International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 656–669
https://dx.doi.org/10.21449/ijate.635675
Published at http://www.ijate.net http://dergipark.org.tr *Research Article*

Developing an Item Bank for Progress Tests and Application of Computerized Adaptive Testing by Simulation in Medical Education

Ayşen Melek Aytuğ Koşan[1,*] **Nizamettin Koç**[2],
Atilla Halil Elhan[3], **Derya Öztuna**[3]

[1]Çanakkale Onsekizmart University; School of Medicine, Medical Education and Informatics Department, Çanakkale, Turkey
[2]Ankara University School of Education, Measurement and Evaluation Department, Ankara, Turkey
[3]Ankara University School of Medicine, Biostatistics Department, Ankara, Turkey

ARTICLE HISTORY

Received: 21 October 2019
Revised: 26 November 2019
Accepted: 17 December 2019

KEYWORDS

Computer Adaptive Test,
Rasch Models,
Progress Test,
Medical Education,
Item Bank Development

Abstract: Progress Test (PT) is a form of assessment that simultaneously measures ability levels of all students in a certain educational program and their progress over time by providing them with same questions and repeating the process at regular intervals with parallel tests. Our objective was to generate an item bank for the PT and to examine the possible fit of CAT for PT application. This study is a descriptive study. 1206 medical students participated. During the analysis of the psychometric properties of PT item bank, "the Rasch model for dichotomous items was used". Several CAT simulations were performed by applying various stopping rules of different standard errors. CAT simulation estimates were compared with the estimates generated from the original calibration of the Rasch model where all items were included. After Rasch analysis, a unidimensional PT item bank consisting of 103 items was obtained. The item bank reliability was calculated as 0.77 with Person Separation Index (PSI) and Kuder-Richardson Formula 20 (KR-20). A high correlation between θ estimations obtained from paper-and-pencil (θRM) and CAT applications (θ_{CAT}) was detected for simulation conditions ([N(0,1)] and [N(0,3)]) at the end of our analysis. In CAT, estimation can be made with an average of 14 questions (reduced 86,4%) and 17 questions (reduced 83,4%) [for N(0,1) and [N(0,3) respectively] with reliability of 0,75. This study reveals that it is possible to develop an appropriate item bank for the PT, and the difficulty of administering large number of items in PT can be scaled down by incorporating CAT application.

1. INTRODUCTION

A Progress Test (PT) is a type of assessment that simultaneously measures the ability levels of all students in a certain educational program and their progress over time during that program by providing them with the same questions and repeating the process at regular intervals with parallel tests (Freeman, 2010). Following a blueprint geared toward the cognitive learning objectives anticipated at the end of the curriculum, a question sample that is representative of

CONTACT: Ayşen Melek AYTUĞ KOŞAN ✉ aysenay1@yahoo.com Canakkale Onsekizmart University; School of Medicine, Medical Education and Informatics Department, Canakkale, TURKEY

ISSN-e: 2148-7456 /

all the disciplines and content areas is used in the PT. Due to its contributions to education, PT is used in medical education worldwide, however it has numerous drawbacks as it demands lots of manual effort from the human resources based on the time needed for its preparation, the implementation itself, and evaluation of the results. It also consists of numerous questions, making it a less attractive method for students as it results in their exhaustion (Wrigley. 2012). Implementing PT with the Computerized Adaptive Testing (CAT) method would cut down on the time and human effort needed in the evaluation of medical knowledge as it helps limit the number of items. This also reduces the bulkiness of the process and in turn the negative feelings associated with it.

Purpose of the study; this study aims to develop an item bank for the PT used by the Ankara University School of Medicine (AUSM) and to investigate the capability of CAT for evaluating the medical knowledge in AUSM students

2. METHOD

Sample/Working Group/Participants

This study is a descriptive study. PT is a component of the evaluation system in AUSM. Participants volunteered for PT and received bonus points on their final examinations. The data used for this study was obtained from the results of a PT taken in the 2010-2011 academic year. One thousand two hundred six medical students participated in the PT in grades 1-5 at the AUSM (89.7% of the total) in 2011.

Data Collection Instruments/Data Collection Methods/Data Collection Techniques

The study was divided into two phases: (i) the development of an item bank using a dichotomous Rasch model; (ii) a simulation study to investigate the performance of CAT application with stopping rules of various standard errors or reliability values; and examination of agreement between ability estimates derived from simulated CAT (θ_{CAT}) application and ability estimates from the Rasch model (θ_{RM}) derived from all the items included in the original calibration.

2.1.1. *Development of item bank*

The PT consisted of "200 multiple-choice questions in single best answer format". Items within each test were classified and matched to the blueprint. The most important component of a CAT is an Item Response Theory (IRT)-based unidimensional and calibrated item bank (Abberger, 2013; Wright & Bell, 1984). To develop an item bank, all 200 items in the chosen PT were considered as candidate items. The first stage of the study, an IRT model, was constructed while examination of the psychometric characteristics of the item bank.

2.1.2. *IRT model selection*

One of the main models of IRT, the Rasch model, produces two different estimates; "the latent trait person estimates" which are independent of the population distribution, and the "item difficulty estimates" which are independent of the person's ability (Andrich, 1988). Examination of the psychometric properties of the item bank and estimation of item parameters, the Rasch model for dichotomous items was performed using the RUMM 2020 software (Andrich, 2003).

To determine the observed data to fit in the Rasch model, numerous statistical processes are used. In this respect, the Rasch analysis included the following steps in this study:

- Distractor analysis
- Unidimensionality and local independence
- Item-model fit
- Invariance of item parameters

- Differential item functioning (DIF)
- Internal consistency reliability of the item bank (Person Separation Index-[PSI] and Kuder-Richardson 20 [KR-20])

2.1.3. *Distractor Analysis*

A Distractor Analysis examines the distractor curves' trend consistent with the Item Characteristics Curve (ICC). As the ability level of the student increases, the correct distractor should follow the general shape of the ICC. Students with higher ability level will likely choose the correct distractor, while the probability value of the other distractors will decrease. Prior to item calibration, item analysis was performed to identify potential item problems by Rasch for dichotomous data (Andrich, 2003).

2.1.4. *Unidimensionality and local independence*

Items to be entered into the item bank were also required to meet unidimensionality and local independence assumptions. Unidimensionality means that all items the test is composed of measure only a single construct.

In this study Principle Component Analysis (PCA) of residuals obtained from the Rasch model was used to examine the unidimensionality assumption. In PCA analysis, if there is no meaningful pattern in the residuals, then it is concluded that the unidimensionality assumption is met. PCA of residuals comprised an item residual correlation matrix. Through this matrix, the correlation between the items and the first residual factor are examined to identify two subsets of items (the positively and negatively correlated items). Difference between the each person estimate obtained from these positively and negatively correlated item sets is compared by independent t-tests. To meet the unidimensionality assumption, the percentage of tests outside the range ±1.96 should not to exceed 5% of total number of tests (Elhan, 2010; Pallant & Tennant, 2007; Tennant & Pallant, 2006).

2.1.5. *Test of fit to the model*

"Rasch item fit" statistics showed how accurately the test data fit the Rasch measurement model (Linacre, 2000). Overall quality of fit for Rasch models was measured regarding the following:

- Overall item fit statistic
- Overall person fit statistics
- Item-trait interaction

Overall item and overall person fit statistics transformed to a z-score. If the items and person data meet the model expectation, it is anticipate that the mean will be approximately zero and the standard deviation will be one.

Other fit statistic which is applies in this study is an item-trait interaction statistic. This statistic is reported as a chi-square and showed the characteristic of invariance across the trait. A non-significant chi-square indicates that the hierarchical ordering of the items do not vary across the trait, denoting the requirement of invariance is met.

Further to these overall summary fit statistics, individual item fit statistics and person fit statistics were applied by using residuals and chi-square statistics. In the individual item fit statistics and person fit statistics that are based on the standardized residuals, the computed residual value of "z" should range between ±2.5, indicating a satisfactory fit to the model. Consequently, the tests of the individual item/person fit were also conducted based on chi-squares. For a given item/person, several chi-squares are computed, and then these chi-square values are totaled to give the overall chi-square for the item. If the p value calculated from the overall chi-square is less than 0.05 (or Bonferroni adjusted value), then the item is considered unfit for the model (Öztuna, 2008; Pallant & Tennant, 2007; Tennant & Conaghan, 2007). Bonferroni corrections were implemented to fit statistics (Bland & Altman,1995).

2.1.6. *Differential item functioning*

To be entered into the item bank, items were required to be free of differential item functioning (DIF). The model fit is affected by item bias. DIF appears when different groups score differently on a specific item, given the same location value of the latent trait (Andrich & Hagquist, 2012; Hambleton, 1991; Teresi, 2000). In this study, a variance-based statistical analysis was performed to test DIF and artificial DIF by grades by using RUMM 2020 software (Andrich, 2003).

2.1.7. *Reliability and content validity of item bank*

Reliability was studied with the Person Separation Index (PSI) and Kuder-Richardson Formula 20 (KR-20). PSI indicated whether the test discriminates students into groups according to their ability, and a PSI of 0.7 or more evidence a fit with the Rasch model KR-20 ranged between 0 and 1, where the value of 1 indicated perfect reproducibility of person placements (Fisher, 1992; Nunnally & Bernstein, 1994; Tavakol & Dennick, 2012). Experts from different medical specialties and measurement and evaluation experts examined the content validity of the item bank.

2.1.8. *Simulation study to investigate the performance of CAT application and agreement between Rasch and simulated CAT derived estimations Computerized Adaptive Testing (CAT)*

After developing the calibrated item bank, the next stage of simulated CAT application was carried out. In CAT application, a set of questions was administered to each student according to their ability level by using a computer package program. For this purpose, the questions in the item bank with the median difficulty level were administered, and the program estimated the students' ability level (θ_{CAT}) and its standard error. After this estimation, the next most appropriate item was selected that maximized the information for θ estimate, and then the program re-estimated the students' ability level (θ_{CAT}) and its standard error. The CAT application program selected the questions for each student, according to his or her individual performance during the test. If the predefined stopping rule was fulfilled, the assessment was finished; if it was not fulfilled, the standard error of the given item administered and the ability level were re-estimated until the stopping rule was met (Bjorner, 2007; Wainer, 2001).

2.1.9. *Simulated CAT applications*

In this study item parameters obtained from the Rasch analysis were used to derive responses of 1000 students/simulee showing two different normal distributions with N(0:1) (mean=0, standard deviation=1) and N(0:3) (0 mean=0, standard deviation=3) by the RUMMss simulation software. These data were simulated to meet Rasch model expectations (Marais & Andrich, 2007). Students' responses generated by the simulation program were used to estimate student ability level using all the items (θ_{RM}), while student ability levels (θ_{CAT}) were estimated by CAT application using the SmartCAT module (Öztuna, 2008; 2012).

Selection of the first question: The question with the median difficulty level in the item bank

- Ability level (θ) estimation: Expectation a Posteriori (EAP) (Wang &Vispoel, 1998)
- Item selection: Maximum Likelihood Weight Information
- Stopping rule: Different standard errors levels (0.50, 0.40 and 0.30)

Estimations from the simulated CAT application (θ_{CAT}) were compared with the ability levels obtained from the Rasch analysis based on all items (θ_{RM}) in the item bank. In this procedure, Spearman correlation coefficient, Bland-Altman limits of agreement (Bland & Altman, 1986; 1999), and Interclass Correlation Coefficient (ICC) statistics were used (Shrout & Fleiss, 1979).

3. RESULTS

A total of 1206 students of AUSM answered 200 items of PT. Distractor analysis was conducted on these 200 items and because of the problems such as “too obvious correct answer” and “discordance pattern of correct answer with ICC”, 33 items were discarded from the item bank (Figure 1). Analyses were performed on the remaining 167 items.

Table 1. Fit of "general medical knowledge” item bank to Rasch model (after rescoring)

Item No	B	SE	Individual Item Fit Residual	X^2	df	p
1 (i1)	0.450	0.059	2.712	4.047	9	0.908
2 (i2)	0.666	0.059	1.835	5.951	9	0.745
3 (i)	0.161	0.060	-1.078	9.683	9	0.377
4 (i5)	-0.124	0.062	1.751	10.036	9	0.348
5 (i7)	-0.286	0.064	-0.627	5.949	9	0.745
6 (m8)	-0.146	0.063	0.531	3.161	9	0.958
7 (i9)	1.242	0.062	0.439	7.917	9	0.543
8 (i10)	0.988	0.060	1.027	14.087	9	0.119
9 (i12)	0.883	0.059	1.377	7.983	9	0.536
10 (i13)	1.047	0.060	1.967	10.128	9	0.340
11 (m15)	-0.271	0.064	-1.685	12.958	9	0.165
12 (m19)	1.079	0.060	-0.913	10.384	9	0.320
13 (i21)	-2.769	0.154	-0.364	11.385	9	0.250
14 (i22)	0.809	0.059	1.266	4.721	9	0.858
15 (i24)	-0.349	0.065	1.223	15.112	9	0.088
16 (i25)	-0.547	0.068	1.118	16.137	9	0.064
17 (i26)	-1.334	0.085	-0.578	7.916	9	0.543
18 (i30)	-0.673	0.070	0.921	14.898	9	0.094
19 (i31)	0.201	0.060	2.894	10.344	9	0.323
20 (i33)	0.039	0.061	-0.207	10.740	9	0.294
21 (i35)	0.249	0.060	3.432	20.162	9	0.017
22 (i36)	0.944	0.060	2.187	10.416	9	0.318
23 (m37)	-0.281	0.064	2.182	10.786	9	0.291
24 (i39)	0.471	0.059	-1.055	6.963	9	0.641
25 (i41)	-0.221	0.063	0.363	9.704	9	0.375
26 (i43)	-0.747	0.071	0.290	5.509	9	0.788
27 (i44)	0.395	0.059	2.411	13.970	9	0.123
28 (i45)	-1.465	0.089	0.094	9.199	9	0.419
29 (i46)	0.569	0.059	1.739	7.000	9	0.637
30 (i48)	-0.194	0.063	0.568	8.184	9	0.516
31 (i50)	-0.686	0.070	-0.082	6.296	9	0.710
32 (i52)	0.015	0.061	1.565	11.500	9	0.243
33 (i54)	0.450	0.059	0.222	17.966	9	0.036
34 (i55)	-1.791	0.101	-0.563	11.493	9	0.243
35 (i56)	-0.389	0.065	-0.717	8.558	9	0.479
36 (i58)	0.050	0.061	2.614	16.004	9	0.067
37 (i59)	-0.049	0.062	0.130	6.533	9	0.686
38 (i60)	-0.997	0.076	-1.006	8.024	9	0.532
39 (i61)	-1.450	0.089	-0.344	8.329	9	0.501
40 (i62)	-0.279	0.064	-1.013	12.447	9	0.189
41 (i63)	-1.157	0.080	0.307	16.995	9	0.049

Table 1. Continues

42 (i65)	-2.064	0.113	0.238	9.990	9	0.351
43 (i69)	-0.405	0.066	0.030	7.078	9	0.629
44 (i70)	0.257	0.060	-0.040	11.881	9	0.220
45 (i73)	0.204	0.060	2.389	6.662	9	0.672
46 (i76)	1.123	0.061	1.872	10.796	9	0.290
47 (i79)	-0.311	0.064	1.186	5.011	9	0.833
48 (i87)	-0.274	0.064	-0.481	9.212	9	0.418
49 (i88)	-0.940	0.075	-0.706	9.410	9	0.400
50 (i89)	1.415	0.063	0.721	12.853	9	0.169
51 (m92)	-0.150	0.063	-0.029	10.393	9	0.320
52 (m102)	0.167	0.060	-1.330	10.783	9	0.291
53 (i103)	0.725	0.059	-0.572	14.471	9	0.107
54 (i104)	0.283	0.060	-1.840	12.909	9	0.167
55 (i105)	-1.646	0.096	-0.916	18.944	9	0.026
56 (i106)	1.431	0.064	0.055	4.844	9	0.848
57 (i113)	-0.228	0.063	-0.738	11.141	9	0.266
58 (i114)	0.801	0.059	0.885	5.515	9	0.787
59 (i115)	0.167	0.060	-0.269	6.465	9	0.693
60 (i116)	1.131	0.061	1.534	13.360	9	0.147
61 (i117)	0.198	0.060	1.845	7.156	9	0.621
62 (i119)	0.502	0.059	-1.374	17.268	9	0.045
63 (i121)	0.358	0.059	0.590	5.691	9	0.770
64 (i122)	0.322	0.059	-1.923	18.530	9	0.030
65 (i123)	-1.127	0.080	-1.039	12.310	9	0.196
66 (i124)	-1.177	0.081	-1.308	16.593	9	0.055
67 (i125)	-0.296	0.064	1.033	9.038	9	0.434
68 (i128)	-0.269	0.064	-1.051	10.836	9	0.287
69 (i129)	0.616	0.059	2.739	12.926	9	0.166
70 (i130)	0.561	0.059	0.957	11.488	9	0.244
71 (i131)	-0.765	0.072	-1.099	9.396	9	0.402
72 (i133)	0.743	0.059	-1.062	10.213	9	0.334
73 (i139)	0.191	0.060	1.022	16.195	9	0.063
74 (i140)	-1.369	0.086	-0.807	9.810	9	0.366
75 (i141)	0.879	0.059	1.681	8.408	9	0.494
76 (i142)	1.127	0.061	1.700	7.514	9	0.584
77 (i145)	1.299	0.062	2.395	15.809	9	0.071
78 (i148)	0.687	0.059	0.771	4.412	9	0.882
79 (i151)	-0.695	0.070	-0.107	5.019	9	0.833
80 (m155)	-0.103	0.062	0.011	5.553	9	0.784
81 (i156)	-0.649	0.069	-0.254	14.803	9	0.096
82 (i157)	-0.116	0.062	-0.003	5.754	9	0.764
83 (i158)	0.139	0.060	0.490	10.426	9	0.317
84 (i160)	0.064	0.061	0.884	6.378	9	0.702
85 (i161)	0.718	0.059	2.828	18.111	9	0.034
86 (i165)	-0.278	0.064	0.167	7.022	9	0.635
87 (i166)	0.718	0.059	0.650	13.473	9	0.142
88 (m171)	0.180	0.060	0.138	11.755	9	0.227
89 (i172)	-0.672	0.070	0.355	8.026	9	0.532
90 (i177)	0.329	0.059	1.157	6.059	9	0.734

Table 1. Continues

91 (i178)	0.494	0.059	0.572	6.733	9	0.665
92 (i180)	0.209	0.060	2.800	11.833	9	0.223
93 (i181)	0.653	0.059	-0.988	6.108	9	0.729
94 (i182)	0.104	0.061	0.932	8.226	9	0.512
95 (m183)	0.376	0.059	-1.006	10.950	9	0.279
96 (i185)	1.341	0.063	0.137	14.669	9	0.100
97 (i188)	0.159	0.060	-0.433	10.147	9	0.339
98 (i191)	-0.068	0.062	0.280	2.790	9	0.972
99 (i192)	0.173	0.060	1.213	11.649	9	0.234
100 (i193)	-0.370	0.065	-0.896	14.136	9	0.118
101 (i194)	-1.185	0.081	0.159	6.196	9	0.720
102 (i195)	-0.888	0.074	-0.318	7.108	9	0.626
103 (i200)	0.696	0.059	-1.364	11.517	9	0.242

B: Item Difficulty, SE: Standard Error, df= Degrees of Freedom

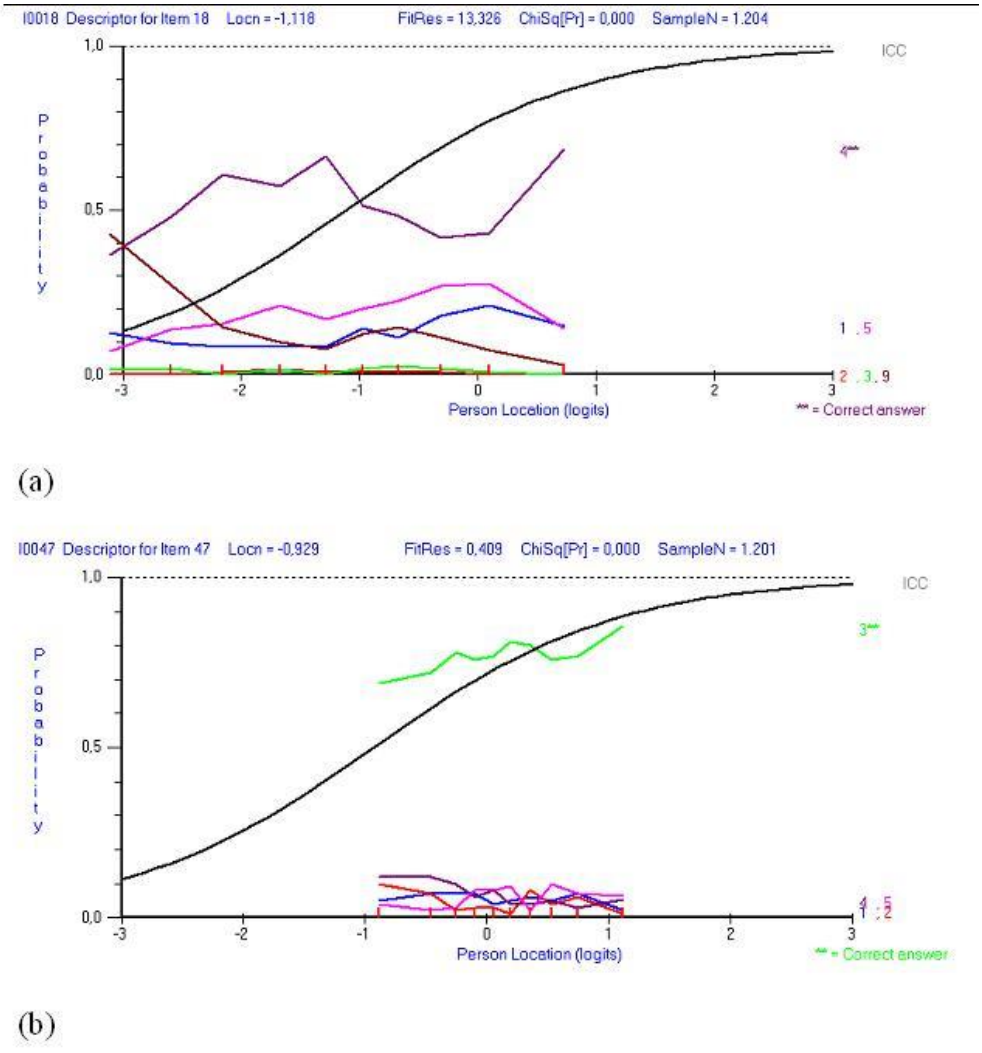

Figure 1. Item analysis og item 18 (a) and 47 (b)

3.1. Development of item bank (internal construct validity)

In order to be entered into the item bank, items were required to satisfy Rasch model expectations, including being free of DIF and having unidimensionality and local independence. Sixty-four of 167 items were omitted as not fitting the Rasch model. For the remaining 103 items, p values from chi-square were less than Bonferroni adjustment fit level of 0,0005 (0.05/103=0,0005). In addition, the standard residual values were within the ±2.5 range. These statistics indicated that the items fit the Rasch model (Table 1).

When the overall fit statistics were tested, it was found that the overall mean item fit residual was 0.402 (SD 1.234) and the overall mean person fit residual was 0.008 (SD 0.893). Since both values met this expectation, this indicated that the items and persons fit the model. The "item-trait interaction" statistic reported that the chi-squared value was non-significant [chi-squared = 1049.33 (0.003); given a Bonferroni adjustment fit level of 0,0005)], meaning that the hierarchical ordering of items was invariant across the trait. DIF was tested for academic grades of students, and it was found that all the items were DIF-free.

When the unidimensionality of the 103-item item bank was examined by PCA, there was no pattern violating this assumption (t=%4.6; CI %3.4-%5.7). When the assumption of local independence was tested, there was no pair of items that had a residual correlation of 0.30 or more. For the person-item threshold distribution, person and item locations were logarithmically transformed and plotted on the same continuum.

Figure 2 shows person and item locations on the x-axis. Figure 2 also demonstrates that the item bank was well targeted with the mean of the persons at 0.597 on the logit scale, and few people were outside of the operational range of the scale. As seen in the graphic, the person distribution (top of the figure) was well matched by the item distribution (bottom of the figure).

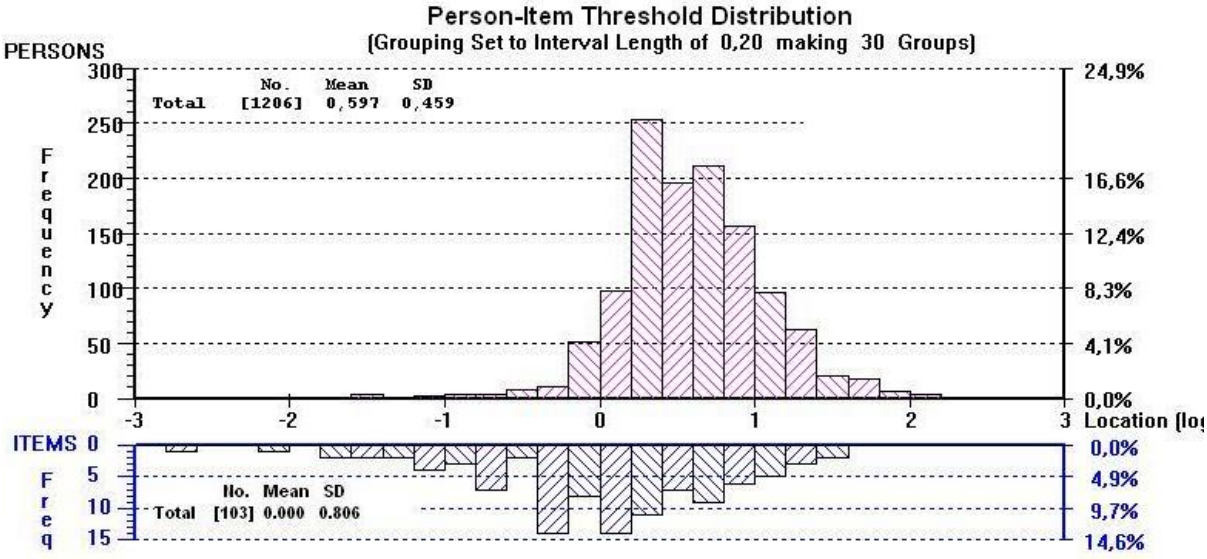

Figure 2. Person-item threshold distribution (103 items)

PSI and KR-20 of the item bank were computed as 0.77. Since the threshold of acceptance for PSI is 0.7, the computed value indicated that it is possible to statistically differentiate between two groups of respondents. This result showed that the items effectively separated the persons.

3.2. Content validity of item bank

When experts from several medical specialties and measurement and evaluation experts examined the content validity of the item bank, they concluded that the item bank contained enough questions for a representative and balanced sampling of the prescribed blueprint.

3.3. Simulation analysis

In this study, simulated CAT application was conducted to evaluate the agreement between θ_{CAT} and θ_{RM}.

3.4. Simulated CAT application [N (0,1)]

The number of items used in CAT, which was carried out with responses derived from 1000 individuals from the distribution of mean with 0 and variance with 1, ranges between 11 and 45 for various standard error levels as shown in Table 2.

Table 2. Descriptive Statistics for Number of Items Administered in CAT application and Correlation and Agreement between θ_{RM} *and* θ_{CAT}*estimations [N (0,1)]*

Stopping rule	Mean number of items (±SD) [Median (min-max)] usedin CAT	r	ICC (95% CI)
Standard Error: 0.30 Reliability: 0.90	45 (±3) [44 (43-50)]	0.975**	0.989 [0.988-0.990]
Standard Error:0.40 Reliability: 0.84	24 (±3) [23 (23-50)]	0.940**	0.971 [0.967-0.974]
Standard Error: 0.50 Reliability: 0.75	14 (±0.8) [14 (13-21)]	0.886**	0.941 [0.933-0.948]
Standard Error:0.548 Reliability: 0.70	11 (±0.53) [11 (11-16)]	0.868**	0.928 [0.919-0.937]

SD: Standard Deviation, r: Correlation Coefficient, ICC: Intraclass Correlation Coefficient., CI: Confidence Interval, **: p<0.001

In CAT applications, an estimation with a reliability of 0.75 can be made using 14 questions (reduced by % 86,4). When compared to paper and pencil tests (based on all items in the bank), CAT resulted in a 56.3-88.5% decrease in the number of items.

The research findings illustrated that there is a high (r=0.868-0.975 and ICC=0.928-0.989, respectively) correlation and agreement (for standard error 0.3, 0.4, 0.5, 0.548) between θ estimations obtained from paper-and-pencil (θ_{RM}) and CAT applications (θ_{CAT}) and these findings are statistically significant. Ninety-five percent ranges of agreement between θ_{CAT} and θ_{RM} according to the Bland-Altman approach were -0.39 to 0.39, -0.64 to 0.65, -0.83 to 0.83, and -0.92 to 0.96 when the stopping rule was set to standard errors of 0.30, 0.40, 0.50, and 0.548, respectively. In addition, 942 of 1000, 953 of 1000, 960 of 1000, and 935 of 1000 converged estimates were within the 94-96% agreement limits for different standard error, respectively (Figure 3).

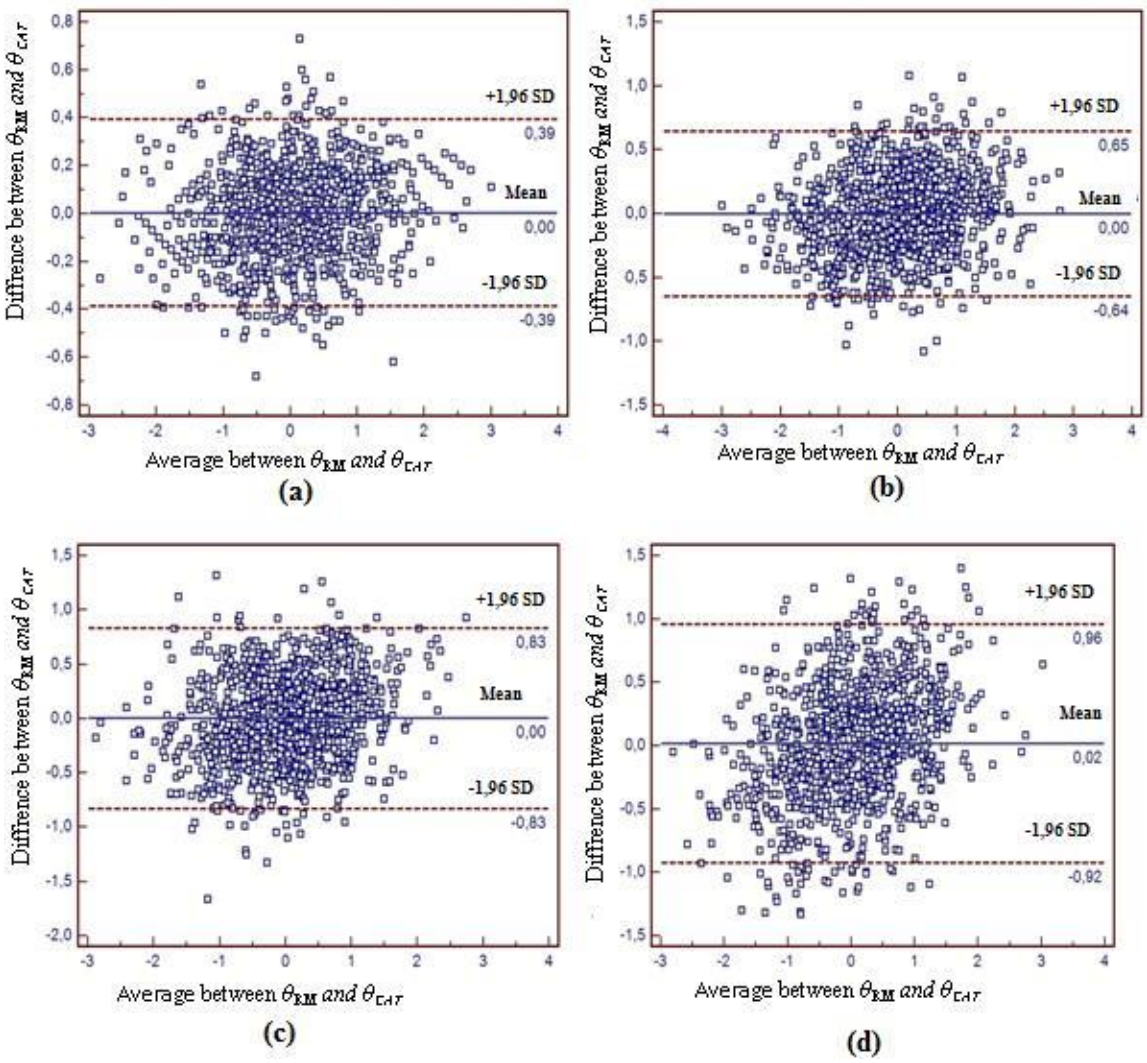

Figure 3. Bland-Altman Plot for [N (0,1)] (a) SE=0.30, (b) SE=0.40, (c) SE=0.50, (d) SE=0.54

3.5. Simulated CAT application [N (0,3)]

The number of items used in CAT, which was carried out with responses derived from 1000 individuals from the distribution of mean with 0 and variance with 3, ranges between 12 and 75 for various standard error levels (Table 3).

Table 3. Descriptive Statistics for Number of Items Administered in CAT application and Correlation and Agreement between θ_{RM} *and* θ_{CAT}*estimations* [N (0,3)]

Stopping rule	Mean number of items (±SD) [median (min-max)] usedin CAT	r	ICC (95% CI)
Standard Error: 0.30 Reliability: 0.90	75 (±27) [76 (42-103)]	0.998**	0.999 [0.999-0.999]
Standard Error: 0.40 Reliability: 0.84	35 (±15) [27 (22-68)]	0.992**	0.995 [0.995-0.996]
Standard Error: 0.50 Reliability: 0.75	17 (±3) [15 (15-23)]	0.984**	0.986 [0.984-0.987]

In CAT applications, an estimation with 0.75 reliability can be made using 17 questions (reduced by %83.4). When compared to paper and pencil tests, CAT amounted to a reduction in the number of items administered by 27.6-88.3%.

The research findings illustrated that there is a high (r=0.984-0.998 and ICC 0.928-0.989, respectively) correlation and agreement (for standard error values 0.3, 0.4, 0.5, 0.548) between θ estimations obtained from paper-and-pencil (θ_{RM}) and CAT applications (θ_{CAT}) and these findings are statistically significant.

Ninety-five percent ranges of agreement between θ_{CAT} and θ_{RM} according to Bland-Altman approach were -0.27 to 0.28, -0.57 to 0.56, -0.96 to 0.94, and -1.21 to 1.28 when the stopping rule was set to standard error of 0.30, 0.40, 0.50, and 0.548, respectively. In addition, 921 of 1000, 948 of 1000, 973 of 1000, and 978 of 1000 converged estimates were also within the 94-96% agreement limits for different standard error, respectively (Figure 4).

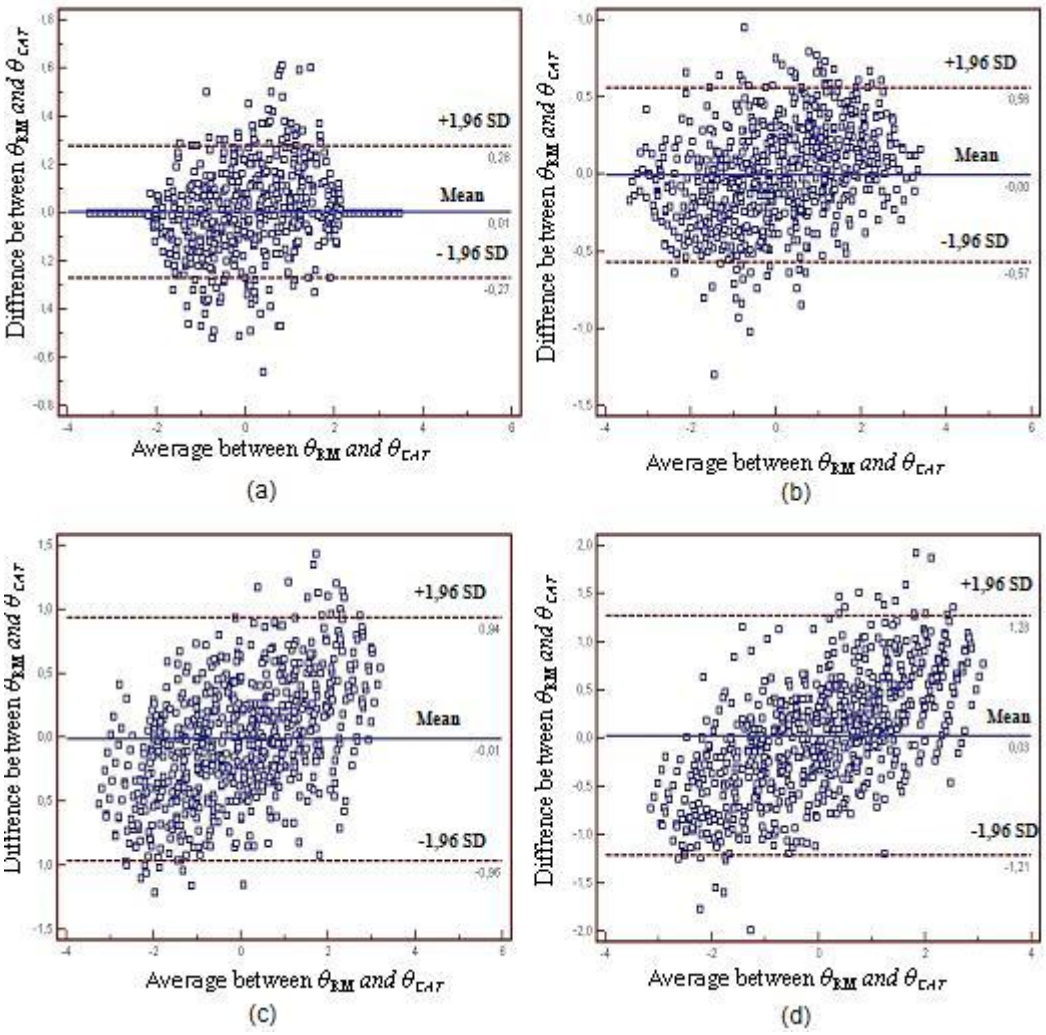

Figure 4. Bland-Altman Plot for [N (0,3)] (a) SE=0.30, (b) SE=0.40, (c) SE=0.50, (d) SE=0.54

4. DISCUSSION and CONCLUSION

This study aimed to demonstrate whether the Rasch model is an alternative to Classical Test Theory, in order to improve the PT in medical education. This study will provide the initiative to investigate the potential for applying CAT in PT.

PT has some disadvantages; firstly, the bulkiness may cause demotivation, boredom, and tiredness. Furthermore, if the questions are too difficult, it could discourage students, while

questions that are too easy could be uninteresting for students. In addition, items that are too easy or too difficult do not give enough information about students' ability. Previous research for CAT usage demonstrated that CAT reduces these disadvantages by shortening test length and providing a flexible test format. CAT application offers students a set of questions that are matched to their ability levels, thus providing the examinees with individualized examinations.

The present findings showed that 103 items formed a unidimensional item bank. The ability of the students estimated by CAT was highly correlated with those of the full item set. The average number of items needed to estimate ability was only 13.6% of the full item set of PT/ paper-pencil based PT (for ([N(0,1) and 0.75 reliability], and 14,6 % of the full item set PT (for ([N(0,3) and 0.75 reliability]. θ_{CAT} and θ_{RM} correlated and agreed well for both populations ([N(0,1) and ([N(0,3)]). This study demonstrated that the test length could be shortened without decreasing reliability.

The follow-up of such an approach raises developmental challenges. In this study, the number of items included in the item bank was less than the average for CAT standards. For CAT application, item banks should contain a large number of items considered important to work on the item bank by the CAT designer. Previous studies have been based on item numbers ranging from less than 100 to several thousand items. In this study, however, the number of items was relatively low, with student ability and item difficulty distribution mirroring each other (as shown in Figure 1), and items distributed across the range of the trait being measured. For this reason, although the item bank for this study was relatively small, the results suggested that CAT worked well.

This study intended to discuss the steps required to build an item bank for PT. At the end of this process, a unidimensional item bank that represented "general medical knowledge" was developed for CAT application. However, PT builds on a blueprint that includes specific subdomains (such as the cardiovascular system or the discipline of anatomy) as a part of the overall domain. Extensive feedback and patterns of knowledge growth within specific subdomains could be provided to students and other stakeholders in addition to overall knowledge build-up. To provide detailed feedback and knowledge growth patterns, the items could be divided into unidimensional subsets, and several item banks could be constructed as a possible solution for PT CAT application. In addition, multidimensional CAT procedures based on multidimensional IRT (MIRT) might be another solution for PT CAT. As a result, a reduction of over 80% of the items in CAT format of PT could test its potential to follow candidates' progress practically through educational programs.

CAT application's utility for medical course assessment has been demonstrated in this study. To the authors' knowledge, there have been no other studies about the CAT application in PT for medical school. It should be emphasized that the purpose of this study was not to investigate whether PT CAT based on 103 items should replace the current paper-pencil based PT. This study aimed to discuss the steps to build an item bank and illustrated the utility of CAT implementation as an example.

This study showed that it is possible to develop an appropriate item bank for the PT using the Rasch model, and that the difficulty administering large number of items in PT can be reduced by CAT application. The results of this study will encourage the implementation of CAT in medicine and in other disciplines.

Acknowledgements

This article is extracted from doctorate dissertation entitled "Developing an Item Bank for Progress Test and Application of A Computerızed Adaptive Testing by Simulation in Medical Education" (Advisor: Prof. Dr Nizamettin Koç, Co-advisor Atilla Halil Elhan PhD Dissertation, Ankara University Institute of Educational Sciences, Ankara/Turkey, 2013)

This study has previously been presented in *International 8th Statistics Congress (ISC2013), Antalya, 2013.*

Conflicts of Interest

The authors report no conflicts of interest. The authors alone areresponsible for the content and writing of this article.

ORCID

Aysen Melek AYTUG KOSAN https://orcid.org/0000-0001-5298-2032
Nizamettin KOÇ https://orcid.org/0000-0002-3308-7849
Atilla Halil ELHAN https://orcid.org/0000-0003-3324-248X
Derya ÖZTUNA https://orcid.org/0000-0001-6266-3035

5. REFERENCES

Abberger, B., Haschke, A., Wirtz, M., Kroehne, U., Bengel, J., & Baumeister, H. (2013). Development and evaluation of a computer adaptive test to assess anxiety in cardiovascular rehabilitation patients. *Archives of Physical Medicine and Rehabilitation, 94*(12), 2433-2439. Doi: 10.1016/j.apmr.2013.07.009

Andrich, D. (1988). Rasch models for measurement. The USA: Sage Publications Inc.

Andrich D, Lyne A, Sheridan B, Luo G. RUMM2020. Perth: RUMM Laboratory Pty Ltd. 2003 Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010). Progress testing internationally. *Medical Teacher, 32*(6), 451-455. Doi: 10.3109/0142159X.2010.485231

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37*(3), 387-416. Doi: 10.3102/1076998611411913

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ, 310*(6973), 170. Doi: 10.1136/bmj.310.6973.170

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310. Doi: 10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160. Doi: 10.1177/096228029900800204

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16*(1), 95-108. Doi: 10.1007/s11136-007-9168-6

Elhan A. H., Küçükdeveci A. A., & Tennant A. (2010). The rasch measurement model. Franchignoni F. (Ed.) *Research issues in Physical & Rehabilitation Medicine*. Advances in Rehabilitation. Maugeri Foundation 19, 89-102

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238.

Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010) Progress testing internationally. *Medical Teacher*. 2010. *32*(6), 451-455. Doi: 10.3109/0142159X.2010.485231

Hambleton, R. K. (1991). *Fundamentals of item response theory*. The USA: Sage publications.

Linacre, J. M. (2000). Computer adaptive testing: A methodology whose time has come. Chae, S.-Kang, U. Jeon, E. Linacre, JM (eds.): *Development of Computerised Middle School Achievement Tests*, MESA Research Memorandum.

Marais, I., & Andrich, D. (2007). *RUMMss. Rasch unidimensional measurement models simulation studies software.* The University of Western Australia, Perth.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Öztuna, D. (2008). Kas iskelet sistem sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması (Implementing computer adaptive testing

method to estimate disability levels in musculoskeletal system disorders). (Doctoral Dissertation). Ankara Üniversitesi Sağlık Bilimleri Enstitüsü. Ankara

Öztuna D. (2012). *A computerized adaptive testing software (CAT): SmartCAT*. European Rasch Training Group (ERTG) Meeting, 17-19 April 2012, Leeds, UK.

Pallant, J. F., & Tennant, A. (2007). An introduction to the rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1-18. Doi: 10.1348/014466506X96931

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420. Doi: 10.1037/0033-2909.86.2.420

Tavakol, M., & Dennick, R. (2012). Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Medical Teacher, 34*(3), e161-e175. Doi: 10.3109/0142159X.2012.651178

Tennant, A., & Pallant, J. (2006). Unidimensionality matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048-1051.

Tennant, A., & Conaghan, P. G. (2007). The rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research, 57*(8), 1358-1362. Doi: 10.1002/art.23108

Teresi, J. A., Kleinman, M., & Ocepek‐Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine, 19*(11‐12), 1651-1683. Doi: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651: AID-SIM453>3.0.CO;2-H

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., & Steinberg, L. (2001). Computerized adaptive testing: A primer. *Qual Life Res, 10*, 733-734. Doi: 10.1023/A:1016834001219

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331-345. Doi: 10.1111/j.1745-3984.1984.tb01038.x

Wrigley, W., Van Der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. Medical Teacher, 34(9), 683-697. Doi: 10.3109/0142159X.2012.704437