ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 430-450

Silent Predictors of Test Disengagement in PIAAC 2012

Münevver İLGÜN DİBEK *

Abstract
Although the effects of test disengagement on the validity of the scores obtained from the data set have been examined in many studies, the predictors of the disengaged behaviors received relatively limited scholarly attention in low-stakes assessment, in particular, in international comparison studies. As such, the present study with a twofold purpose sets out to determine the best fitted explanatory item response theory model and examine the predictors of test disengagement. The data were collected by using items measuring literacy and numeracy skills of adults from different countries such as Norway, Austria, Ireland, France, Denmark, Germany, and Finland participated in PIAAC 2012. The results of the model with item and person characteristics demonstrated that adults tended to be disengaged on very difficult items. Similarly, age has a negative effect on test-taking engagement for adults in several countries such as France and Ireland, while several predictors such as educational attainment, readiness to learn, and the use of ICT skills at home and work had positive effects on test engagement. In addition, females exhibit a higher level of engagement in Norway. Overall, the findings suggested that the effect of the predictors on disengagement depended on the domain and country. So, this study brings further attention that the role of test disengagement should be a prerequisite practice before reaching a conclusion from international large-stake assessments.

Key Words: Explanatory item response theory model, low-stakes assessment, PIAAC, test disengagement.

INTRODUCTION

Examinees are not always motivated to put their full effort into responding to test items, especially in low-stakes settings, such as the Programme for the International Assessment of Adult Competencies (PIAAC). (e.g., Finn, 2015; Wise & DeMars, 2010). The reason why low test motivation is often seen in low-stakes assessments can be revealed by expectancy-value models (e.g., Eccles & Wigfield, 2002). More specifically, as indicated by these models, achievement motivation is closely affected by factors, such as expectancy and value. The former factor is defined as the individual's expectation of achievement in responding to the test items and will be low if the item is too difficult relative to the ability of the individual. In the most general sense, the latter factor is related to the perceived importance and usefulness of the test. However, there is not a straightforward explanation since there are different aspects of value components, such as attainment value, intrinsic value, utility value, and perceived costs (Eccles & Wigfield, 2002). Both the combination of them and each aspect separately is considered to be low in low-stakes assessments. This is because, although there is a need to make a sufficient effort to respond to the test items correctly, the intrinsic motivation of some of the respondents is low, and the results obtained from the test are not vital for the respondents. Therefore, this results in a contradiction. There will be serious problems when the lower levels of motivation of individuals give rise to a low test effort (Wise & DeMars, 2010). These invalid responses cause construct-irrelevant variance and distortion of psychometric features (e.g., Rios, Guo, Mao, & Liu, 2017), leading to the misinterpretation of the results obtained from the data set (Nagy, Nagengast, Becker, Rose, & Frey, 2018). To put it in different words, the true scores of the individuals are contaminated by a systematic source of error due to their level of engagement in the test (Braun, Kirsch, Yamamoto, Park, & Eagan, 2011). In addition, disengagement gives rise to (a) inflated item difficulties, as well as deflated item discriminations (e.g., van Barnevald, 2007), (b) biased item and test information estimates (e.g., van

* Ass. Prof. Dr., TED University, Faculty of Education, Ankara-Turkey, munevver.ilgun@tedu.edu.tr, ORCID ID:0000-0002-7098-0118

To cite this article:
İlgün-Dibek, M. (2020). Silent predictors of test disengagement in PIAAC 2012. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 430-450. doi: 10.21031/epod.796626.

Received: 17.09.2020
Accepted: 13.12.2020

Barnevald, 2007), (c) inflation of reliability estimates based on classical test theory (CTT) (e.g., Wise & DeMars, 2009), (d) erroneous flagging of differential item functioning (e.g., Wise & DeMars, 2010), and (e) decreased correlations with external variables (e.g., Wise, 2009).

Although test disengagement has been characterized in different ways in the literature, rapid guessing is the most widely used and validated one (Wise, 2015). According to Schnipke and Scrams (2002), rapid guessing behavior is the fast response of the test takers to the test items in a way that does not allow them to understand the content of the item. To determine whether this method is being implemented, Schnipke and Scrams (2002) proposed that respondents are divided into two groups according to their solution behavior or rapid guessing behavior. In this approach, the focus is on time elapsed between presenting the item to the respondents and the respondent's response to the item. If the test-taker responds in a period below a certain response time threshold, it means that s/he is displaying rapid guessing behavior. The main challenge in this situation is determining which responses to items are rapid guessing and which responses are solution behaviors.

Alternative Methods for Measuring Test Disengagement

Whatever the reason for the occurrence of disengaged behavior, measuring this behavior accurately and efficiently is crucial given the sizable validity problems that occur due to test disengagement. Test disengagement is determined by computing item response time thresholds that differentiate engaged and disengaged responses. To determine test disengagement, constant threshold and item-specific thresholds are proposed in the literature. For example, as a constant threshold, the frequently used method is the three-second rule (Kong, Wise & Bhola, 2007; Lee & Jia, 2014). The amount of time required to answer the item may vary from item to item (Lee & Jia, 2014). As an example, while respondents can answer an easy item that measures numerical skills faster, they can answer an item that includes long texts with a high reading load and measures verbal skills in a longer time. Thus, researchers tend to use item-specific thresholds, with one of the earliest and most basic approach being the visual inspection method (DeMars, 2007; Wise & Kong, 2005). For each item, the notion is to define the threshold as the judged endpoint of the short time spike in a bimodal response time distribution. In this process, the distributions of the response time of test-takers responding rapidly and those responding more slowly are presented. Although the visual inspection method has various advantages, such as easy interpretation and being evidence-based, there are disadvantages; e.g., being subjective, time-consuming, and not applicable in cases where there is no bi-model distribution (Lee & Jia, 2014; Rios et al., 2017).

Another method for determining item-specific thresholds was the one used by Lee and Jia (2014) on items in multiple-choice format. For each item, the proportion correct conditional on the response time is determined. The response time threshold is defined as proportion correct greater than the chance level for obtaining a correct answer. Since the items included in the PIAAC assessment vary in difficulty and complexity, the amount of time required to give the correct answer will differ. Therefore, considering the advantages of item-specific response time thresholds shown in previous research (e.g., Wise, 2006), the current study adopted this approach.

While much is known about the impact of disengagement on observed test scores, little is known about the impact of an item and personal characteristics on the disengagement of individuals. Some individuals consistently exhibit more disengaged behaviors than others. Determining the person and item as a source of variation could be used for examining individual differences.

Relationship Between Test Disengagement and Person- and Item-Level Variables

Considering the effects of test disengagement on the observed scores of individuals, the reasons for individuals' disengagement have become the focus of attention. Differences between individuals in terms of test disengagement show that it is crucial to take the person as a source of variation in disengagement (Wise, 2009). Therefore, examining the role of person-level variables on test disengagement is beneficial in terms of explaining these differences. To evaluate this situation in terms of large-scale applications, the results of these applications are not of vital importance for individuals

(Asseburg & Frey, 2013; Sundre & Kitsantas 2004; Wise, 2009). Therefore, according to the expectancy-value theory, individuals will attribute the same value to the areas measured in these practices. Consequently, there will be no individual differences in terms of test engagement. However, individuals' perceived expectations about their ability to answer items correctly change from one person to another, depending on the several characteristics that they have. In this regard, gender can affect their perception of the capability, and thus their engagement. Several studies in the literature indicate that males exhibit disengaged behaviors more frequently than females (e.g., DeMars, Bashkov & Socha 2013). Females tend to spend more time answering the items (Setzer, Wise, van den Heuvel, & Ling, 2013).

Although the education level and age of individuals may have a significant effect on the time they spend responding to an item in the test, it has been observed that the literature does not focus on this issue sufficiently. The investigation of this effect would help shed light on solving some unanswered questions in education. For example, highly educated individuals are committed to achieving several tasks and thus have sufficient competency (Organisation for Economic Co-operation and Development [OECD], 2016b); therefore, they may spend more time responding to an item. In addition, older adults may have the necessary knowledge and skills and tend to respond faster to items due to biological factors, such as fatigue and boredom so that they can complete the assessment as soon as possible (Xie, 2003).

Individuals’ readiness to learn has an effect on their disengagement levels. It is closely related to whether adults have sufficient motivation, cognitive skills, and learning strategies to learn a task, feel curious about it, are interested in learning, look for associations among ideas, and believe that they can cope with a problem that they face (Smith, Rose, Smith & Ross-Gordon, 2015). Although the extent to which individuals have the characteristic to be measured by that test plays an important role in responding to a test item, in some cases, various factors also have a critical effect on responding behaviors. When these factors are not taken into account, invalid interpretations can be obtained by only looking at test scores (Nagy et al., 2018), At this point, considering that the test items in the PIAAC are given in a computer environment regardless of which domain measurement, the familiarity of the individuals with various technological elements such as computers and the internet will also have an effect on the individuals' behavior of responding to the test items as if they were insidious, silent factors. In other words, as a source of variation in the engagement levels of respondents, familiarity with information and communications technology (ICT) can also affect respondents’ engagement. The frequent use of the ICT skills of individuals makes them familiar with computers, which increases the motivation, concentration and achievement of individuals in computer-based assessments (Mastuti & Handoyo, 2017). In addition, the extent to which the individuals use various skills at home and work can have an effect on how much effort they applied when responding to tests.

In the literature, it has been stated that several item-level variables have an impact on individuals’ disengagement levels. According to the expectancy-value theory, if the individuals perceive an item as difficult by taking into consideration their competence, their engagement in the testing situation will be negatively affected. Some studies revealed that individuals put more effort into items which had moderate difficulty relative to their ability (Asseburg & Frey, 2013).

In conclusion, the importance of addressing these variables can be explained by analogy with the area above and below an iceberg. While there is only a small part of the total mass above the iceberg, there is a large part of it below, and this controls all the movements of the iceberg. At this point, the same logic can be used to explain the disengagement behaviors of individuals. In other words, in this study, these variables that make up the area under the disengagement as an iceberg will play an important role in explaining the disengagement behavior of individuals. To narrow the focus even more, when the effect of these person and item-level variables on disengagement is ignored, the difference in test scores due to disengagement could not be determined correctly (Braun et al., 2011). Thus, investigation of to what extent these variables explain the disengagement behavior is crucial.

It seems, however, that there has been extensive research on the topic of test-taking effort. Many of these endeavors possess several limitations: focusing on relatively homogenous populations based in a single country (Goldhammer, Martens & Lüdtke, 2017). To date, there have been very few studies that have examined potential differences in test-taking effort between countries in international assessments

(Rios & Guo, 2020), although their personal characteristics largely differ by culture/country (Brown & Harris, 2016). Also, regardless of the number of response categories, studies using traditional IRT models provide information on various individual or item-related characteristics such as respondents' abilities, cognitive levels, achievements, or difficulty and discrimination. Still, they are insufficient to identify systematic effects resulting from the design of the measurement process. In other words, they do not reveal common variability across items or individuals depending on the design of the measurement process or measurement tool. However, this information is very important in determining construct-irrelevant variance originating from various reasons such as cognitive, cultural, and biological factors (AERA, APA, NCME, 2014). Since data were collected in a nested design in the PIAAC study, analyses were done using explanatory item response theory models (EIRT), which allow to include several item and person characteristics as first-level and second-level units, respectively. Thus, this study begins to close this gap in the literature taking a closer look at the predictors of the test disengagement of adults from different countries. Examination of predictors provides the opportunity to obtain more detailed and appropriate results about the factors behind the disengagement of examines.

Purpose of the Study

The aim of this study was to examine the role of several item- and person-level variables on engaged responses in the domains of literacy and numeracy assessed in PIAAC 2012. Investigation of examines' responses on these domains is crucial since, in the most basic sense, the skills regarding numeracy and literacy contribute to the development of various high-level thinking skills, such as analytical thinking, understanding the information in a particular field. In particular, numeracy means more in everyday life than the mathematics we learn at school. In addition, the skills in these areas are used in many areas, from real life to education, business life, and communication with authorized persons (OECD, 2013c). Thus, in order to investigate examines' responses in terms of their engagement in tests requiring numeracy and literacy skills, the answers to the two related research questions were sought:

1. Which of the explanatory item response theory (EIRT) models (baseline model, a model with person characteristics, a model with the item characteristic, and a model with all person and item characteristics and the interaction between them) is best fitted to the PIAAC 2012 subdata?

2. To what extent does the engagement of adults in responding to items included in PIAAC 2012 be explained by person and item characteristics?

METHOD

Sample and Population

The target population of this study included all non-institutionalized adults between age 16 and 65 residing in the country at the time of data collection and participated in Round 1 of PIAAC 2012. In this study, the reason for the selection of countries participating in Round 1 is the high number of countries participated in this round and to increase the representation and generalizability of the results. Another reason for choosing Round 1 is that the t-disengagement rates of the countries participating in only this round are clearly examined in relation to each other in the official report (OECD, 2019), which ensures that the selection of data sets is based on evidence.

In PIAAC, probability sampling was used (OECD, 2013b). In the present study, countries were selected according to their rates of t-disengagement, which represents situations where a respondent spends less time than specified as an item-specific threshold (OECD, 2019). Therefore, in the term "t-disengagement", "t" stands for threshold. More precisely, the percentage of individuals showing t-disengagement in countries participating PIAAC 2012 varies between 8.4% and 33.4%. In the grouping of countries, the percentage of individuals showing t-disengagement in a country is compared to the average percentage of individuals showing t-disengagement in all countries participating in PIAAC 2012. For example, if the percentage of the individuals showing t-disengagement in a country is above the average percentage of t-disengagement, which is 15.70%, this country is classified as the country with a high percentage of t-disengagement. Accordingly, in addition to two countries such as France (21.50%) and Ireland (20.40%) with the highest percentage of individuals with t-disengagement, two countries such as Denmark (14.50%) and Germany (12.30%) where the percentage of individuals with t-disengagement is close to the average were selected. Also, three countries with the least percentage of individuals showing t-disengagement were selected (OECD, 2017) to represent better the pattern observed in the countries that participated in PIAAC 2012. From these examines, the ones who took the computer-based assessment of PIAAC 2012 were included as participants of this study. As a result, the sample of the current study includes 29959 adults from seven countries in total. Specifically, the frequency of these participants by the variables of the interest and countries were presented in Table 1.

Table 1. Frequency of Adults by Variable and Country

	Austria	*Denmark*	*Germany*	*France*	*Ireland*	*Norway*	*Finland*
Variables	*n(3830)*	*n (6048)*	*n(4510)*	*n(2758)*	*n (4058)*	*n(4292)*	*n(4463)*
Gender							
Male	1932	2942	2271	1372	1831	2942	2230
Female	1898	3106	2239	1386	2227	3106	2233
Highest level of schooling							
Less than high school	-	888	695	193	534	888	632
High school	-	2407	1899	1063	847	2407	1659
Above high school	-	2668	1876	1484	2656	2668	2156
Not definable	-	85	40	18	21	85	16
CBA Core score for stage 2							
3	51	104	75	64	75	104	65
4	249	378	296	216	283	378	246
5	927	1422	1120	817	1042	1422	1040
6	2603	4144	3019	1661	2658	4144	3112
Age in 10-year bands							
24 or less	825	965	1023	258	657	965	849
25-34	822	851	921	690	1113	851	995
35-44	899	1170	943	784	1217	1170	874
45-54	832	1182	1026	696	639	1182	908
55 plus	452	1880	597	330	432	1880	837
Index of readiness to learn							
All zero response	1	5	1		1	5	3
Lowest to 20%	463	374	563	135	428	374	167
More than 20% to 40%	840	1052	1232	437	790	1052	545
More than 40% to 60%	841	1348	1107	726	857	1348	967
More than 60% to 80%	829	1542	869	790	954	1542	1381
More than 80%	856	1727	738	670	1028	1727	1400
Index of use of ICT skills at work							
All zero response	153	188	200	162	113	188	188
Lowest to 20%	450	668	487	460	362	668	668
More than 20% to 40%	515	893	571	535	443	893	893
More than 40% to 60%	581	923	690	638	437	923	923
More than 60% to 80%	555	796	639	586	477	796	796
More than 80%	404	912	366	377	582	912	912
Valid skip	1172	1668	1557		1644	1668	1668
Index of use of ICT skills at home							
All zero response	19	10	19	3	10	4	117
Lowest to 20%	581	479	566	257	579	337	579
More than 20% to 40%	708	879	762	689	820	711	830
More than 40% to 60%	826	1290	1002	708	788	1045	708
More than 60% to 80%	848	1529	1101	630	766	1186	592
More than 80%	710	1742	893	471	763	938	428
Valid skip	138	119	167	-	332	71	1209

Data Collection Instruments

In PIAAC 2012, whether the surveys to be used as data collection tools will be applied in the computer environment or in the form of paper and pencil is determined according to the success of the respondents in two tests that measure their ICT skills. If the respondents fail to reach a certain level in the first stage, they will be redirected to the paper-based core section. Furthermore, if the respondents who were successful in the first task fail the subsequent short test, they only participate in the paper-based assessment. To participate in the computer-based assessments, the respondents must pass both tests.

The data collection instrument of the present study contained the literacy and numeracy surveys administered in the computer-based assessment of PIAAC 2012 (Round 1). Fifty-eight items were included in the literacy survey assessing adults' ability to read digital texts, as well as traditional print-based texts. Additionally, 56 items were included in the numeracy survey assessing the adults' ability to use, apply, interpret, and communicate mathematical information. For each domain, the distribution of items by context was presented in Table 2 (OECD, 2016a).

Table 2. Distribution of Items by Context

Survey	Context	Number	%
Literacy	Work	10	17
	Personal	29	50
	Community	13	23
	Education	6	10
	Total	58	100
Numeracy	Everyday life	25	45
	Work-related	13	23
	Society and community	14	25
	Further learning	4	7
	Total	56	100

In order to get evidence for the reliability of the test scores, how much variance is explained by the model for each cognitive domain was computed. Accordingly, reliability coefficients of the results obtained from literacy and numeracy domains range from .86 to .90 (OECD, 2013b). These values are found to be acceptable because they are more than .60, which is the minimum cut-off criteria in social sciences (Zikmund, Babin, Carr, & Griffin, 2010).

Explanatory item-level and individual-level variables

Studies (Bridgeman & Cline, 2000; Masters, Schnipke, & Connor, 2005; and Yang, O'Neill, & Kramer, 2002) examining the factors that have an influence on the time individuals spend on responding to a test item have considered item difficulty, item type, content area, degree of abstraction, etc. as an item level variable. However, in this study, since not all items and thus their characteristics are released by the OECD, only the item difficulty variable (OECD, 2013b) is considered the item-level variable as taken by the similar study of Goldhammer et al. (2017).

The cognitive pre-test is a kind of short test given to examinees to determine whether they are directed to full computer-based assessment of PIAAC. It includes three literacy and three numeracy items of low difficulty. If the examines failed from this test, they will be given the reading components of the assessment. On the other hand, if they achieve this test, they will take the full assessment (OECD, 2013b).

In PIAAC, there are several demographic variables regarding examinees. One of them is gender. More precisely, in this assessment, examinees are required to provide information about their gender. Also, there is an item which assesses examinees' age in 10-year bands such as 24 or less, 24-34, 35-44, 45-54, and over 55. Another demographic variable assessed in PIAAC is educational attainment, which refers to the highest level of schooling. This categorical variable includes categories such as less than high school, high school, and above high school.

In PIAAC, examinees' readiness to learn is also measured. Specifically, there are six items focusing on the extent to which the examinees deal with problems and tasks they encounter. With these questions, they are asked how often they relate a new idea to the real-life situation and what they learned before, they are willing to learn something new, try to learn hard things in all details, and search for additional information to make it understandable when something they don't understand (Perry, Helmschrott, Konradt, & Maehler, 2017).

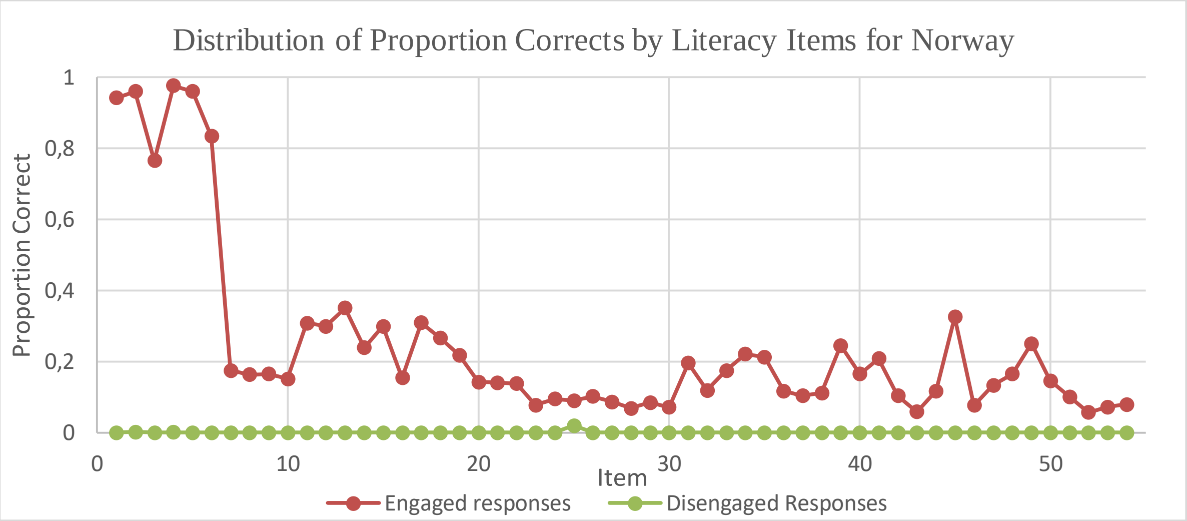
One of the variables measured in PIAAC is the use of ICT at work. There are a set of questions about the frequency of the use of computers or the internet as part of their job. More precisely, these questions focus on the use of e-mail, the internet for understanding job-related issues, conducting transactions on the internet, participating in real-time discussions on the internet, and the use of spreadsheets and word processing and the use of a programming language to program or write computer code. For measuring the use of ICT at home, the same questions were exposed to the examinees. However, this time these questions focus on the frequency of doing these activities in everyday life. All in all, examinees are divided into subcategories according to their frequency of using ICT at work, from those who use it least to those who use it most (OECD, 2015).

Data Analysis

The following procedure was followed to identify disengaged behaviors. If the time taken to respond to an item is below the threshold, it is considered that insufficient effort has been made for that item. To compute item-specific thresholds, the proportion correct greater than zero (P+>0%) method was used. Before seeking answers for the research questions, the time spent on the item was converted to a dichotomous engagement indicator (0 = disengaged, 1 = engaged) as an item response variable depending on whether the response time was below or above the response time thresholds. The variables cognitive pre-test score and item difficulty were centered and scaled to make a more meaningful interpretation of interaction effects.

Validity checks

In the present study, two validity checks were used to ensure that the threshold procedure employed accurately identified disengaged responses. In the first validity check, the engaged and disengaged response behaviors were compared in terms of their proportion correct (e.g., Wise & Kong, 2005; Wise & Ma, 2012). In order for the threshold determination process to be valid, the proportion correct for engaged behavior should be higher than the chance level, and the proportion correct for disengaged behavior should be at the level of chance. Considering that the items measuring verbal and numerical skills of adults in the PIAAC application have many response options, the probability of finding the correct answer by chance is very close to zero or zero. In the present study, the distributions of the observed proportion correct for responses classified as engaged or disengaged using the proportion correct conditional method (P+>0%) were examined for each domain and country. Accordingly, it was proven that the proportion correct for disengaged response behavior was found to be close to zero or zero, whereas the proportion correct for engaged response behavior was much higher. As an example, the distribution of the proportion correct scores of the engaged and disengaged individuals in Norway for each domain is presented in Figure 1.



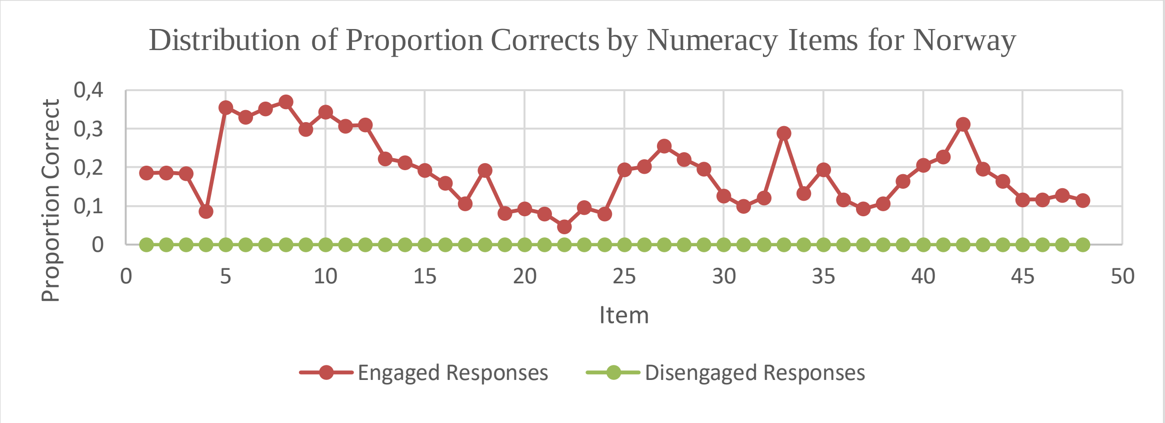


Figure 1. Distributions of the Proportion Correct Scores of Engaged and Disengaged Responses

In the upper part of Figure 1, the red line shows the proportion correct for engaged response behavior while the lower green line represents the corresponding proportion correct for disengaged response behavior. Figure 1 clearly shows that the proportion correct scores of the engaged individuals were higher than those of the disengaged individuals in Norway. A similar pattern was also observed in the other selected countries.

Another validity check for each item and domain was the examination of the association between the proficiency scores of individuals and the proportion correct of engaged and disengaged behaviors (e.g., Lee & Jia, 2014). According to the proficiency scores, individuals are divided into different groups referred to as score groups. In order for the threshold determination process to be valid, it is expected that there must be a positive relationship between the proportion correct and proficiency scores of the engaged responses for each item. No such relationship is expected for disengaged behaviors.

In the current study, the participants were divided into six score groups ranging from low competency to high competency as defined by PIAAC competency levels (OECD, 2013a) for both domains. Regardless of which plausible value is taken for examinees, individuals are at the same competency level defined by PIAAC. Furthermore, the plausible values were not used in the main analysis, but only as a proof of validity check. Therefore, in order to provide ease in calculations and interpretations, in assigning people to score groups, the mean of the adults' 10 plausible values regarding both domains was used. For each item, the relationship between the proficiency scores of the participants (i.e., an average of plausible values) and the proportion correct scores of engaged and disengaged response behaviors were investigated. Figure 2 shows the related findings for the selected literacy and numeracy items.

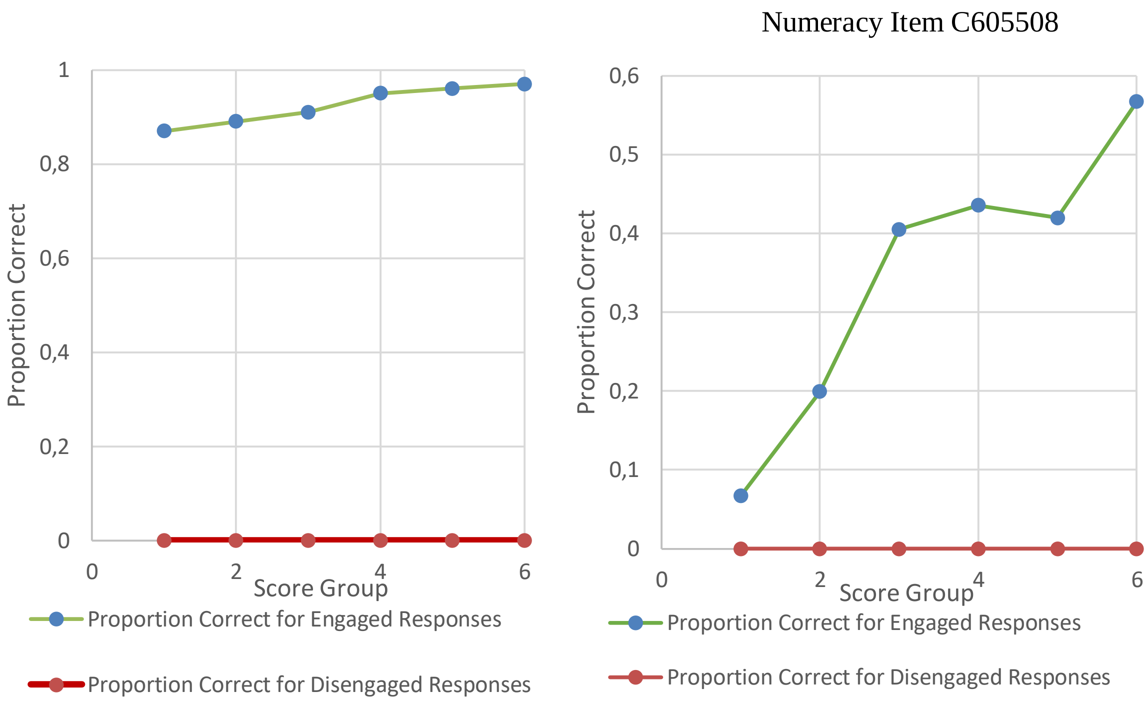


Figure 2. Association between the Score Groups and Proportion Correct Scores in Selected Literacy (C301C05) and Numeracy Items (C605508)

In both figures, the upper and lower lines show the association between the score groups and proportion correct scores for engaged and disengaged response behaviors, respectively. As expected, the association between the score group (plausible values) and the proportion correct for engaged response behavior was positive for all items in both domains.

Once the validity of the procedure for determining a threshold was proven, a 1-parameter logistic (1PL) item response model for each domain with dichotomous engagement indicators (0 = disengaged, 1 = engaged) was tested as an item response variable. 1PL models assume uni-dimensionality and equal discriminations across items. To determine the item fit, information-weighted (Infit) and unweighted (Outfit) mean-squared residual-based item fit statistics were inspected. If the infit and outfit values are between .5 and 1.5, it shows that the item fits the data (de Ayala, 2009). Thus, for each country and domain, very few items that did not fit the data were removed from the data set, which will not distort the representativeness of items. Specifically, for the countries Norway, Austria, Denmark, Germany, and Ireland, nine items were removed from the literacy survey, while seven items were removed from the numeracy survey. Furthermore, for Finland, three items were not included in the analysis of the responses to the literacy survey, while seven of the items were removed from the numeracy survey. Lastly, for France, the numbers of the items excluded from the data sets regarding the domain of literacy and numeracy were six and four, respectively.

Different EIRT models were constructed due to their flexibility to include the effect of the item and person-level variables simultaneously (Briggs, 2008). These models can be used for measurement and explanation purposes. The EIRM approach defines individuals as clusters, items as the repeated observations, and item responses as the dependent variable within a multilevel structure. In other words, the EIRT is of the multilevel models in which individuals' item responses are considered as the first-level factors, individuals are considered as second-level factors, and the individuals' and/or items' characteristics are included as predictors (De Boeck & Wilson, 2004).

Accordingly, after testing the baseline model, Model 0 and Model 1 with personal characteristics, such as educational attainment, gender, age group, cognitive skill, readiness to learn, and use of ICT skills at home and work were tested. Model 2 included only item difficulty since an item characteristic was being tested. Finally, the full Model 3 was tested with item- and person-level variables and the interaction of item difficulty with cognitive skill. After running all models, likelihood-based fit statistics, such as the likelihood-ratio (LL) statistics, Akaike's information criterion (AIC), and the Bayesian information criterion (BIC), were determined. All models were estimated in the R environment (R Core Team, 2016). The TAM package (Kiefer, from the "lme4" package (Bates, Maechler, Bolker & Walker, 2015) was used to test explanatory item response models. The intra-class correlation (ICC) for each domain and country was computed to determine the proportion of variance in the dependent variable and the test-taking engagement that is attributed to personal differences. ICC is calculated by dividing the random effect variance by the total variance (Hox, 2002).

RESULTS

Model-Fit

For both literacy and numeracy domains, four explanatory IRT models were tested, and the LL, BIC, and AIC values were examined to determine the most appropriate IRT model for PIAAC 2012. There is no general rule about which model (the most complex or simpler) will fit the data. Therefore, in this study, although it was not predicted that Model 3 would definitely fit better before, it is predicted that item and individual-level variables may be effective on individuals' engagement levels. When the results were examined, it was found that Model 3 fitted the PIAAC 2012 data best because of the lower values of these indices. Therefore, the results of Model 3 were taken into consideration in this study. The model-fit results were presented in Table 3.

Table 3. Model-fit Results for Literacy and Numeracy Domains

			Literacy			*Numeracy*	
Country	*Model*	*AIC*	*BIC*	*LL*	*AIC*	*BIC*	*LL*
Austria	Model 0	172115.3	172145.7	-86054.6	172105.0	172135.5	-86049.5
	Model 1	172120.5	172394.4	-86033.3	172120.4	172394.2	-86033.2
	Model 2	170062.5	170123.3	-85025.2	169870.9	169931.8	-84929.4
	Model 3	169761	170065.3	-84850.5	169543.1	169847.4	-84741.6
Denmark	Model 0	265707.4	265739.2	-132851	264451.8	264483.6	-132222.9
	Model 1	265736.3	266054.2	-132838	264481.8	264799.8	-132210.9
	Model 2	262570.4	262634	-131279	261320.8	261384.4	-130654.4
	Model 3	261618.6	261968.3	-130776	260473.2	260823.0	-130203.6
Germany	Model 0	203498.3	203529.2	-101746	201059.9	201090.8	-100527.0
	Model 1	203523.5	203832.7	-101732	201082.9	201392.1	-100511.4
	Model 2	201159.3	201221.1	-100574	198410.1	198472.0	-99199.1
	Model 3	200581.1	200921.1	-100258	197927.6	198267.7	-98930.8
France	Model 0	166548.8	166578.4	-83271.4	166045.9	166075.7	-83020.0
	Model 1	166367.6	166634.2	-83156.8	165999.2	166267.3	-82972.6
	Model 2	166427.4	166476.8	-83208.7	157061.9	157111.3	-78526.0
	Model 3	165945.2	166231.5	-82943.6	157029.8	157316.1	-78485.9
Ireland	Model 0	177264.2	177294.8	-88629.1	181819.0	181849.6	-90906.5
	Model 1	177291.4	177587.2	-88616.7	181843.9	182170.3	-90889.9
	Model 2	174959.1	175020.3	-87473.6	179729.1	179790.3	-89858.6
	Model 3	174546.4	174883.0	-87240.2	179283.9	179620.5	-89608.9
Finland	Model 0	182698.3	182729.4	-91346.2	181819.0	181849.6	-90906.5
	Model 1	180477.6	180788.8	-90208.8	181843.8	182170.3	-90889.9
	Model 2	182685.4	182747.6	-91336.7	179729.1	179790.3	-89858.6
	Model 3	180465.1	180807.4	-90199.5	179283.9	179620.5	-89608.9
Norway	Model 0	185127.2	185158.0	-92560.6	189895.0	189925.8	-94944.5
	Model 1	185146.7	185454.4	-92543.4	189924.3	190232.0	-94932.2
	Model 2	182599.0	182660.5	-91293.5	187528.2	187589.7	-93758.1
	Model 3	182067.7	182406.2	-91000.9	186871.6	187210.1	-93402.8

Differences in Test Engagement

For each country, the results regarding the effects of the item- and person-level factors on test-taking engagement are presented in Tables 4 and 5.

Table 4. Results regarding Items Assessing Literacy

Variables	Subcategory	Austria	Finland	Norway	Denmark	Germany	France	Ireland
Intercept		-13.9**	-20.86**	-1.08**	-1.13*	-12.98**	-.09**	-10.73**
Difficulty		-	-2.68**	-			-.83**	-
Cognitive pre-test x difficulty		.11**		.15**	.16**	.14**	.16**	.19**
Age	35-44	-	-	-	-		-.07**	-
	45-54	-	-	-	-		-.05**	-
	Over 55	-	-	-.23'	-		-.10**	-
Educational attainment	Above high school	-	.35'	-	-	.22'	-	-
Readiness to learn	Lowest to 20%	12.43**	-3.34**	-	-	12.25**	-	9.45**
	More than 20% to 40%	12.41**	-3.14**	-	-	12.36**	.05'	9.41**
	More than 40% to 60%	12.36**	-2.60**		-	12.2**	.05'	9.54**
	More than 60% to 80%	12.21**	-2.48**	-	-	12.22**	-	9.59**
	More than 80%	12.44**	-2.62**	-	-	12.32**	-	9.59**
Use of ICT at home	lowest to 20%	1.12**	-	-	-	-	-	
	More than 20% to 40%	1.1**	-	-	-	-	-	.78'
	More than 40% to 60%	1.09**	-	-	-	-	-	-
	More than 60% to 80%	1.27**	-	-	-	-	-	-
	More than 80%	1.22**	-.81'	-	-		-	-
Use of ICT at work	lowest to 20%	-	18.16**	-	-	-		-
	More than 20% to 40%	-	23.42**	-	-	-	-	-
	More than 40% to 60%	-	23.42**	-	-	-	-	-
	More than 60% to 80%	-	23.24**	-	-		-	-
	More than 80%		23.06**	-			-	-
ICC		.49	.48	.50	.49	.48	.50	.50

** *p* < .001, **p* < .01, ' *p* < .05

Table 5. Results regarding Items Assessing Numeracy

Variables	Subcategory	Austria	Finland	Norway	Denmark	Germany	France	Ireland
Intercept		-	-1.84**	-3.06**	-1.72**	-1.46*	-	-1.87**
Difficulty	-	-	-	-	-	-	-	-
Cognitive pretest	-	.09'	.07*	-	.08'	-	.03**	.07*
Cognitive pre-test x difficulty		.39**	.13**	-	.51**	.43**	-	.13**
Age	25-34	-	-	.21'	-	-		-
	35-44	-	-	-	-	-	-.06*	-
	45-54	-	-.25*	-	-	-	-.04'	-.24*
	Over 55	-		-			-.09**	
Gender	Female		-	.15*	-	-	-	-
Readiness to learn	lowest to 20%	-	9.92**	-	-	-	-	9.95**
	More than 20% to 40%	-	9.82**	-	-	-	.05'	9.84**
	More than 40% to 60%	-	9.80**	-	-	-	.05'	9.83**
	More than 60% to 80%	-	9.80**	-	-	-	.05'	9.83**
	More than 80%	-	9.81**	-	-	-	-	9.84**
Use of ICT at home	lowest to 20%	-	-	1.28*	-	-	-	-
	More than 20% to 40%	-	-.79'	1.42**	-	-	-	-.79'
	More than 40% to 60%	-	-	1.43**	-	-	-	-
	More than 60% to 80%	-	-	1.47**	-	-	-	-
	More than 80%	-	-.86'	1.4*	-	-	-	-.86'
Use of ICT at work	lowest to 20%	-.35*	-	-	-	-	-	-
ICC		.49	.50	.50	.51	.48	.50	.49

** $p < .001$, *$p < .01$, ' $p < .05$

As shown in Table 4, the difficulty of items measuring literacy had a negative effect on the engagement of participants in France (-.93) and Finland (-2.68), showing that when the item difficulty increased, adults tended not to give sufficient time to the items. On the other hand, the difficulty of items measuring numeracy was found to have no significant effect on the engagement of the adults. In addition to the main effect of item difficulty on engagement, the interaction between item difficulty and cognitive skill was also significant. Specifically, the effect of item difficulty on engagement was higher among strong test-takers who put more effort into solving items than poor test-takers who did not put sufficient effort into items.

Age had a statistically significant on the engagement of participants in literacy items in France and Norway. Specifically, as the age of the French participants increased, they tended to be disengaged. Additionally, there was a particularly strong decrease in the engagement rate of the oldest group, participants aged 55 or above in Norway (-.23). A similar pattern was also found for the domain of numeracy. Moreover, the significant negative effect of age on the engagement of the adults taking the numeracy items was observed in the countries of Ireland and Finland.

The highest level of educational attainment was associated with higher engagement in Germany (.22) and Finland (.35). In other words, individuals with a high level of education in Germany spent more time answering the questions. When the results were examined in terms of the numeracy domain, as shown in Table 5, it was found that educational attainment had no significant effect on engagement.

As it was clearly seen in Table 4, for Austria, Germany, France, and Ireland, the adults' readiness to engage in learning activities had a positive effect on their engagement on items addressing literacy skills. However, this was not the case for the participants from Finland. The adults' readiness to engage in learning activities which require the use of literacy skills had a negative effect on their engagement. The finding was that the adults who were highly ready to learn put insufficient effort into answering the items. For the domain of numeracy, as s presented in Table 5, a similar pattern observed for literacy domain was also found in France and Ireland in terms of the effect of adults' readiness to learn on their engagement levels. That is, as the level of readiness to learn of the adults increased, their test-engagement levels also increased when responding to the items assessing numeracy items.

For the literacy domain, Table 5 shows that the effect of the use of ICT skills at home of individuals from each category in Austria on their engagement levels was positive and significant, suggesting that the test-takers who more frequently used ICT skills at home exhibited a higher level of engagement. In contrast, the use of ICT skills at home was negatively associated with the adults' engagement in numeracy in Ireland (-.79) and Finland (-.79), but the use of ICT skills at home for each category of the individuals in Norway was positively related to the students' engagement in numeracy.

When the effect of the use of the ICT skills of individuals at work was examined across all countries, according to Table 4, it was found that in Finland, those who more frequently used ICT skills at work tended to be more engaged while responding to the items measuring literacy. On the other hand, this was not the case for the field of numeracy. A negative and significant effect (-.35) of the use of ICT skills at work on the engagement of individuals in Austria was found, suggesting that the adults who used ICT skills frequently at work tended to be disengaged when answering the items in the test. When the findings regarding gender were considered, it was determined that for only the field of numeracy, in Norway, being female (.15) was found to be positively related to test-taking engagement.

For each country and domain, as presented in Tables 4 and 5, the ICC values taking into account the adults' test-taking engagement differences at the person level were found to be similar to each other. Specifically, approximately 50% of the variation in engagement levels of individuals was attributable to differences between subjects.

DISCUSSION and CONCLUSION

This study aimed to determine which of the explanatory IRT model was the best fit for the analysis of the PIAAC sub-data. In addition, the present study aimed to investigate the effect of person- and item-level factors depending on the analysis of the model that best fitted the data. To achieve these aims, predictions were created utilizing different models for the domains of literacy and numeracy.

The conclusion of this study is that there is increasing disengagement in more difficult items measuring literacy skills, thus indicating that individuals spend little time on very difficult items (OECD, 2013a). When individuals perceive an item to be very difficult, they may tend to stop trying to understand and respond to the item very quickly. Considering that the data in this study belonged to the low stake assessment, the low motivation of the participants may have played a role in this outcome. Furthermore, whether a particular item is perceived as 'too difficult' depends on the cognitive level of the adult. The reason behind this finding is that there is a significant and positive effect of the interaction between cognitive pre-test and item difficulty on test engagement (Wise & Kingsbury, 2015). In other words, the significant effect of the interaction between item difficulty and cognitive pre-test shows that individuals tend to engage in relation to their cognitive skills.

Older adults tend to exhibit a higher propensity to disengage in both fields. Increasing disengagement by older test-takers in items in technology-rich environments may be related to their lower levels of ICT experience and skills (OECD, 2013a). They have more difficulty than their younger counterparts in using computers due to age-associated changes in visual, perceptual, psychomotor, and cognitive abilities. Older people with insufficient experience with computers may also have a negative attitude toward computer usage (Xie, 2003), which may cause disengaged behaviors in testing.

Additionally, the present study revealed that more educated individuals were more engaged in the items assessing literacy. This finding is supported by the study of Goldhammer, Martens, Christoph, and Lüdtke (2016), in which the effect of educational attainment on the individual' disengagement was investigated. There may be several reasons for this result. Firstly, compared to individuals who are less educated, highly educated individuals are relatively more proficient and more likely to respond to more difficult items. Secondly, since those with higher education are more accustomed to testing and assessment environments; thus, they may get less tired than test takers with lower education levels. As a result, the former do not stop trying to give an answer to an item. Lastly, people with a high level of education may have a stronger sense of commitment to completing the assessment, which makes them put more effort into solving the items. Those people with a low level of education may have difficulty in understanding the items. They may not have sufficient literacy and numeracy skills (OECD, 2019), which can result in a tendency to respond to items quickly.

Individuals who are more ready to learn tend to exhibit more engagement in the items. The reason behind these results might be related to the composite feature of the readiness to learn, which consists of attitudinal or emotional, cognitive, behavioral, and, to a lesser extent, personality or dispositional components (Smith, Rose, Ross-Gordon & Smith, 2015). Therefore, individuals who are more ready to learn are more attentive, willing, and motivated to learn. Thus, they can easily concentrate on the items and complete them without getting bored (Eccles & Wigfield, 2002).

The current study concluded that adults who frequently used ICT skills at home and work engaged more than the adults that rarely used ICT skills. This finding is in line with the literature that suggests individuals with strong ICT skills engage more in a technology-enriched environment (Bergdahl, Nouri & Fors, 2019). This can be explained by familiarity with ICT which has an effect on the motivation and engagement of individuals (OECD, 2019).

It is concluded that gender has a significant effect on adults' engagement in items assessing numeracy skills, suggesting that engagement can be seen as a domain-specific construct (Goldhammer et al., 2016); for example, in Norway, females exhibit a higher level of engagement. This finding is also supported by the study of Marrs and Sigler (2012). They found that females tended to engage in the material at a deeper level, whereas males tended to display minimal effort.

Interpreting the results regarding literacy obtained from this study in terms of country groups according to t-disengagement percentages shows that the use of ICT skill had no effect, except for the test-taking engagements of countries with a low t-disengagement percentage. On the other hand, for the numeracy domain, there were several similarities in the effect of person-level factors on the same country groups. For example, the effect of age and readiness to learn on countries with a high t-disengagement percentage was similar. For the numeracy domain, age had a negative effect on test-taking engagement for adults in both France and Ireland, whereas readiness to learn had a positive effect. Additionally, it was concluded that some personal-level variables (age, gender, readiness to learn, and use of ICT skills at home and work) did not have an effect on the test-taking engagement of countries with a relatively moderate t-disengagement percentage.

To make more accurate evaluations, it is suggested that assessment practitioners should manage disengagement by identifying disengaged responses when obtaining test scores and filtering such responses in the data. Additionally, adults can be provided with valuable feedback regarding their performance (DeMars et al., 2013). One or more of these methods can be used for the validity of the results obtained from low-stake assessments. Underestimating disengaged responses may have significant negative consequences due to the potential high-stakes nature of international assessments for educational stakeholders and policymakers. By demonstrating the differential predictors of disengaged responses by country, this study revealed the potential for educational stakeholders to make inaccurate inferences when comparing subgroup performance across countries. For example, when comparing performance by gender, it is possible that score differences observed between males and females across countries may be confused with test-taking effort as opposed to true differences. Since such effects may be investigated as a basis for constructing national education policy reform, it is crucial that disengaged responses are identified and filtered before performing operational analyses (e.g., item analyses) and research analyses (Rios & Guo, 2020). These recommendations are some examples of how the results of this new study can be used and how they can benefit practitioners. However, in any case, the most important message that can be derived from this study is that the source of the differences in the scores of individuals in low-stake assessments may be their disengagement levels. Future research can be conducted to explore the extent to which these factors developed in recent years are effective in disengagement under low-stakes conditions.

The findings from this study offer practical uses; however, they are limited in a number of ways. Firstly, in this study, a selection was made from countries with different levels of disengagement, but not all countries participating in PIAAC 2012 were included. The findings of the present study cannot be generalized to adults; thus, further similar research is required. Secondly, this study used only one method to determine response time thresholds. Since there are many other methods to detect disengaged behaviors, future research can be conducted to compare the effectiveness of these methods. Despite the limitations of this study, it is considered that it draws further attention to the role of test-taking effort in international assessments and contributes to the discussion of investigating test-takers' effort as part of standard operational practices.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and

Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48

Bergdahl, N., Nouri, J., & Fors, U. (2019). Disengagement, engagement and digital skills in technology-enhanced learning. *Education and Information Technologies*, 149. doi:10.1007/s10639-019-09998-w.

Braun, H., Kirsch, I., Yamamoto, K., Park, J., & Eagan, M. K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89-118. doi: 10.1080/08957340801926086.

Bridgeman, B., & Cline, F. (2000). *Variations in mean response time for questions on the computer-adaptive GRE General Test: Implications for fair assessment.* GRE Board Professional Report No. 96-20P. Princeton, NJ: Educational Testing Service.

Brown, G. T. L., & Harris, L. R. (2016). *Handbook of human and social conditions in assessment*. New York, NY: Routledge.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. doi:10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practices in Assessment, 8*, 69-82.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015(2),*1–17. doi: 10.1002/ets2.12067

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education, 5*(18), 1-25. doi: 10.1186/s40536-017-0051-9.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC.* Vol. 133. In: OECD Education Working Papers. Paris: OECD Publishing.

Hox J. 2002. *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules*. R package version 1.99–6. Retrieved from http:// CRAN.R-project.org/package=TAM

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(1), 1–24. doi: 10.1186/s40536-014-0008-1.

Marrs, H., & Sigler, E. A. (2012). Male academic performance in college: The possible role of study strategies. *Psychology of Men & Masculinity, 13*(2), 227-241.

Masters, J., Schnipke, D. L., & Connor, C. (2005, April). *Comparing item response times and difficulty for calculation items*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Mastuti, E., & Handoyo, S. (2017, October). *Effects of individual differences on the performance in computer-based test (CBT).* Paper presented at the 3rd ASEAN Conference on Psychology, Counselling, and Humanities (ACPCH). Malang, Indonesia.

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: an IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling, 60*(2), 165–187.

Organisation for Economic Co-operation and Development. (2013a). *OECD skills outlook 2013: First results from the survey of adult skills.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2013b), "The methodology of the Survey of Adult Skills (PIAAC) and the quality of data", in *The*
Survey of Adult Skills: Reader's Companion, OECD Publishing, Paris.

Organisation for Economic Co-operation and Development. (2013c). *What the Survey of Adult Skills (PIAAC) measures in The Survey of Adult Skills: Reader's Companion*, OECD Publishing, Paris. doi: https://doi.org/10.1787/9789264204027-4-en

Organisation for Economic Co-operation and Development. (2015). *Adults, Computers and Problem Solving: What's the Problem?, OECD Skills Studies*, OECD Publishing, Paris, https://doi.org/10.1787/9789264236844-en.

Organisation for Economic Co-operation and Development. (2016a).Technical report of the Survey of Adult Skills (PIAAC) (2nd edition). OECD, Paris,

Organisation for Economic Co-operation and Development. (2016b). *The Survey of Adult Skills: Reader's Companion*, Second Edition, OECD Skills Studies, OECD Publishing, Paris. doi: 10.1787/9789264258075-en.

Organisation for Economic Co-operation and Development. (2017). *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, GESIS Data Archive, Cologne, doi:10.4232/1.12955.

Organisation for Economic Co-operation and Development. (2019), *Beyond Proficiency: Using Log Files to Understand Respondent Behaviour in the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris. doi: 10.1787/0b1414ed-en.

Perry, A., Helmschrott, S., Konradt, I., & Maehler, D. B. (2017). *User Guide for the German PIAAC Scientific Use File*: Version II. (GESIS Papers, 2017/23). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-54438-v2-7

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*. Advance online publication. doi: 10.1080/08957347.2020.1789141

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1),74-104. doi: 10.1080/15305058.2016.1231193

Smith, M C., Rose, A.D., Smith, T. J.& Ross-Gordon, J. M. (2015, May). *Adults' readiness to learn and skill acquisition and use: An analysis of PIAAC*. Paper presented at the 56th Annual Adult Education Research Conference. Manhattan, KS.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a largescale assessment. *Applied Measurement in Education, 26*(1), 34–49. doi: 10.1080/08957347.2013.739453

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26. doi:10.1016/S0361-476X(02)00063-2.

Team, R. C. (2016). *R: A language and environment for statistical computing* (Version 3.1.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Xie, B. (2003). Older adults, computers, and the Internet: Future directions. *Gerontechnology,2*(4), 289-305.

van Barnevald, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31*(1), 31–46. doi:10.1177/0146621606286206

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19*(2), 25-114.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*(3), 152–166. doi:10.1353/jge.0.0042

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252. doi:10.1080/08957347.2015.1042155

Wise, S. L. , & DeMars, C. E. (2009). A Clarification of the effects of rapid guessing on Coefficient α: A note on Attali's "reliability of speeded number-right multiple-choice tests". *Applied Psychological Measurement , 33*(6), 488–490. doi:10.1177/0146621607304655

Wise S. L. & DeMars C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27-41.

Wise, S. L., & Kingsbury, G. G. (2015). *Modeling student test-taking motivation in the context of an adaptive achievement test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Yang, C. L., O'Neill, T. R., & Kramer, G. A. (2002). Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement, 3*(3), 282-299.

Zikmund, W.G., Babin, B.J., Carr, J.C. and Griffin, M. (2010). *Business research methods,* Canada:South-Western Cengage Learning.

PIAAC 2012'de Test Katılımının Sessiz Yordayıcıları

Giriş

Bireylerin düşük riskli uluslararası değerlendirmelerde güdülerinin düşük olması testteki maddeleri cevaplamaya yeterince zaman ayırmamalarına neden olmaktadır (Wise ve DeMars, 2010). Bu durum testin psikometrik özelliklerin bozulmasına (Rios, Guo, Mao, & Liu, 2017)v e veri setinden elde edilen sonuçların yanlış yorumlanmasına yol açmaktadır (Nagy, Nagengast, Becker, Rose ve Frey, 2018). Daha doğrusu, bireylerin gerçek puanlarına, teste katılım seviyelerine bağlı olarak sistematik bir hata karışmaktadır (Braun, Kirsch, Yamamoto, Park ve Eagan, 2011). Bunun yanı sıra, bireylerin teste yeterince zaman ayırmamaları (a) madde güçlük ve ayırıcılık parametrelerinin olduğundan daha yüksek (van Barnevald, 2007) (b) madde ve test bilgi fonksiyonlarının yanlı olarak (van Barnevald, 2007), (c) klasik test teorisine dayalı güvenilirlik tahminlerinin olduğundan yüksek (Wise & DeMars, 2009), (d) değişen madde fonksiyonun yanlış (Wise & DeMars, 2010) ve (e) değişkenler arası korelasyonların daha düşük (Wise, 2009) kestirilmesine neden olmaktadır.

Bireylerin testteki maddelere yeterince zaman ayırmamasının nedeni, bu davranışın doğru ve verimli bir şekilde ölçülmesi, testteki maddelere yeterince zaman ayırmamadan kaynaklanan büyük geçerlilik sorunları göz önüne alındığında çok önemlidir. Geniş ölçekli uygulamalardan biri olan PIAAC değerlendirmesine dâhil edilen maddeler zorluk ve karmaşıklık açısından farklılık gösterdiğinden, doğru cevabı vermek için gereken süre birbirinden farklı olacaktır. Bu nedenle, avantajları göz önünde bulundurularak, bu çalışmada testteki maddelere katılım gösteren ve göstermeyen davranışları belirlemede maddeye özgü tepki süresi eşikleri kullanılmıştır (Wise, 2006).

Bireylerin testteki madde üzerinde harcadıkları zaman konusunda kapsamlı araştırmalar yapılmış olsa da, bu çabaların çoğu tek bir ülkede bulunan nispeten homojen popülasyonlara odaklanmıştır (Goldhammer, Martens & Lüdtke, 2017). Kişisel özellikler kültüre veya ülkeye göre büyük ölçüde farklılık gösterse de (Brown ve Harris, 2016) uluslararası değerlendirmelerde ülkeler arasında ülkelerin teste harcadıkları zaman açısından potansiyel farklılıkları inceleyen çok az çalışma yapılmıştır (Rios ve Guo, 2020). Genel olarak, bu çalışma, farklı ülkelerden yetişkinlerin katılımını etkileyen faktörleri daha yakından inceleyerek alan yazındaki bu boşluğu kapatmaya katkıda bulunmaktadır. Bu bağlamda, bu çalışma, PIAAC uygulamasında ele alınan sözel ve sayısal becerilerle ilgili alanlara ilişkin maddelere harcanan zaman üzerindeki çeşitli madde ve birey düzeyindeki değişkenlerinin rolünü incelemeyi amaçlamaktadır. Bu doğrultuda, bu çalışmada cevap aranan araştırma soruları şu şekildedir:

1) Açımlayıcı madde tepki modellerinden hangisi (temel model, birey düzeyindeki değişkenlerinin dâhil edildiği model, madde düzeyindeki değişkenin dâhil edildiği model ve bütün madde ve birey düzeyindeki değişkenlerin ve bunlar arasındaki etkileşimin dâhil edildiği model) PIAAC alt verilerine en iyi uyumu sağlamaktadır?

2) Maddelere katılım gösteren yanıtlar birey ve madde düzeyindeki değişkenlerle açıklanabilir mi?

Yöntem

Çalışmanın hedef evreni veri toplama sırasında ülkede ikamet eden ve PIAAC 2012'ye katılan 16 ila 65 yaşları arasındaki yetişkinler içermektedir. Olasılıklı örnekleme yöntemi kullanılmıştır. Çalışmanın örneklemini teste katılmama düzeylerine seçilen ülkeler oluşturmaktadır. Buna göre, katılmama düzeyi yüksek olan ülke grubundan iki ülke (Fransa ve İrlanda), orta olan iki ülke (Danimarka ve Almanya) ve düşük olan üç ülke (Avusturya, Finlandiya ve Norveç) çalışmaya dâhil edilmiştir.

Çalışmada veri toplama aracı olarak sözel ve sayısal becerileri ölçen anketler kullanılmıştır. Bilgisayar tabanlı değerlendirmeye katılan yetişkinlerin dijital metinleri okuma becerilerinin yanı sıra geleneksel basılı metinleri de değerlendiren sözel becerileri ölçen ankette 58 madde dâhil edilmiştir. Ek olarak, yetişkinlerin matematiksel bilgileri kullanma, uygulama, yorumlama ve iletme yeteneklerini değerlendiren sayısal becerileri ölçen ankette 56 madde dâhil edilmiştir (OECD, 2016).

Maddelere özgü eşik parametrelerini belirlemek için sıfırdan büyük doğru cevaplama oranı (P +>% 0) yöntemi kullanılmıştır. Araştırma sorularına cevap aramadan önce ikili puanlanan yeni bir değişken tanımlanmıştır. Buna göre, maddeye harcanan zaman, madde eşik parametresinin altında veya üstünde bir değer almasına göre yeniden kodlanmıştır (0= katılım göstermemiş, 1= katılım göstermiş). Madde güçlük parametreleri kestirimlerin kolaylaşması açısından 100'e bölünerek yeniden ölçeklendirilmiştir. Etkileşim etkisini belirlemek için açıklayıcı madde tepki modellerinin analizleri sırasında bilişsel ön test puanları ve madde güçlük parametreleri ölçeklendirilmiştir.

Madde eşik parametrelerinin belirlenmesi sürecinin testteki maddelere yeteri kadar katılım göstermeyen ve gösteren bireyleri doğru bir şekilde ayırıp ayırmadığını belirlemek için iki tane geçerlik kontrolü yapılmıştır. Birinci geçerlik kontrolünde, doğru cevaplama oranları katılım göstermiş ve göstermemiş bireyler açısından karşılaştırılmıştır. Geçerli bir belirleme sürecinde, PIAAC uygulamasındaki maddelerin çok sayıda tepki seçeneklerinin olduğu düşünüldüğümde katılım gösteren bireylerin doğru cevaplama oranlarının dağılımı sıfırdan büyük iken katılım göstermeyen bireylerin doğru cevaplama oranlarının dağılımının sıfır veya sıfıra çok yakın olması beklenir. Bu çalışmada da bu durum doğrulanmıştır. Bir diğer geçerlik kanıtı olarak ise farklı yeterlik gruplarında katılım göstermemiş ve göstermiş bireylerin doğru cevaplama oranları karşılaştırılmıştır. Maddelere katılım gösteren bireyler için yeterlik puanları ile doğru cevaplama oranları arasında pozitif yönde ilişki çıkması beklenirken katılım göstermeyen bireyler için manidar bir ilişkinin çıkması beklenmez. Bu çalışma da bu durum doğrulanmıştır.

Her bir alan için 1-parametreli lojistik modeller bireylerin katılım düzeylerini gösteren ve yeniden oluşturulan ikili puanlanan değişkenin varlığında test edilmiştir. Modele uyum sağlamayan maddeler veri setinden çıkarılmıştır. Dört farklı açıklayıcı madde tepki kuramı modeli madde ve birey düzeyindeki değişkenlerin etkilerini aynı anda incelenmesini sağlaması nedeniyle test edilmiştir. Veriye uyum sağlayan modelin belirlenmesinde çeşitli uyum iyiliği indekslerinden yararlanılmıştır. Verilerin analizinde tek boyutluluğu belirlemede R yazılımında "TAM" paketi (Kiefer et al., 2016) ve açıklayıcı madde tepki kuramı modellerinin analizinde ise “lme4” paketi (Bates et al., 2015) kullanılmıştır. Yetişkinlerin maddelere katılım düzeylerindeki varyansı açıklamada bireyler arası farklılıkların etkisini belirlemek için sınıf içi korelasyon katsayıları her bir ülke ve her bir alan için hesaplanmıştır.

Sonuç ve Tartışma

Veriye en iyi uyum sağlayan modelin hem madde düzeyinde hem de birey düzeyindeki değişkenlerin dâhil edildiği model olduğu bulunmuştur. Bu çalışmada, bireylerin çok zor maddelere çok az zaman ayırdıkları sonucuna ulaşılmıştır (OECD, 2013). Kişiler maddenin çok zor olduğunu algıladıklarında, denemeyi bırakıp maddeye çok çabuk cevap verme eğiliminde olabilirler. Bu çalışmadaki verilerin düşük riskli bir değerlendirmeye ait olduğu düşünüldüğünde, bu durumda katılımcıların düşük motivasyonu rol oynamış olabilir. Ayrıca belirli bir maddenin “çok zor” olarak algılanıp algılanmaması yetişkinlerin bilişsel düzeyine bağlıdır. Bu durum, bu çalışmadan elde edilen sonuçlardan biri olan bilişsel ön test ile madde zorluğu arasındaki etkileşimin test katılımı üzerinde anlamlı ve olumlu bir etkisinin olmasıyla da desteklenmektedir (Wise ve Kingsbury, 2015).

Daha yaşlı yetişkinlerin her iki alanda da daha yüksek düzeyde katılmama eğilimi gösterdikleri sonucuna varılmıştır. Teknoloji açısından zengin ortamlardaki değerlendirmelerde nispeten yaşı büyük olan katılımcılarının artan ilgisizliği, teknolojiyle ilgili deneyim ve becerilerinin daha düşük olmasıyla açıklanabilir (OECD, 2013). Bu yüzden özellikle bilgisayar kullanıma yönelik olumsuz tutuma sahip olabilir (Xie, 2003). Bu durum ise onların testteki maddelere yeteri düzeyde katılmamalarına neden olabilir.

Ayrıca bu çalışma, daha eğitimli bireylerin sözel becerileri değerlendiren maddelere daha fazla zaman harcadıklarını ortaya çıkarmıştır. Bu bulgu, Goldhammer, Martens, Christoph ve Lüdtke'nin (2016) eğitim düzeyinin bireyin testteki maddelere katılmamaları üzerindeki etkisinin araştırıldığı çalışmasıyla desteklenmektedir. Bu sonucun birkaç nedeni olabilir. İlk olarak, eğitim düzeyi yüksek olan bireyler, eğitim düzeyi düşük olan bireylere göre görece daha yetkin olduklarından, onlardan daha zor maddelere cevap vermeleri istenebilir. İkinci olarak, test ve değerlendirme ortamlarına daha alışkın oldukları için diğer katılımcılara göre daha az yorulabilirler. Sonuç olarak, maddeye cevap vermeye çalışmaktan vazgeçmeme eğilimi gösterebilir.

Öğrenmeye daha hazır olan bireyler, maddeleri cevaplamada yeterince zaman harcamaktadırlar. Bu durum, öğrenmeye daha hazır olan bireylerin daha dikkatli, daha istekli ve öğrenmeye güdülü olmasıyla açıklanabilir. Böylece maddeler üzerinde kolayca odaklanabilir ve sıkılmadan tamamlayabilirler (Eccles ve Wigfield, 2002).Mevcut çalışmada, BİT becerilerini evde ve işte sıklıkla kullanan yetişkinlerin, BİT becerilerini nadiren kullanan yetişkinlere kıyasla testte yer alan maddeleri cevaplamada yeterince zaman harcadıkları sonucuna varılmıştır. Bu bulgu, yüksek düzeyde BİT becerilerine sahip bireylerin teknolojiyle zenginleştirilmiş ortamlarda daha fazla katıldıklarını belirten alan yazınla paralellik göstermektedir. (Bergdahl, Nouri & Fors, 2019). Bu, bireylerin güdüsü ve katılımı üzerinde etkisi olan BİT’e olan aşinalık ile açıklanabilir (OECD, 2019).

Bu çalışmada cinsiyetin, yetişkinlerin sayısal becerilerini değerlendiren maddelere katılımı üzerinde önemli bir etkisinin olduğu sonucuna varılmıştır ve bu, maddelere katılımın alana özgü bir yapı olduğunu göstermektedir (Goldhammer, Martens & Lüdtke, 2016). Daha açık olarak belirtmek gerekirse, Norveç'teki kadınlar maddelere cevap vermede daha yüksek düzeyde katılım sergilemektedir. Bu bulgu, Marrs ve Sigler'in (2012), kadınların kendilerine verilen göreve daha yüksek düzeyde katılma, erkeklerin ise minimum çaba gösterme eğiliminde olduğunu belirten çalışmasıyla uyumludur.

Daha doğru değerlendirmeler yapmak için, uygulayıcılar test puanlarını hesaplarken ve verilerdeki bu tür yanıtları belirleyerek filtreleyebilir. Ayrıca yetişkinlere performanslarıyla ilgili değerli geri bildirimler de sunulabilir (DeMars, Bashkov & Socha, 2013). Düşük riskli değerlendirmelerden elde edilen sonuçların geçerliliği için bu yöntemlerden bir veya daha fazlası kullanılabilir. Bununla birlikte, her durumda, bu çalışmadan çıkarılabilecek en önemli mesaj, bireylerin sonucuna dayalı olarak önemli kararların alınmadığı (geçti-kaldı, veya seviye atlama gibi) değerlendirmelerdeki puanlarındaki farklılıkların kaynağının, bireylerin maddelere yeterince zaman ayırmama davranışı olabileceğidir.

Bu çalışmadan elde edilen bulgular, pratik açıdan faydalı olmasına rağmen birkaç yönden sınırlıdır. İlk olarak, bu çalışmada, farklı seviyelerde katılmama düzeyindeki ülkelerden bir seçim yapılmasına rağmen, PIAAC 2012'ye katılan tüm ülkeler bu çalışmaya dahil edilmemiştir. Bu çalışmanın bulguları bütün yetişkinlere genellenemeyebilir. Bu nedenle, bulgular gelecekteki araştırmalarda tekrarlanmalıdır. İkinci olarak, bu çalışmada tepki süresi eşiklerini belirlemek ve böylece katılmama ve katılma davranışlarını sergileyen bireyleri ayırt etmek için yalnızca bir yöntem kullanılmıştır. Katılmama davranışı sergileyen bireyleri tespit etmek için başka birçok yöntem vardır. Dolayısıyla, bu yöntemlerin etkinliğini karşılaştırmak için araştırmalar yapılabilir.