Estimation of science topic area scores for Iranian eight grader in TIMSS 2007

Masoud Kabiri

Research Institute for Education, Iran

Abstract: Results of TIMSS were published in two levels- average scale score (math or science scores) and domains score (cognitive or content domains score). While knowledge about domain scores is useful, more detailed information needs about students' performance. In present article, the focus is on estimation of topic area science scores for Iranian eight graders in TIMSS 2007. Based on items in each topic areas, 12 estimations were done by applying Plausible Values methodology. Results showed that compared to average science scale score of Iran, some topic areas such as "forces and motion"; "physical states, changes in matter, electricity and magnetism"; and "earth processes, cycles and history" had been achieved better than others. In other side, "light and sound"; "cells, their functions and human health"; and "ecosystems" had not been achieved well. Furthermore, some topic areas performed better in higher benchmarks whereas others are good in lower benchmarks, considering relatively same averages. In addition, although private schools are significantly better than public schools; but, higher averages of girls rather than boys were not significant in all topic areas. Finding suggested that more detailed knowledge about performance of topic areas or even topics help to give diagnostic information for curriculum developers and policy makers.

Keywords: science education, plausible values, diagnostic information

Introduction

TIMSS provides valuable information about educational system in a comparative perspective. Performance of students in international assessment is an important indicator about what students know and can be able to do. Reasonably, information of assessment should came back to modify educational systems. Providing feedbacks from performance of students help especially in curriculum developing. Therefore, TIMSS helps policy makers and curriculum developers in order to detect educational defects, but more feedbacks needed regarding the information.

There are increasing demands to access more detailed information from large scale assessment. The information likely detects students' shortages in learning and leads to proper learning.

Although, TIMSS shows many of Strengths and weaknesses of educational systems in a comparative perspective, limited diagnostic information are provided to change curriculum. In other words, curriculum developers do not aware what to do about the results. That is, estimation of TIMSS is limited to a total average scale score (for example, science average scale score) and several cognitive and content domain scores. No estimation is specified in topic and/or topic area. It is assumed that estimation in more detailed levels let opportunities to provide this kind of information about students' achievement.

Having diagnostic information is important demands of curriculum developers and policy makers. So far, many diagnostic approaches are proposed in educational assessment. They located in cognitive diagnostic assessment (Leighton & Gierl, 2011; Rupp, Templin, & Henson, 2010). In present paper, plausible values are used to achieve detailed information about topic areas. Applying plausible values draw such needs and inspect achievement of students, regarding intended curriculum.

In large scale assessments, more extended performances are expected to assess, so large numbers of items are needed. Due to students' limitations to ask whole items, matrix sampling is used and each student asks a part of items. According to this consideration, an analysis technique should be applied to combine data from all students in order to estimate of population parameters. Plausible values adapt to such situation which proposed by Mislevy and Sheehan (1989) based on imputation theory of Rabin (1987). In this technique, performance of students are treated as a missing value and are applied in limited subset of items and background questionnaires, in conjunction with measurement model of Rabin' multiple imputation to generate ability distribution of students in population (Rutkowski, Gonzalez, Joncas, & von Davier, 2010).

Plausible values have several aspects. First, because the true values are unknown, and because we do not have accurate estimates of individual performance on short tests, plausible values are a very useful tool for generating values that have more accurate statistical properties than do observed scores for subgroup comparisons (Von Davier, Gonzalez, & Mislevy, 2009). Second aspect of plausible values relates to adequateness for determine parameters of population. In other words, when the purpose of assessment is to describe populations rather than to measure individuals assigning student scale scores and then analyzing scale scores to estimate population characteristics does not provide correct results (Monseur & Adams, 2009; Wu, 2005), and consequently, aggregations of individual student scores can lead to seriously biased estimates of population characteristics (Foy, Galia and Li, 2008). So, instead of calibration of individual abilities and aggregation of them to estimate parameters of population, response information and background data are used to estimate parameters of population directly. Therefore, goal of plausible values is not to assign a mark,

grade, or score to individual students, but to describe a group of students in terms of their long-term expected performance and the variability of them.

The relatively small number of items per block and the relatively small number of blocks per test booklet mean that the accuracy of measurement at the individual level of these assessments is considerably lower than the level of accuracy common for individual tests used for diagnosis or admission purposes (von Davier, Gonzalea and Mislevy, 2009).

Some advantages for applying of plausible values are specified. Plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated (Foy, Galia, & Li, 2008). Also, there is variation among five generated plausible values for each student which is assignable to ambiguity of students' proficiencies estimation. This component of uncertainty adds to sample error to generate unbiased standard error. in other words, if the measurement at the individual level contains error, then this error should be taken into account in the computation of population statistics and their standard errors (Wu, 2005) In other hand, rather large number of items to achieved satisfaction convergence is a disadvantage of plausible values(Adams & Carstensen, 2002; Monseur & Adams, 2009).

Present article is designed to estimate topic area' scores of science for eight grade Iranian students. So, comparison between students' achievement in science topic areas is main purpose of the study. Also, an investigation percentage of students reaching to international benchmarks is another purpose of article. Furthermore, gender and school type comparisons are last consideration of the study.

Method

Sample

National school sample in TIMSS 2007 was 220 schools. 12 schools withdrew because there were closed at the date of data gathering. 4046 students were listed in 208 reminded schools, but 65 absent students were in sample, consequently, 3981 students were last sample in grade eight. Therefore, Iranian students in grade eight who participated in TIMSS 2007 contained sample of present study.

Coverage rate in Iran was 100%. 0.5% of schools in grade eight were excluded in school level. Participation rate was in school and class level 100% and in student level 98%, so total participation rate became 98%.

Measures

344 science items of TIMSS 2007 were used in order to data analysis. They consisted of 183 multiple choice and 161 created response items. There were divided to 86 items developed in 2007 cycle, 124 trend and 134 bridge items. They were assigned to 46 topics and 19 science topic areas.

Data analysis

Plausible values methodology was applied to analyze the data. 18 variables from student, teacher, and administrator questionnaires were perception of school climate, emphasis on homework, class limitations due to student factors, working conditions, perception of school safety, preparation to teach topics, teachers' major area of study, and time to teach science from science teacher file; time doing homework, attitudes towards science, valuing science, science self-concept, and perception of being safe in school from student background file; good attendance in school, availability of school resources in science education, and perception of school climate from school file; along with gender and type of school (private or public). PARSCALE, DESI and IDB Analyzer were applied to analyze data.

After estimation, linking phase related to TIMSS score was conducted according to procedure noted in Technical Report (Foy, et al., 2008). Furthermore, weighting sampling and jackknife repeated replication (JRR) techniques were used to more accurate results.

Data sources

Main data source was student achievement file of Iran (bsairnm4). In addition, in order to carry out plausible values, also some information was added from science teacher, administrator and student background files (respectively btmirnm4, bcgirnm4, and bsgirnm4).

Results

Based on allocation of items on topics, there were 5 to 25 items in each topic area and 1 to 11 items in each topic. According to acceptable standard error of TIMSS (7.5) and related studies, scores were estimated for topic areas which at least 4 items were asked by students. With regards to aligning of items in booklets, scales were estimated with more than 12 items. From 19 topic areas, 8 ones met the above criteria and contained 10 to 25 items. In addition, some similar topic areas with similar theme were combined to create common scale. Four combined topic area were "cells and their functions and human health"; "light and sound"; "physical states and changes in matter and electricity and magnetism"; "earth's structure, physical features; resources, their use, conservation; and earth in the solar system and the universe". Therefore, 12 scales were considered for objective of study.

In order to create prior distribution, 18 variables were selected from student, teacher and school background data files. There were science teachers' perception of school climate, emphasis on science homework, science class limitations on instruction due to student factors, adequate working conditions, teachers' perception of school safety, feeling of preparation to teach science topics, science teachers' major area of study, and weekly implemented times for science from science teacher file; spent time doing science homework, attitudes towards science, valuing science, science self-concept, and students' perception of being safe in school from student background file; student good attendance in school, availability of school resources in science education, and principals' perception of school climate from school file. In addition, gender of students and type of school (public/private) were added to above variables.

Table 1 showed averages and standard errors of discrimination, location and guessing parameters for all examined topic areas. The parameters were derived from item parameters in Olson et al. (2008).

Table 1. Averages and standard deviations of parameters of items in each topic areas

Content domain	scale	Discrimination		Location		Guessing	
		average	Standard deviation	average	Standard deviation	average	Standard deviation
biology	Characteristic and life process of living things	0.932	0.281	0.587	0.418	0.263	0.084
	Cells and their functions and human health±	0.909	0.229	0.492	0.625	0.23	0.043
	Life cycles, reproduction and heredity	1.122	0.535	0.715	0.443	0.183	0.014
	Diversity, adaptation, and natural selection•						
	Ecosystems	1.07	0.344	0.522	0.478	0.254	0.038
chemistry	Classification and composition of matter	1.142	0.673	0.656	0.887	0.258	0.053
	Properties of matter•						
	Chemical change	0.842	0.272	0.741	0.593	0.293	0.013
physics	Physical states and changes in matter and electricity and magnetism±	0.79	0.184	0.963	0.523	0.244	0.032
	Energy transformations, heat, and temperature	0.889	0.215	0.8	0.487	0.232	0.046
	Light and sound±	1.013	0.296	0.743	0.466	0.274	0.058
	Forces and motion	1.092	0.313	0.702	0.814	0.218	0.034
Earth science	Earth's processes, cycles, and history	1.004	0.349	0.58	0.543	0.27	0.061
	Earth's structure, physical features; resources, their use, conservation; and earth in the solar system and the universe±	0.915	0.233	0.662	0.433	0.251	0.091

± Scale is consisted of two or three topic areas.

• There is no estimation in this scale.

Table 1 showed that averages of slopes (discrimination parameters) were closed to each others. Lowest slope belonged to physical states, changes in matter, electricity and magnetism and in other side classification and composition of matter was a steepest slope. Low range of slopes (about 0.35) indicated that slope of topic areas were similar. Same situation was

observed in location of scales' items. That is, lowest value was related to cells, their functions and human health (0.492) and highest was physical states and changes in matter, electricity and magnetism (0.963). Moreover, guessing values had limited range (between 0.182 and 0.293). In general, parameters of scales' items did not differ very much so comparison between scales can be conducted.

Estimation of scores for topic areas (scales) was provided in table 2 along with number of covered items, number of asked students and Jackknife Repeated Replication technique (JRR) of standard error. In each scale, number of items and responded students presented also.

Table 2. Achievement of students in topic areas

Content domain	scale	items	N	score	JRR standard error
biology	Characteristic and life process of living things	15	282	457	4.7
	Cells and human health	20	853	432	3.8
	Life cycles, reproduction and heredity	10	273	459	6.5
	Diversity, adaptation, and natural selection	25	1992	437	3.1
chemistry	Classification and composition of matter	22	1717	452	3.1
	Chemical change	12	282	473	8.4
physics	Physical states, electricity and magnetism	17	1696	481	3.4
	Energy transformations and temperature	13	280	466	4.9
	Light and sound	11	273	427	8.6
	Forces and motion	13	289	500	9.8
Earth science	Earth's processes and history	18	860	478	5.1
	Earth's structure and resources	22	1415	461	4.4

According to table 2, forces and motion in physics domain had highest score (average scale score of 500) and physical states, their changes in matter, electricity, and magnetism (physics domain); earth's processes, cycles, and history (earth science domain); and chemical change (chemistry domain) had next ranks, respectively. In other side, the weakest topic areas

were cells, their functions and human health (biology domain) and light and sound (physics topic).

In content domain comparison perspective, it was indicated that varying scores were in physics content domain. Whereas, forces and motion; and physical states and changes in matter were two highest topic areas, but energy transformations and temperature was around national average scale score and light and sound was lowest topic area. Figure 1 showed topic areas' scores in graphical view. In the figure, it can be seen that 95% Confidence Interval for Average (±2SE) along with point of 5th, 25th, 75th and 95th Percentiles of Performance.

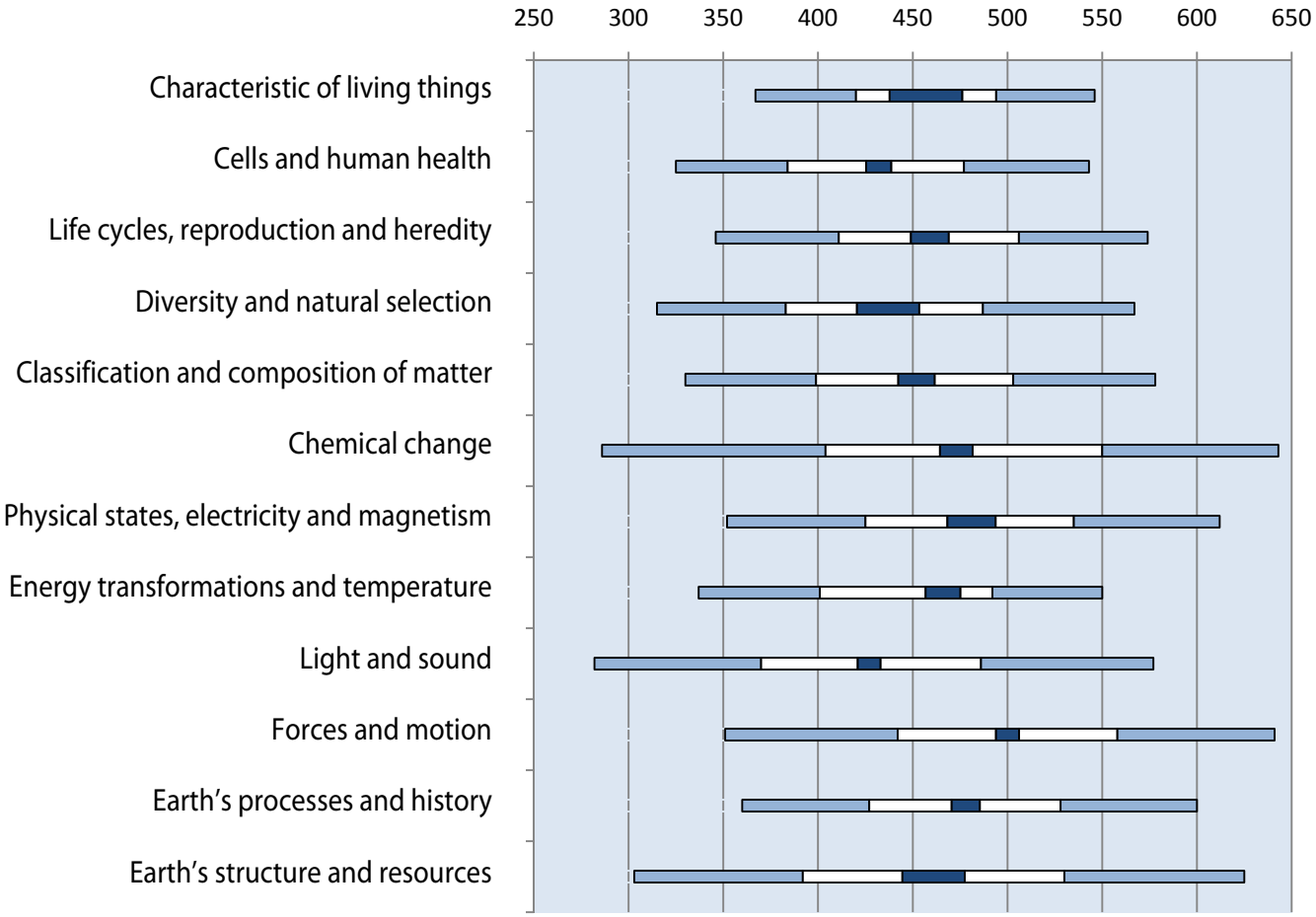

Figure 1. Distributions of science topic areas scales.

Table 3. Multiple Comparison of topic areas scale scores

scale	score	Forces and motion	Physical states	Earth's processes	Chemical change	Energy transformations	Earth's structure and use	Life cycles and heredity	Characteristic of living things	Classification of matter	Ecosystems	Cells and Human health	Light and Sound
Forces and motion	500			↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
Physical states and Electricity	481					↑	↑	↑	↑	↑	↑	↑	↑
Earth's processes and history	478	↓					↑	↑	↑	↑	↑	↑	↑
Chemical change	473	↓							↑	↑	↑	↑	↑
Energy transformations, heat, and temperature	466	↓	↓						↑	↑	↑	↑	↑
Earth's structure and use	461	↓	↓	↓							↑	↑	↑
Life cycles and heredity	459	↓	↓	↓							↑	↑	↑
Characteristic of living things	457	↓	↓	↓							↑	↑	↑
Classification and composition of matter	452	↓	↓	↓	↓	↓						↑	↑
Ecosystems	437	↓	↓	↓	↓	↓	↓	↓	↓				
Cells and Human health	432	↓	↓	↓	↓	↓	↓	↓	↓	↓			
Light and Sound	427	↓	↓	↓	↓	↓	↓	↓	↓	↓			

↑ Average achievement significantly higher than comparison topic area
↓ Average achievement significantly lower than comparison topic area

The above table revealed multiple comparisons of topic areas' scores ($p<0.05$). Some scales did not differ due to relative large standard error and/or small difference.

Comparison of achievement for some groups is another main point of present study. This comparison was conducted in type of schools and students' gender. Tables 4 and 5 stressed the results.

Table 4. The comparison of students' topic areas achievement between public and private schools

scale	Public schools		Private schools		Standard error of difference	t
	n	score	n	score		
Characteristic and life process of living things	211	455	61	495	12.461	3.255
Cells and human health	676	430	177	462	7.243	4.505
Life cycles, reproduction and heredity	213	457	60	490	14	2.413
Diversity, adaptation, and natural selection	1577	434	415	475	7.494	5.423
Classification and composition of matter	1361	449	356	496	6.314	7.379
Chemical change	221	468	61	537	19.658	3.513
Physical states, electricity and magnetism	1339	476	357	538	6.651	9.23
Energy transformations and temperature	222	444	58	479	12.647	2.8
Light and sound	213	422	60	495	17.854	4.085
Forces and motion	230	497	59	533	22.59	1.595
Earth's processes and history	679	475	181	521	8.504	5.347
Earth's structure and resources	1119	457	296	509	8.389	6.201

The results demonstrated private schools significantly better achieved than public schools in all topic areas (except in forces and motion the difference was not significant). Chemical change had largest difference. In general, topic areas' score for private schools were higher than international scale score in 5 of 12 topic areas including, earth's processes, cycles, and history; forces and motion; chemical change; physical states and changes in matter. It is worth mentioned that national average scale score (453) was lower than international scale score. Figure 2 showed comparison between private and public schools, visually.

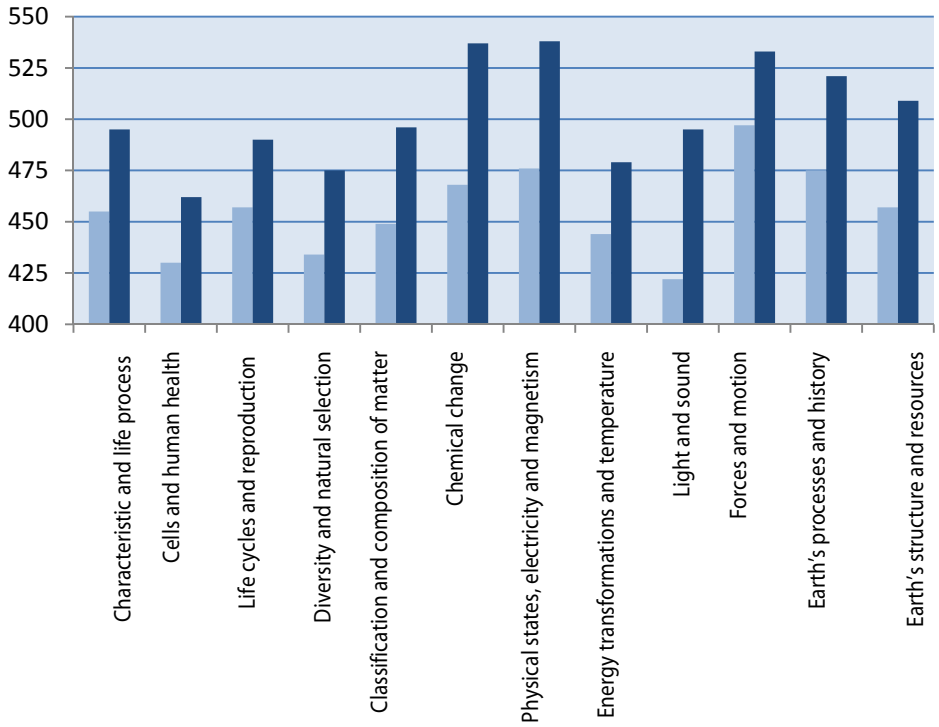

Figure 2. The comparison of students' between public and private schools

Like to type of schools comparing, a comparison was done between achievement of girls and boys in examined topic areas. The results present in table 5 and figure 2.

Table 5. The comparison topic areas' achievement between boys and girls

scale	girls		boys		Standard error of difference	t
	n	average	n	average		
Characteristic and life process of living things	130	460	152	455	10.162	0.533
Cells and human health	382	440	471	425	6.758	2.179
Life cycles, reproduction and heredity	122	465	151	453	11.43	1.049
Diversity, adaptation, and natural selection	891	441	1101	433	6.265	1.226
Classification and composition of matter	766	463	951	443	6.075	3.151
Chemical change	130	492	152	456	18.145	2.022
Physical states, electricity and magnetism	762	482	934	479	7.214	0.324
Energy transformations and temperature	123	452	157	442	12.163	0.83
Light and sound	122	427	151	427	15.323	0.018
Forces and motion	135	504	154	496	17.305	0.487
Earth's processes and history	385	486	475	472	8.73	1.565
Earth's structure and resources	627	461	788	460	8.86	0.095

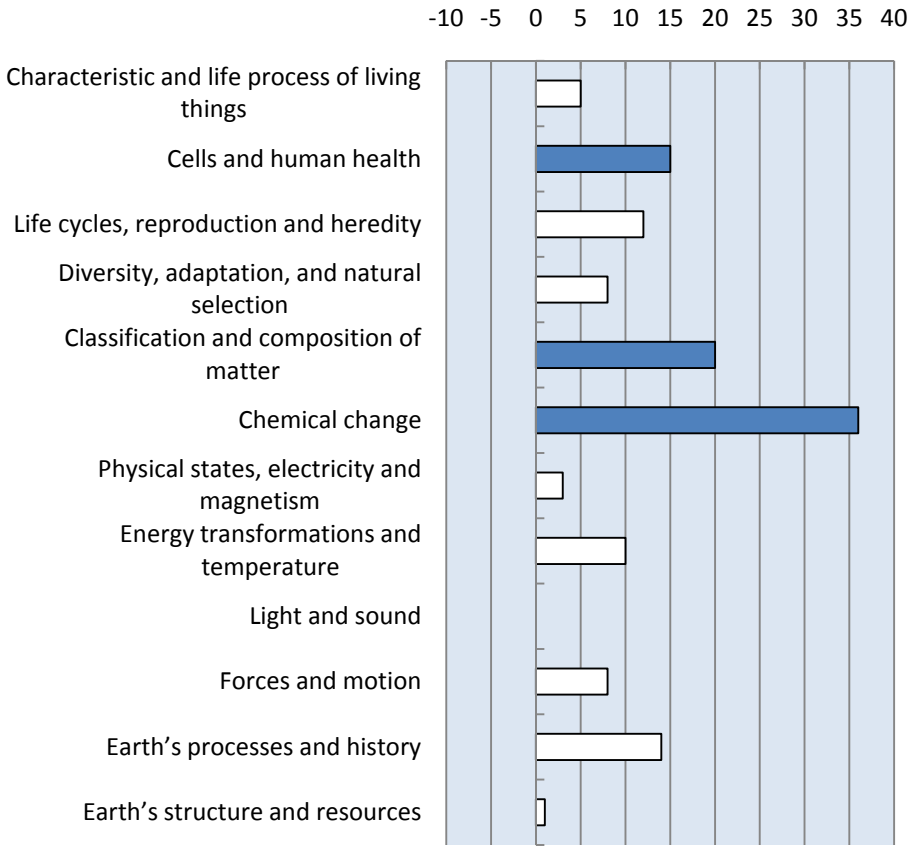

Figure 3. The comparison topic areas' achievement between boys and girls

According results, although girls revealed higher achievement than boys in almost topic areas, but the differences were not significant in many of them. Regarding standard error, the differences were significant only in 3 scales including cells and human health; classification matter; and chemical change. Highest difference was in chemical change that girls performed better than boys about 36 scores. In other side, there is no difference between boys and girls in light and sound.

Identifying the percentages of students reaching the TIMSS benchmarks in each topic areas was another point of view of present study. Table 6 indicated percentages of students in 5 benchmarks. One benchmark was added to detect percentages of weak-achieved students. So, explored benchmarks were advanced (625), high (550), intermediate (475), low (400) and weak (325).

Table 6. Percentages of students reaching the TIMSS benchmarks

Scale	Advanced	High	Intermediate	Low	Weak
Characteristic and life process of living things	0	4	38	85	99
Cells and human health	0	4	26	67	95
Life cycles, reproduction and heredity	1	9	40	80	98
Diversity, adaptation, and natural selection	1	7	33	68	93
Classification and composition of matter	1	10	38	75	96
Chemical change	7	25	51	76	91
Physical states, electricity and magnetism	4	20	52	84	98
Energy transformations and temperature	0	5	34	75	97
Light and sound	1	9	28	63	86
Forces and motion	8	28	61	87	98
Earth's processes and history	2	16	51	86	98
Earth's structure and resources	5	19	43	72	92

As the table showed student reaching to benchmarks differed in topic areas. This difference was in part due to variety average scale point of topic areas. However, exploration of scale distributions showed remarkable points. That is, average scale point of earth's processes and history was higher than chemical change, whereas, more students in chemical change than earth's processes and history reaching to advanced and high benchmarks (7% vs. 2% in advanced benchmark and 25% vs. 16% in high benchmark). This stated that chemical change could better lead to deep and advanced learning rather than earth's processes, but in the same time more general instruction had been provided in earth's processes rather than chemical change.

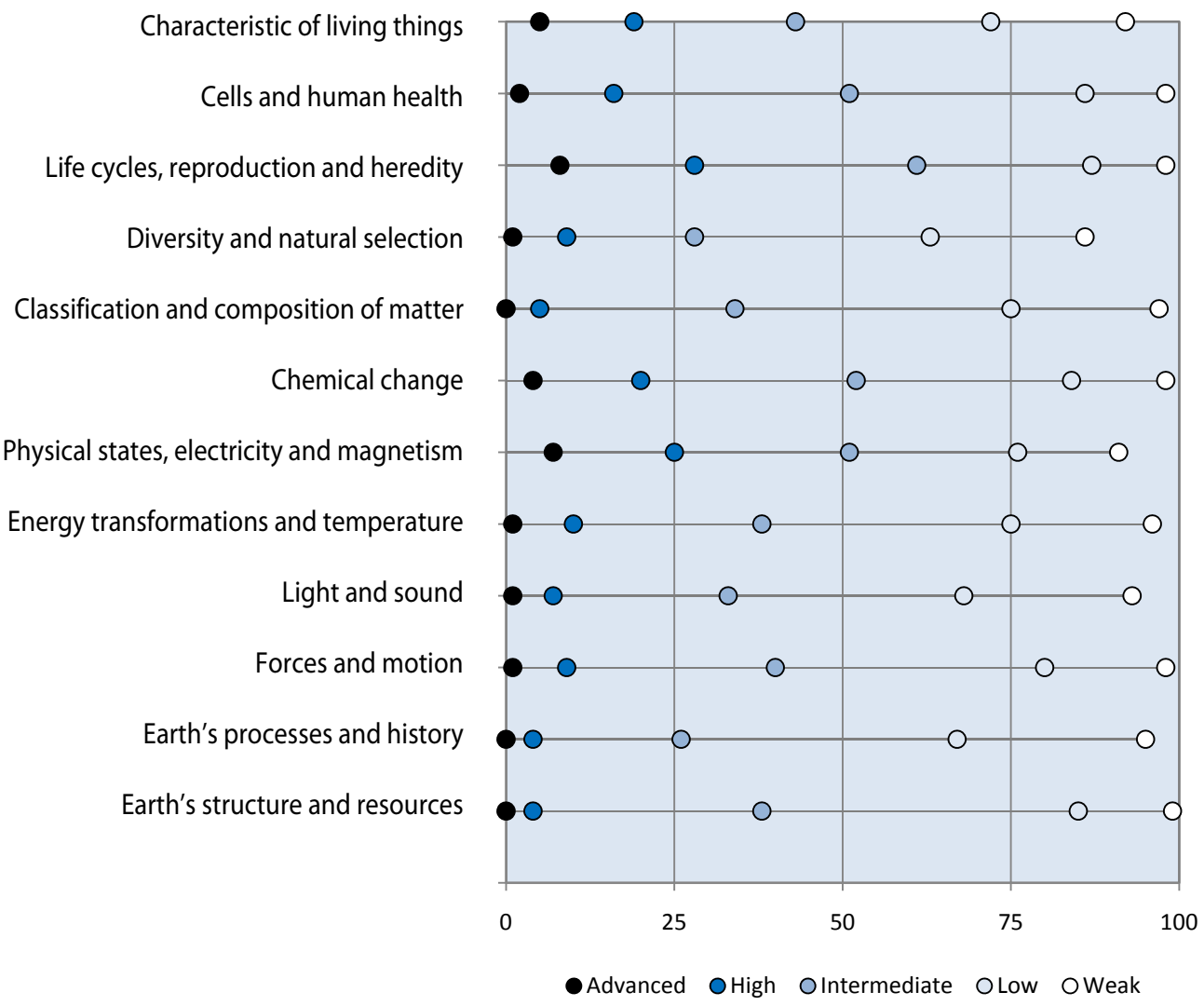

Figure 3. Percentages of students reaching the TIMSS benchmarks

Reletively low percentage of students reached in weak benchmark in some higher achieved topic areas. For example, 91% of students reached in weak benchmark in chemical change and the percentage was lower than most topic areas, reversely, the highest percentage of students reached in advanced benchmark in same scale. Similar situation was in earth's structure and resources. In other side, the percentage of higher benchmarks in Energy transformations, heat, and temperature was not much, but most of students (97%) reached to weak benchmark. These findings showed that some in topic areas emphasis was placed on higher order concepts and ordinal information was neglected.

Conclusion

The present study was designed to estimate topic areas scores for science course of TIMSS 2007 in grade eight of Iranian sample. Although, TIMSS national average scale score along with content and cognitive scale scores provide valuable information about national educational system, however more detailed information about students' performance can guide policy makers, especially curriculum developers to modify and adapt curriculum in order to better achievement. With this point of view, analysis was concentrated into estimate topic areas scores which concerns both appropriate reliability and more detailed scores.

To estimate topic areas scores first number of science items in each topic areas counted and 8 topic areas and 4 combined topic areas were determined. Then topic areas' scale scores were estimated by plausible values and scale linking was conducted. Finally gender and types of school's comparisons were accomplished along with benchmarking analysis.

Results showed that there were variations of topic areas' scores in Iranian eight grader students. ordered from high to poor there were: forces and motion; physical states, changes in matter, electricity and magnetism; earth's processes, cycles and history; chemical change; energy transformations, heat, and temperature; earth's structure, physical, conservation and earth in the solar system; life cycles, reproduction, and heredity; characteristics, classification and life processes of organisms; classification and composition of matter; ecosystems; cells, their functions and human health; and light and sound. Considering topic areas according to their content domains, it revealed that the best and worst performances were seen in physics content domain (forces and motion in contract to light and sound). Another finding of study revealed that private schools significantly better than public schools, but, higher scores of girls rather than boys were not significant in all topic areas (just 3 of 12 scales). In addition, percentages of student in five benchmarks (advanced, high, mediate, low and weak) in topic areas showed differences between distributions of percents in benchmarks. For example, topic area's average scale score of earth processes, cycles and history was higher than chemical change; however, percentage of reached students in high and advanced benchmarks for chemical change were higher than earth processes.

Variation of topic areas' scores is expected due to different content domain scores; however, current results indicated large variation inside of content domains. Based on results, the most variation was observed in physics content domain. Although some explanations about topic presentation or topic sequence can be mentioned, but, further studies should be designed to investigate reasons of variety of performance, especially in physics. Moreover, results showed that emphasis of lessons' presentation were different in topic areas. Assuming same average scale score for topic areas, in some topic areas contents were designed to concentrate advanced learning and less focus on general education, whereas, in other topic areas students were instructed to lower order of learning instead of advanced learning.

Difference between public and private schools was a symptom that performance students were due to implementation of curriculum. Spite of private schools had same contents and curriculum because of centralized education system in Iran, but they performed better than public schools. They used higher standards of education; such as more educational equipments, more trained teachers, smaller size of class and better family's social-economic statues. Indeed, they implemented curriculum better than public schools, consequently, presentation of topics cannot be considered as a single explanation of variety of students' achievement in topic areas, but, more reasons such as implementation of curriculum should be noticed.

No significant gender differences in most topic areas' scores illustrated the situation was same for both boys and girls. Still, in three topic areas, especially in chemical changes girls performed very better than boys. The gender differences finding should be explained with helps to another studies.

In summary, examining the TIMSS results in more detailed perspective can help to diagnose student defects. Although, information about cognitive and content domains can be useful, however, we need to know about students' achievement of topics in content domains and/or activities in cognitive domains. Albeit, the accuracy issue may be challengeable in this perspective, but, test modifications and developments in TIMSS assessment design can achieved us to this goal. Furthermore, adaptation of TIMSS to cognitive diagnostic assessment is another strategy to achieved more detailed assessment design.

References

Adams, R., & Carstensen, C. (2002). Scaling Outcomes. In R. J. Adams & M. Wu (Eds.), *PISA 2000 technical report*. Paris: Publications de l'OCDE.

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, I. V. S. Mullis & A. Arora (Eds.), *TIMSS 2007 technical report*. Boston: IEA TIMSS & PIRLS.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: the role of cognitive models*. Cambridge: Cambridge University Press.

Monseur, C., & Adams, R. (2009). Plausible values: how to deal with their limitations. *Journal of applied measurement, 10*(3), 320.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Boston TIMSS & PIRLS International Study Center.

Olson, J. F., Martin, M. O., Mullis, I. V. S., & Arora, A. (2008). *TIMSS 2007 technical report*. Boston: IEA TIMSS & PIRLS.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data. *Educational Researcher, 39*(2), 142.

Von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What is plausible values and why are they useful? *Issues and Methodologies in Large-Scale Assessments* (Vol. 2, pp. 9-36). Hamburg: IEA-ETS Research Institute.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128.