSS 22. These proportions could be considered as classical test theory item difficulty indices for each item in each score category. In the present study, three different response probabilities (RP) were used to determine the exemplars: 50% RP: The items answered correctly by at least 50% of the participants in each performance level were selected as exemplar items; 67% RP: The items answered correctly by at least 67% of the participants in each

performance level were selected as exemplar items; 80% RP: The items answered correctly by at least 80% of the participants in each performance level were selected as exemplar items. For example, at the third performance level (41 - 60), the proportion of correct response for item 3 was calculated as 60.2%. This item was not chosen as an exemplar according to the empirical based method using 67% and 80%, while it was selected as an exemplar item according to empirical based method using 50%.

Determination of exemplars according to the model-based method: In the present study, ConstructMap 4.6 (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2010) program was used which gives the total raw score of the students, student ability estimation and item difficulty values on Wright map. The program analyzes 1-0 item scores based on the Rasch model of item response theory. The Wright map shows student ability scores and item difficulty values on the same scale. In addition, raw scores can be reported on this map. Items were given in the order related to their difficulty indices and item clusters were investigated to decide the cut scores for each performance level.

RESULTS

Psychometric Properties and Item Bias Analysis

Internal consistency analysis

The Cronbach's Alpha reliability coefficient value in the proficiency exam reading part calculated as .814 with 19 items. This value indicated a good internal consistency (George & Mallery, 2003). The corrected item-total correlation coefficient of each item was higher than .200 indicated that all items correlated with total score as expected.

Descriptive statistics

Reading test consisted of 19 items that were scored dichotomously. The reading score of students ranged from 0 to 19 (M = 10.06, SD = 4.38). Reading scores were normally distributed, with skewness of 0.15 and kurtosis of -0.86. Students were 391 men and 239 women (men: M = 9.94, SD = 4.23; women: M = 10.24, SD = 4.62). An independent-samples t-test indicated that reading scores of men and women were not significantly different ($t_{(628)} = 0.831$, $p > .05$, $d = 0.07$).

Table 1. Descriptive Statistics of the Reading Test

N	Mean	Standard Deviation	Standard Error of The Mean	Skewness	Kurtosis
630	10.06	4.38	.17	0.15	-0.86

Factor structure

Reading test aimed to measure one dimensional reading ability of students (See Figure 1). Therefore, confirmatory factor analysis was conducted to test whether 19 items reading test was unidimensional as it was proposed (see Table 2). The results showed that RMSEA, CFI and TLI values indicated an acceptable fit of the data to the unidimensional model (RMSEA = .054 < .060; CFI = .918 > .900). Thus, confirmatory factor analysis findings indicated that the proposed model was supported by the collected reading test data.

Table 2. One-dimensional Confirmatory Factor Analysis Results

χ^2/df	RMSEA	CFI	TLI
2.836***	.054	.918	.908

***p < .001.

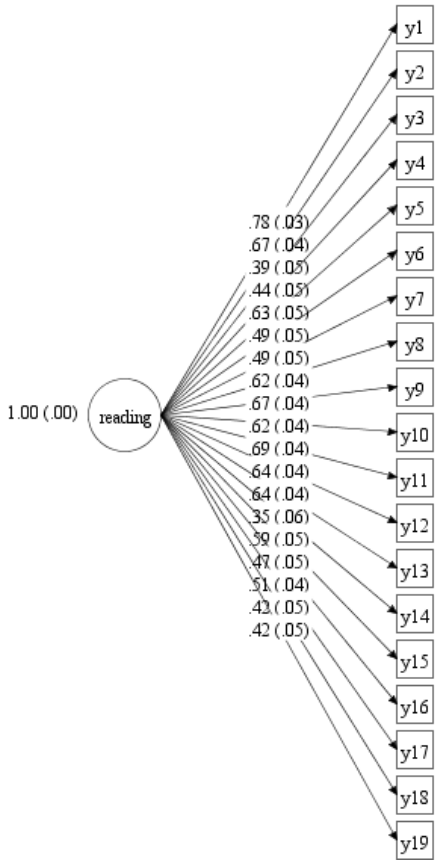

Figure 1. The Proposed Structure of Reading Test

Item bias

In this section, gender related DIF results based on Logistic Regression and Structural Equation Modeling DIF detection methods were presented. DIF results using LR method was presented in Table 3. The results indicated that none of the reading items showed DIF for gender groups. SEM DIF results are presented in Table 4. In comparing answers of girls and boys, none of the reading items showed DIF for gender groups either. Therefore, using two different DIF detection methods, it was concluded that reading test did not contain any DIF items for gender groups which was a fairness indicator of the test.

Table 3. Logistic Regression DIF Results

Item No	**Girls-Boys ΔR^2**
01	.004
02	.007
03	.002
04	.009
05	.006
06	.001
07	.001
08	.005
09	.007
10	.001
11	.001
12	.006
13	.004
14	.001
15	.004
16	.001
17	.003
18	.003
19	.002

Table 4. SEM DIF Results

Model	χ^2/df	RMSEA	CFI	ΔCFI	TLI	ΔTLI	DIF ITEMS
Configural	1.483**	.039	.956		.950		None
Scalar	1.464**	.038	.956	.000	.952	-.002	

**p < .01.

Item parameters according to classical and item response theory

In Table 5, item difficulty and item discrimination indices calculated by classical test theory and item response theory were reported. According to classical test theory item analysis statistics, the difficulty of the items were ranged from .31 to .85 with the mean value of .53; and the discrimination index was ranged from .24 to .53 with the mean value of .39. One parameter item response theory (Rasch model) results produced item difficulty indices ranging from -1.90 to 1.12 with the mean value of 0.00. These values indicated that the reading test had medium level difficulty.

Table 5. Item Parameters According to Classical and Item Response Theory

Item	Item Difficulty Index	Item Discrimination Index	b Parameter
1	.53	.53	0.03
2	.52	.43	0.04
3	.63	.28	-0.48
4	.34	.29	0.99
5	.32	.44	1.12
6	.38	.34	0.77
7	.68	.33	-0.79
8	.60	.44	-0.34
9	.59	.48	-0.30
10	.59	.44	-0.28
11	.52	.50	0.05
12	.60	.45	-0.33
13	.56	.47	-0.13
14	.31	.24	1.18
15	.85	.35	-1.90
16	.47	.34	0.29
17	.54	.37	-0.03
18	.65	.31	-0.59
19	.39	.30	0.72
Total	.53	.39	0.00

Defining Performance Level Descriptors

Identifying exemplar items using empirical method

Using RP 50, RP 67 and RP 80, exemplar items for each score interval (0 - 20; 21 - 40; etc.) were decided (see Table 6). Exemplar item grouping results were affected from chosen response probability. While an item was located to lower score intervals in RP 50, the same item was generally located to higher score intervals in RP 80. For score interval of 0 - 20, none of the items were located. This means that students who got a score between 0 and 20 in reading part could not achieve none of the items on general. In the next section how these item classifications were used to define performance level descriptors was explained. Additionally, the hierarchical structures were observed for RP 50, RP 67 and RP 80. If an item was located in one of the score interval (answered correctly by students in this score interval with required percentage) then the item was achieved by students in above score intervals with required percentage, too.

Table 6. Exemplar Items in Empirical Method

PL	n	RP 50	RP 67	RP 80*
0-20	31	-	-	-
21-40	174	15	15	-
41-60	186	3, 7, 8, 9, 10, 12, 13, 18	7	15
61-80	156	1, 2, 6, 11, 16, 17	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	1, 7, 8, 9, 10, 12
81-100	83	4, 5, 14, 19	4, 5, 6, 14, 16, 19	2, 3, 5, 11, 13, 16, 17, 18, 19

PL: performance level, RP: response probability. * Item 4, 6 and 14 could not be classified to any PL for RP 80.

Performance level descriptors using empirical method

In Table 6, exemplar items were reported with different response probabilities to show how each response probability affected the classification. In order to define performance level descriptors, RP 67 was selected. RP 50 was justified as the number of students at a particular score interval can do a task exceeds the number of students who cannot do the task (Zwick et al., 2001). However, RP 50 is criticized as being too low for a standard. Kolstad et al., (1998) stated that "if one is going to say that people with a particular score on an assessment can successfully perform a particular assessment task, one wants to be fairly sure that a substantial majority of them can do it" (p. 11). RP 80 could be used if the aim of the test requires higher percentage correct values. RP 80 was considered to be too stringent (Kolstad et al., 1998). In this study, three items (Item 4, 6 and 14) could not be located to any score interval for this reason. In RP 67 two third of the students were required to answer the item correctly in related score interval. RP 67 was justified as being consistent with the mastery notion (Kolstad et al., 1998) and maximizing the information of the correct response under several IRT models (Huynh, 2006). Therefore, performance level descriptors were defined using exemplar items under RP 67. The performance level descriptors were defined by three experienced scholars.

Results showed that students in score interval 0 - 20 could not show any reading ability measured in this test. Students in score interval 21 - 40 "can recognize a detail from context by using more frequently used vocabulary item (from k1 band) in the question root as an explicit clue". The ability of students in score interval 41 - 60 could be exemplified as, in addition to previously described ability, "can recognize a detail from context by using frequently vocabulary item (from k1 band) in the question root as an explicit clue". There was a small difference between these two abilities and for these groups only one item was located. For score intervals 61 - 80 and 81 - 100, there were more items. This might indicate that this test could better differentiate between score intervals of 0 - 60, 61 - 80 and 81 - 100 which is reasonable in a sense that a student should get overall score of 60 to be successful. Students in score interval 61 - 80 "can infer a detail by using an explicit clue in the text" whereas students in score interval 81 - 100 "can infer the meaning by using implicit clues in the text with less frequently used vocabulary" in addition to previously described abilities. It is also important to note that these structures are based on a probabilistic view in which a student in a score interval could have these abilities with at least 67% probability.

Cross validation of exemplar items in empirical method

As empirical method is based on percentages calculated according to classical test theory and as classical test theory is affected from different samples, the dataset was divided randomly into two to cross validate the results. In Table 8 and Table 9 these results were reported. In sample 1, for RP 50 and RP 67 only one item was located to different score interval whereas for RP 80, two items were mislocated (0.95, 0.95, and 0.89 convergence ratios, respectively). In sample 2, for RP 50 and RP 67 two items were located to different score interval whereas for RP 80, four items were located differently (0.89, 0.89, and 0.74 convergence ratios, respectively). These results showed that RP 80 was affected from sample change compared to RP 50 and RP 67. This finding also justified not selecting RP 80 for defining performance level descriptors.

Table 7. Performance Level Descriptors in Empirical Method

Level	PL	n	RP 67%	Performance Level Descriptors
1	0-20	31	-	-
2	21-40	174	15	• Can recognize a detail from context by using more frequently used vocabulary item (from k1 band) in the question root as an explicit clue.
3	41-60	186	7	• Can recognize a detail from context by using frequently vocabulary item (from k1 band) in the question root as an explicit clue.
4	61-80	156	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	• Can recognize a detail from context by using more frequently used vocabulary item (from k2 band) in the question root as an explicit clue. • Can follow the development of text structure and decide from where in the text each sentence is removed by using an explicit clue. • Can reach a conclusion by using an implicit clue in the text. • Can infer a detail by using an explicit clue in the text.
5	81-100	83	4, 5, 6, 14, 16, 19	• Can follow the development of text structure and can decide from where in the text each sentence is removed by using an implicit clue. • Can infer the meaning by using explicit clues in the text. • Can infer the meaning by using implicit clues in the text with less frequently used vocabulary. • Can infer writer's attitude and viewpoint.

Table 8. Cross Validation of Exemplar Items in Empirical Method-Sample 1

PL	n	RP 50	RP 67*	RP 80**
0-20	20	-	-	-
21-40	85	15	15	-
41-60	98	3, 7, 8, 9, 10, 12, 13, 18	7	15
61-80	61	1, 2, 6, 11, 16, 17, **19**	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	1, **2**, 7, 8, 9, 10, 12
81-100	47	4, 5, 14	4, 5, 6, 16, 19	3, **4**, 5, 11, 13, 16, 17, 18, 19

PL: performance level, RP: Response probability. * Item 14 could not be classified to any PL for RP 67. ** Item 6 and 14 could not be classified to any PL for RP 80

Table 9. Cross Validation of Exemplar Items in Empirical Method-Sample 2

PL	n	RP 50	RP 67*	RP 80
0-20	11	-	-	-
21-40	89	15	15	-
41-60	88	3, 7, 8, 9, 10, 12, **16,** 18	7, **18**	15
61-80	95	1, 2, 6, 11, **13**, 17	1, 2, 3, 8, 9, 10, 11, 12, 13, 17	1, 7, 8, 9, 10, 12, **18**
81-100	36	4, 5, 14, 19	5, 6, 14, 16, 19	2, 3, 11, 13, **14**, 17

PL: performance level, RP: Response probability. * Item 4 could not be classified to any PL for RP 67. ** Item 4, 5, 6, 16, 19 could not be classified to any PL for RP 80

Identifying exemplar items using model-based method using ConstructMap

ConstructMap 4.6.0 program was used to get Wright Map (See Figure 2). Wright Map provided ability level of students (ranging from -3 to +3), raw score associated with this ability levels, number of students in each ability level (denoted by X's) and item numbers ordered based on difficulty estimation done based on item response theory. The next step is to decide item groups by setting cut points. Among several approaches about how to decide cut points, The Construct Mapping method (Draney & Wilson, 2009) was used to identify the exemplar items. The Construct Mapping method was selected as experts defining performance level description (panelists) were given items' location and related scale scores. Panelists examined the data and items and selected the best locations for cut scores.

In the study, panelists investigated item clusters in the Wright Map and grouped items as given in Table 10. Then the scale scores intervals (theta) were reported for each level with RP67. These scale scores were estimated using the item response theory. Items were investigated in content and cognitive processes and performance level descriptors were provided. The results provide hierarchical structure for cognitive processes.

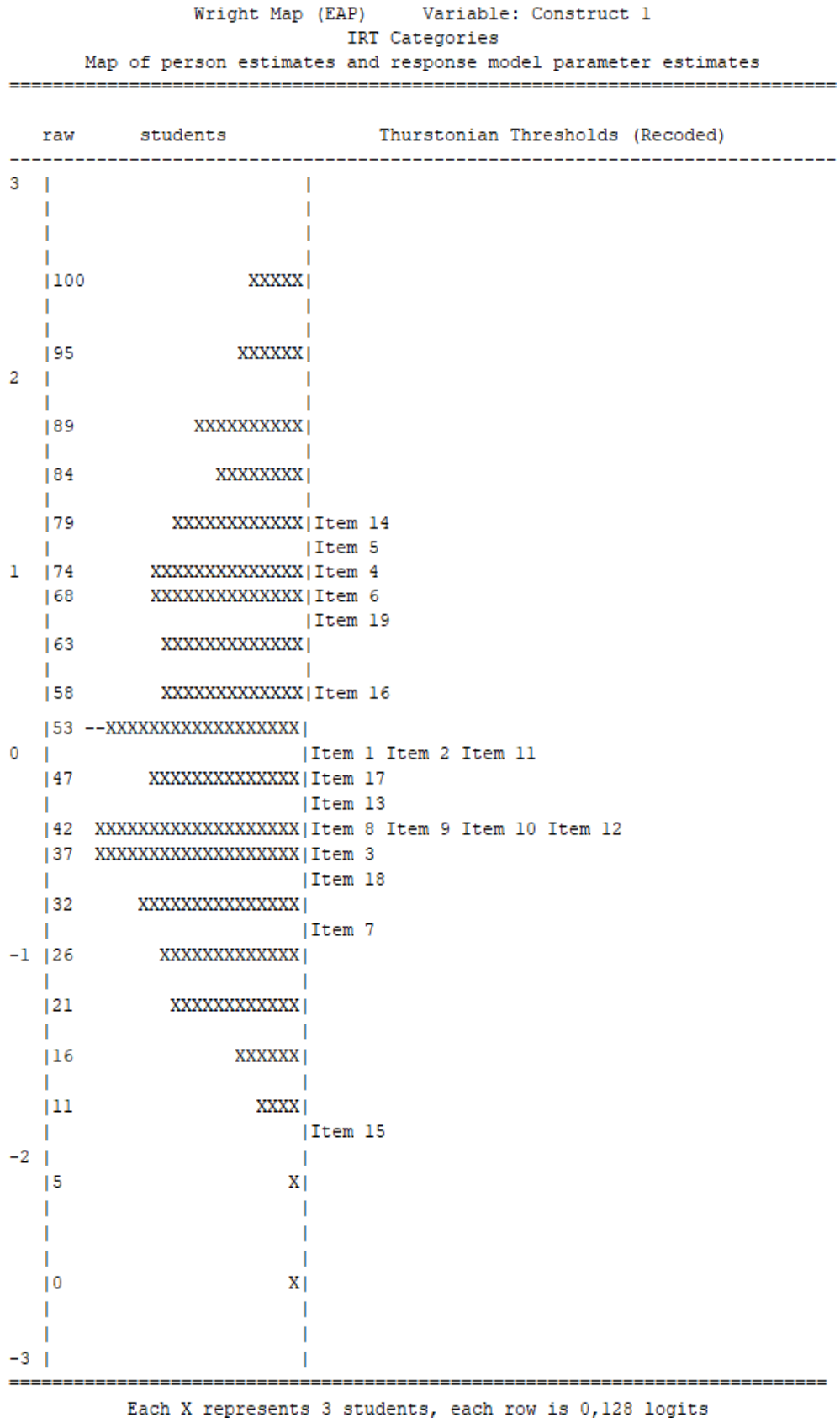

Figure 2. Wright Map Obtained by ConstructMap Program

Table 10. Item Grouping According to Construct Mapping Method

Level	Items	Theta Score Interval RP 67	Performance Level Descriptors
1	15	-0.60 and below	* Can recognize a detail from context by using more frequently used vocabulary item (from k1 band) in the question root as an explicit clue.
2	7	-0.60 and 0.00	* Can recognize a detail from context by using frequently vocabulary item (from k1 band) in the question root as an explicit clue.
3	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	0.00 and 0.90	* Can recognize a detail from context by using more frequently used vocabulary item (from k2 band) in the question root as an explicit clue. * Can follow the development of text structure and decide from where in the text each sentence is removed by using an explicit clue. * Can reach a conclusion by using an implicit clue in the text. * Can infer a detail by using an explicit clue in the text.
4	16	0.90 and 1.25	* Can infer writer's viewpoint.
5	4, 5, 6, 14, 19	1.25 and above	* Can follow the development of text structure and can decide from where in the text each sentence is removed by using an implicit clue. * Can infer the meaning by using explicit clues in the text. * Can infer the meaning by using implicit clues in the text with less frequently used vocabulary. * Can infer writer's attitude.

DISCUSSION and CONCLUSION

This study aimed to raise awareness about the importance of criterion referenced assessment via showing how performance level descriptors in a high-stake exam in Turkey could be defined. Giving too much emphasis on norm referenced assessment by rankings and almost no emphasis on criterion referenced assessment is continuing to harm the educational system from early years of primary school to university education. Especially national large-scale assessments that aim to select limited number of students among huge number of students to a higher educational institution focuses on norm referenced assessment in Turkey. However, there are national assessments, especially language tests, that aims to decide who are proficient or not, but even the results of these assessments are not reported with the criterion referenced perspective. Therefore, criterion referenced assessment is undervalued. There is a need to use criterion referenced assessment via providing performance level descriptors to integrate assessment results to the instructions and to provide concrete feedback to the stakeholders. Performance level descriptors could be used to follow the development of a student throughout the years of assessments. Therefore, a student who started from lower levels could increase his or her performance over years and this development could create a confidence for the student. Only ranking students is harming majority of the students as top rankings are reserved by top achievers.

One of the reasons of why assessment results based on criterion referenced assessment via performance level descriptors is not popular could be that there are very limited examples of performance level descriptors in Turkish context. Defining performance level descriptors requires more detailed effort and know how compared to providing norm referenced assessment results. This study showed how performance level descriptors could be defined using empirical method and model-based method. Empirical method is based on classical test theory and easier to implement and model-based method is based on item response theory and requires expertise on statistical software. In both methods, in the process of defining the descriptors for the score intervals, there is a hierarchical structure among the item clusters, and items that are located in higher score intervals require higher cognitive demands. As it is known, Wright maps were based on the item response theory, in which the item parameters could be estimated independently from the sample. In the study, we obtained similar results for both empirical method and model-based method. In the relevant literature, similar results were obtained in studies in different fields (e.g. mathematics). In the previous literature, it was found that the results obtained from the empirical method and wright maps were similar (e.g. Arıkan & Kilmen, 2018). As both methods produce similar item rankings and item clusters in this study, teachers could use empirical method to define performance level descriptors for their assessments and measurement experts could use model-based methods to get more stable results.

Teacher groups with limited access to the measurement experts could follow the steps described in the empirical method and could get item clusters and then could describe required abilities by the items. The study showed that with 600 students the findings were consistent with the smaller samples. With smaller number of students, the results could be more sample dependent, but the feedbacks based on performance level descriptors would be still useful for this specific group. Teachers could cooperate with other teachers to increase the number of students in their assessments and group discussion on defining performance level descriptors would be beneficial for them. Testing companies with measurement specialist and bigger schools that have measurement department are advised to use model-based method. Item statistics estimated by item response theory are sample independent which makes them more consistent (Hambleton & Jones, 1993). Cooperating with teachers and experts, Construct Mapping method is useful in defining performance level descriptors based on item analysis and item mapping.

Overall, we showed that it is possible to define performance level descriptors for an English proficiency exam. With the help of verbal descriptors for related score intervals, the exam results will be more meaningful and related feedback will be given to students, parents and school administration. Teachers and administration are expected to use this information to raise the quality of education. The student achievement outcome was defined according to what students can do and cannot do, therefore, overall success of given education throughout the year would be evaluated by these standards. When similar assessment is used for incoming proficiency exams, the outcome could also be comparable in terms of these standards. For students who could not achieve this test could be provided what they can do in addition to what they cannot do. These feedbacks are expected to help these students to shape their remedial studies.

The limitation of this study is that the number of reading items was not that high, and the items were generally loaded above score of 60. As a result, for some score intervals, one item was loaded. Defining performance level descriptors based on a limited number of items would threat the reliability of the findings. Therefore, having more items that have more equal distribution over score intervals would be preferable. Piloting items and selecting items according to pilot item analysis could be beneficial when administrating the items beforehand is possible.

REFERENCES

Arıkan, S., & Kilmen, S. (2018). Sınıf içi ölçme ve değerlendirmede puanlara anlam kazandırma: %70 doğru yanıt yöntemi. *İlköğretim Online, 17*(2), 888-908. doi: 10.17051/ilkonline.2018.419337

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191-204. doi: 10.3102/10769986017002191

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 137–162). Newbury Park, CA: Sage.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. doi: 10.1207/S15328007SEM0902_5

Demirtaşlı, N. (2009). Eğitimde niteliği sağlamak: Ölçme ve değerlendirme sistemi örneği olarak CİTO Türkiye öğrenci izleme sistemi (ÖİS). *Cito Eğitim: Kuram ve Uygulama, 3*, 25-38.

Draney, K., & Wilson, M. (2009). Selecting cut scores with a composite of item types: The Construct Mapping procedure. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 276–293). Maple Grove, MN: JAM Press.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. London: Lawrence Elbaum Associates, Publishers.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220. doi: 10.1207/s15324818ame1702_3

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). Fundamentals of item response theory. Newbury Park CA: Sage.

Hu, L.-T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi: 10.1080/10705519909540118

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20. doi: 10.1111/j.1745-3992.2006.00053.x

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349. doi: 10.1207/S15324818AME1404_2

Karantonis, A. (2017). *Using exemplar items to define performance categories: A comparison of item mapping methods* (Unpublished doctoral dissertation, University of Massachusetts). Retrieved from https://scholarworks.umass.edu/dissertations_2/1101/

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard- setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12. doi: 10.1111/j.1745-3992.2006.00047.x

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2010). ConstructMap 4.6. [Computer software]. Berkeley, CA: BEAR Center.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: American Institutes for Research.

Muthen, B. O., & Muthen, L. K. (2015). Mplus (Version 7.4) [Computer software]. Los Angeles, CA: Muthen & Muthen.

Shulman, L. S. (2009). Assessment of teaching or assessment for teaching? Reflections on the invitational conference. In G. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality*. Thousand Oaks, CA: Sage Publications.

Ullman, J. B. (2001). Structural equation modeling. In B. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed.), (pp. 653-771). Boston, MA: Allyn & Bacon.

Van de Vijver, F. J. R. (2017). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (2nd, rev. ed.). New York, NY: Routledge.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15-25. doi: 10.1111/j.1745-3992.2001.tb00059.x

Ampirik ve Modele Dayalı Yeterlik Tanımları: İngilizce Yeterlik Sınavı Örneği

Giriş

Türkiye'de, test sonucunu daha verimli bir şekilde rapor etmek için yeterlik puan raporlama yöntemlerinin geliştirilmesine ihtiyaç duyulmaktadır. Bir testten alınabilecek puan aralıklarında tanımlanan yeterlikler sınav sonuçlarının anlamlı hale gelmesini sağlamakta ve paydaşlara gerekli geribildirimler verme konusunda yararlı olmaktadır. Bu çalışma, ampirik yöntem ve modele dayalı yöntem (ConstructMaps) kullanılarak, İngilizce yeterlilik testine ait puan aralıklarının nasıl tanımlanabileceğini göstermeyi amaçlamıştır.

Ampirik yönteme (Zwick, Senturk, Wang, & Loomis, 2001) göre yeterlik tanımlamanın ilk aşamasında, önce ölçeğe ilişkin puan aralıkları belirlenir. Ardından, bu puan aralıklarında yer alan öğrenci grupları saptanır. Her bir puan aralığındaki öğrencilerin çoğunluğu tarafından doğru olarak cevaplandırılan (örneğin %50, %65, %67 ve %80) maddeler belirlenir (Beaton & Allen, 1992). Araştırmacı belli bir doğru yanıtlama olasılığı belirleyerek bu olasılık üzerinden her bir puan

aralığındaki maddeleri belirler. Örneğin, doğru yanıt olasılığı %65 olarak belirlenmişse, her bir puan aralığında bireylerin %65'i tarafından doğru şekilde yanıtlanan maddeler bulunur. Her bir puan aralığı için, bu maddelerle ölçülen bilişsel ve içerikle ilgili özellikler belirlenir ve her bir puan aralığı için performans tanımlanır.

Modele dayalı yöntemde, ampirik yöntemde olduğu gibi, maddenin doğru yanıtlanma olasılığı esas alınarak maddeler belirlenir. Modele dayalı yöntemin ampirik yöntemden farkı, Madde Tepki Kuramı Rasch modeline göre doğru cevap olasılıklarının tahmin edilmesidir. Madde tepki kuramına göre, yetenek ve madde parametreleri aynı ölçekte yerleştirilebilir (Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991). Yukarıda belirtildiği gibi, bu iki yöntem arasındaki fark, yanıt olasılıklarının hesaplanma şeklidir. Ampirik yöntemde, yanıtlanma olasılığı klasik test teorisine göre hesaplanırken modele dayalı yöntemde madde tepki kuramına göre hesaplanır.

Yöntem

Çalışma grubu

Türkiye'nin güneybatısındaki bir devlet üniversitesinin 2016-2017 İngilizce yeterlilik sınavına giren 630 lisans öğrencisi bu araştırmanın çalışma grubunu oluşturmaktadır. Öğrencilerin %68'i erkek, %32'si ise kadındır.

Veri toplama aracı

Bu çalışmada, üniversitenin yabancı dil okulu test geliştirme ekibi tarafından geliştirilen İngilizce yeterlilik testi kullanmıştır. İngilizce yeterlilik sınavının dört ana boyutu bulunmaktadır: Okuma, Dinleme, Yazma ve Konuşma. Bu çalışma, bu testin bir kısmını oluşturan okumaya odaklanmaktadır. Okuma bölümü okuma paragraflarını içermektedir. Çeşitli madde formatlarında (eşleştirme, kısa cevap ve çoktan seçmeli) 19 test maddesinden oluşmaktadır.

Verilerin analizi

Ön analiz olarak, okuma testinin iç tutarlılığı Cronbach'ın Alfa güvenilirlik katsayısı kullanılarak hesaplanmıştır. Okuma testi, okuduğunu anlama yeteneğini ölçmek için geliştirilmiştir. Bu nedenle, okuma testinin tek boyutluluğunu test etmek için doğrulayıcı faktör analizi yapılmıştır. Doğrulayıcı faktör analizi MPLUS 7.4 programı ile gerçekleştirilmiştir (Muthen & Muthen, 2015).

Maddelerin bir gruba yanlı olup olmadığını test etmek için madde yanlılığı analizi yapılmıştır. Bu çalışmada, lojistik Regresyon (LR) ve Yapısal Eşitlik Modelleme (YEM) madde yanlılığı yöntemleri kullanılmıştır.

Bu analizlerin ardından yeterlik tanımlama işlemleri yapılmıştır. Bu araştırmada yeterliklerin tanımlanmasında ampirik ve modele dayalı yöntemler kullanılmıştır. Ampirik yöntemde öğrencilerin almış oldukları puanlar beş performans seviyesine ayrılmıştır (0 - 20; 21 - 40; 41 - 60; 61 - 80; 81 - 100). Her bir puan kategorisindeki öğrenciler belirlenmiş ve daha sonra her bir puan kategorisi için her bir maddenin doğru cevaplanma oranı hesaplanmıştır. Bu çalışmada, her bir puan kategorisini temsil eden madde örneklerini belirlemek için üç farklı cevap olasılığı (RP) kullanılmıştır. %50 RP: Her bir performans seviyesinde katılımcıların en az %50'si tarafından doğru olarak cevaplanan maddeler örnek maddeler olarak seçilmiştir. %67 RP: Her bir performans seviyesinde katılımcıların en az %67'si tarafından doğru olarak cevaplanan maddeler örnek maddeler olarak seçilmiştir. %80 RP: Her performans seviyesinde katılımcıların en az %80'i tarafından doğru bir şekilde cevaplanan maddeler örnek maddeler olarak seçilmiştir. Modele dayalı yeterlik tanımlamaları ise Wright haritası üzerinde öğrencilerin toplam ham puanını, öğrenci yetenek tahminlerini ve madde güçlük indekslerini veren ConstructMap 4.6 (Kennedy, Wilson, Draney, Tutunciyan & Vorp, 2010) programı kullanılarak

yapılmıştır. Program Madde Tepki Kuramının Rasch modeline dayanarak 1-0 şeklinde puanlanan maddeleri analiz etmektedir. Wright haritası, öğrenci ölçeği puanlarını ve madde güçlük indekslerini aynı ölçekte göstermektedir.

Bulgular

Yeterlik sınavı okuma testinin Cronbach Alfa güvenirlik katsayısı .81 olarak hesaplanmıştır. Bu değer ölçekten güvenilir sonuçlar elde edildiğinin bir kanıtıdır (George & Mallery, 2003). Okuma testi öğrencilerin tek boyutlu okuduğunu anlama becerilerini ölçmeyi amaçlamıştır. Bu nedenle 19 maddelik okuma testinin tek boyutlu olup olmadığını test etmek için doğrulayıcı faktör analizi yapılmıştır. Yapılan analiz sonucunda ölçeğin tek boyutlu bir yapıda olduğu saptanmıştır (RMSEA = .054 < .060; CFI = .918 > .900). Lojistik Regresyon ve Yapısal Eşitlik Modelleme madde yanlılığı belirleme yöntemlerine dayalı analizler sonucunda okuma maddelerinin hiçbirinin cinsiyet grupları için madde yanlılığı göstermediği saptanmıştır.

Ampirik yönteme göre bulgular incelendiğinde, 0 - 20 puan aralığında öğrencilerin okuduğunu anlama becerisinin tanımlanamadığı saptanmıştır. 21 - 40 puan aralığındaki öğrencilerin soru kökündeki açık bir ipucu olarak daha sık kullanılan kelime hazinesini (k1 bandından) kullanarak içerikten bir detay tanıyabildiği belirlenmiştir. 41 - 60 puan aralığında bir puan alan öğrencilerin ise soru kökündeki açık bir ipucu olarak sık başvurulan kelime hazinesini (k1 bandından) kullanarak içerikten bir ayrıntıyı tanıyabildiği saptanmıştır. 61 - 80 puan arası bir puana sahip öğrencilerin soru kökündeki açık bir ipucu olarak daha sık kullanılan kelime hazinesini (k2 bandından) kullanarak içerikten bir detay tanıyabildiği, metin yapısının gelişimini takip edebildiği metinde açık bir ipucu kullanarak bir sonuca ve detaylara ulaşabildiği görülmüştür. En üst yeterlik düzeyi olan 81 - 100 puan arasında puan alan öğrencilerin ise metin yapısının gelişimini takip edebildiği ve bir ipucu kullanarak her cümledeki metnin nereden çıkacağına karar verilebildiği, daha az kullanılan kelime dağarcığı içeren metinde örtük ipuçlarını kullanarak anlam çıkarabildiği ve yazarın tutum ve bakış açısını yakalayabildiği saptanmıştır.

Modele dayalı bulgulara göre en alt yeterlik basamağının kesim noktası olarak -0.60 puan belirlenmiş, bu puanın altında bir puana sahip öğrenciler için yeterlik tanımları yapılabilmiştir. Ancak yapılan tanımlamalar ampirik yöntemdeki 21 - 40 puan aralığında tanımlanan yeterliklerdir. Diğer bir deyişle, ampirik yöntemde 21 - 40 puan arasında tanımlanan yeterlikler modele dayalı yöntemde en alt yeterlik basamağında tanımlanmıştır. Benzer şekilde ampirik yöntemde 41 - 60 puan aralığında belirlenen yeterlik tanımları da modele dayalı yöntemde -0.60 - 0.00 puan aralığında tanımlanmıştır. 0.00 - 0.90 arasında puan alan öğrencilerin ise soru kökündeki açık bir ipucu olarak daha sık kullanılan kelime haznesini (k2 bandından) kullanarak içerikten bir detay tanıyabildiği, metin yapısının gelişimini takip edebildiği metinde açık bir ipucu kullanarak bir sonuca ve detaylara ulaşabildiği görülmüştür. Bu yeterlik tanımı ampirik yöntemde 61 - 80 puan aralığına denk gelmektedir. 0.90 - 1.25 arasında puan alan öğrencilerin yazarın bakış açısı hakkında çıkarım yapabildiği saptanmıştır. 1.25 puan üzerinde puan alan öğrencilerin metin yapısının gelişimini takip edebildiği ve bir ipucu kullanarak her cümledeki metnin nereden çıkacağına karar verilebildiği, daha az kullanılan kelime dağarcığı içeren metinde örtük ipuçlarını kullanarak anlam çıkarabildiği ve yazarın tutumunu belirleyebildiği görülmüştür.

Sonuç ve Tartışma

Genel olarak değerlendirildiğinde, ampirik yöntem ve modele dayalı yöntem arasında yeterlik tanım basamakları açısından birtakım farklılıklar gözlense de sonuçlar yeterlik tanımlarının hiyerarşik bir şekilde sıralandığını, İngilizce yeterlilik sınavının yeterlik tanımlarının ampirik ve modele dayalı yöntemlerle tanımlanabileceğini göstermektedir. Ampirik yöntem, klasik test teorisine dayanır ve uygulanması kolaydır. Modele dayalı yöntem, madde tepki kuramına dayanır ve istatistiksel yazılım üzerinde uzmanlık gerektirir. Her iki yöntemde de puan aralıkları için tanımlayıcıların tanımlanması sürecinde, madde kümeleri arasında hiyerarşik bir yapı bulunmuş ve daha yüksek puan aralıklarında

bulunan maddeler daha yüksek bilişsel beceriler gerektirmiştir. İlgili literatürde, farklı alanlarda (örneğin matematik) yapılan çalışmalarda benzer sonuçlar elde edilmiş, literatürde, ampirik yöntem ve Wright haritalarından elde edilen sonuçların benzer olduğu bulunmuştur (Arıkan & Kilmen, 2018). Her iki yöntemde de benzer madde sıralamaları ve madde kümeleri oluşturulduğundan, öğretmenler, değerlendirmeler için performans düzeyi tanımlayıcılarını tanımlamada ampirik yöntem kullanabilir, ampirik yöntemde açıklanan adımları takip edebilir ve yeterlikleri tanımlayabilirler. Öğretmenler, diğer öğretmenlerle birlikte, öğrencilerin başarısını arttırmak için iş birliği yapabilir ve performans düzeyi tanımlayıcılarını tanımlamak için bir araya gelebilirler. Ölçme ve değerlendirme alanında uzmanlaşmış kişilerin ise modele dayalı yöntem kullanmaları tavsiye edilebilir. Çünkü madde tepki kuramı ile tahmin edilen madde istatistikleri, örneklemden bağımsızdır ve bu da parametreleri daha tutarlı hale getirir (Hambleton & Jones, 1993).

Türkiye'de geniş ölçekli testlerde bağıl ve mutlak değerlendirmeler yapılmasına rağmen daha çok bağıl değerlendirmeye vurgu yapılmaktadır. Özellikle çok sayıda öğrenci arasından sınırlı sayıda öğrenciyi yükseköğretim kurumlarına seçmeyi amaçlayan ulusal geniş ölçekli değerlendirmeler normlara odaklanmaktadır. Bununla birlikte, ulusal çapta düzenlenen mutlak değerlendirmenin kullanıldığı sınavlardan özellikle dil sınavları kimin yetkin olup olmadığına karar vermeyi amaçlamasına rağmen, kişinin yeterliklerine odaklanan bir rapor sunmamakta, sonuçlar puan ile sınırlı kalmaktadır. Oysa değerlendirme sonuçlarının puan ile sınırlı kalmayarak öğrencilere ve paydaşlara somut bir geri bildirim sağlamak için kullanılması daha yararlı olacaktır. Ayrıca, yeterlik tanımları, yıl boyunca bir öğrencinin gelişimini takip etmek için kullanılabilir. Örneğin, düşük seviyelerden başlayan bir öğrenci, yıl boyunca kendi performansını artıracak çalışmaları yeterlik göstergelerinin inceleyerek bulabilir ve kendi gelişimini başarabildiklerine ve başaramadıklarına odaklanarak kendi kendine hızlandırabilir.

Bu çalışmanın çeşitli sınırlılıkları bulunmaktadır. Sınırlı sayıda öğrenciyle elde edilen bulgular sonuçların genellenebilirliğini azaltmaktadır. Bu nedenle daha büyük örneklemlerde benzer araştırmalar yapılabilir. Okuma maddelerinin sayısının çok yüksek olmaması bazı puan aralıklarına sadece bir maddenin yerleşmesine neden olmuştur. Sınırlı sayıda maddeye dayanarak yeterliklerin tanımlanması bulguların güvenilirliğini tehdit etmektedir. Bu nedenle, daha fazla madde içeren testlerle benzer araştırmalar yapılabilir.