IN-CONTEXT ITEMS IN A NATION WIDE EXAMINATION: WHICH KNOWLEDGE AND SKILLS ARE ACTUALLY ASSESSED?

Nelio Bizzo[1], Ana Maria Santos Gouw[1], Paulo Sérgio Garcia[1], Paulo Henrique Nico Monteiro[1] and Luiz Caldeira Brant de Tolentino Neto[2]
[1] Faculty of Education and EDEVO Research Nucleus, São Paulo University, Brazil
[2] Santa Maria Federal University, Brazil

Abstract: Over seven million Brazilian youngsters enrolled for a National Test (ENEM 2013) aimed at those who are willing to get one of the 170,000 places available in public free universities (Jan 2014). ENEM submits 180 multiple choice questions in a context with supposedly relevant information which should be applied, relying on few or no previous knowledge. It was originally presented as a new possibility for poor students finding a path leading to good quality universities. We prepared two instruments based on real ENEM questions focusing on the same subject matter (biology), which were presented to two randomized groups of high school students. One group received full length questions (n=1,631), and the other received the same questions (n=1,400) with abridged context information, but with the same stem and options. Performance analysis not only showed no statistical significance towards students who were answering full length questions, but also showed that students' performance was significantly higher in three questions with abridged information. Conclusions show that students' performance may rely more heavily on reading and time management skills rather than on previous knowledge or mental skills. Democratization on university access, if any, may be due to the novelty of the test.

Keywords: students' assessment; intelligence assessment, ENEM; time management skills; reading skills.

INTRODUCTION

The Brazilian Ministry of Education (MEC) organizes a National Test for students at the end of High School ("ENEM") since the year of 1998. In 2009 major changes were introduced, which attracted a great number of students, not only those who are actually at the last year of High School. All people who aspire to a university degree seem to have been encouraged to pursue such testing, given the reward introduced for a good score, in the form of a place in a free public university.

Students have been challenged to achieve the highest possible mark, which would enable them to apply for a place in a computerized system ("SISU") provided by MEC, which compares ENEM scores of students and assigns seats in public universities all over the country. In the year 2013 over seven million students were enrolled in ENEM, competing for places in free public universities. In addition to SISU, students can compete for over 170,000 scholarships in private universities ("PROUNI"), which can be as high as 100% of the tuition fees, given some conditions related to students' socioeconomic status. According to official MEC information, about 110,000 students enrolled in the first version of the then national test in 1998, and no one could believe that seven million people would be enrolled the same test fifteen years later (2013 exam), competing for about 170,000 places in public universities throughout Brazil (January 2014).

ENEM is known for avoiding traditional questions, which rely heavily on the recollection of factual knowledge. Since it was launched, it was presented as a new strategy to assess directly students' competencies, defined by an official document as "structural modalities of intelligence" (Franco and Bonamino, 1999:29). The new test was warmly welcomed by the Brazilian press and broadly marketed in "grey literature", which is difficult to quote. Apparently it was taken as a strategy not

only for a new assessment-based educational reform, but also for social reform, as it would help poor students to pursue a path to higher education and, in addition, was aimed explicitly at reaching the job market. MEC presented the test as an opportunity for youngsters to plan their futures, having a "clear idea of their personal and professional potential", as the test "would allow assessing their potential in order to plan future choices" (Zákia and Oliveira, 2003: 884). Even today, the "Novo Enem" ("New ENEM") is officially presented by MEC as a tool for democratization of access to public institutions of higher education, which are free, to promote academic mobility and to induce changes in high school curricula (MEC, 2013).

ENEM was originally based on five competencies and 21 abilities, aimed at reaching an interdisciplinary approach, with no mention to specific school disciplines or subjects. The major reform which took place in 2009 created the "Novo ENEM" ("New ENEM") with a major increase in the number of competencies and abilities under assessment, references to conceptual disciplinary knowledge were introduced, and the total number of questions increased dramatically. The original 63 multiple choice questions (plus an optional written composition) performed in one afternoon became, now in the new version, 180 questions (plus a compulsory written composition) and two days are necessary, with a tight time schedule, which allows three minutes per item. They are taken as unidimensional, as Item Response Theory (IRT) is now applied to establish final scores. However, the major features of items construction seem to be essentially the same: some visual and written context is given, followed by a stem and five options. Recollection of facts and concepts should be rarely necessary, at least in the form of conceptual definitions; the essential information to find the right option is supposedly part of the context given.

Previous research carried out with PISA items, which are also based on a stimulus which "'tells a story' to which the test items relate more or less directly", categorized items according to the level of contextualization (Nentwig, et al, 2009). Items with "high level of contextualization" had stimulus content which was essential for information extraction and processing, whereas items with "low level of contextualization" brought stimulus which was not essential for answering the question. In that piece of research both stimulus *Content* and *Relevance* were taken into consideration in a threefold scale, in which items could have substantial information, which was relevant for item solution (score 2), or could have some text or information but stimulus information was not relevant for solution (score 1). Items could also bring few or no information as stimulus (score 0).

Authors provided examples of items of PISA 2006 in which the "question can be answered – and exclusively so – with the recollection of factual knowledge not related to the stimulus", and were coded 1. Their objective was to carry out further performance studies of selected questions, comparing students of different countries, in order to understand how well German students could extract and process information, rather than find the right answer recollecting factual knowledge.

Data is presented here testing the hypothesis that stimulus in a group of selected ENEM questions was actually relevant for student performance in biology. Instead of simply rating questions on the basis of stimulus *Content* and *Relevance* by judges, as done in the cited article, an additional step was added. Low contextualization questions, corresponding to score 1 of Nentwig et al, 2009, were selected and presented to students in two forms: full length, with the original stimulus, and abridged version, in which stimulus was removed, leaving just the stem and options. Scores on the two groups of students are presented and we discuss methods for identifying possible flawed multiple choice items.

METHODS

A sample of seven questions with low level of contextualization clearly related to biology were selected in the 2009 and 2010 tests (Novo ENEM), which were presented in 2011 to two randomized groups of High School students. One group (n_0=233) was asked to answer original questions (Full ENEM), in a six-page long questionnaire; another group of similar students (n_1=200) was asked to answer the same questions with written stimulus entirely removed, leaving the stem and the very same options, in the form of a three-page long questionnaire (Abridged ENEM). Another three questions (standard questions) were included in the two sets of questionnaires, focusing Biology subjects, with exactly the same brief stem and five options, for comparison purposes. As survey participants were not selected by randomised procedures, these questions would test general biology knowledge of the two groups, ascertaining their proficiency in the field (Biology) was equivalent, and therefore the sample could be reliable for the only purpose of comparing items. According to quota sampling techniques, choice of quota controls would "challenge the quota sampler's ingenuity" , as "quota variables should be strongly related to the survey variables" (…) thereby becoming "substantially homogeneous". As Leslie Kish states, quota sampling is not a standardized scientific method; "rather, each one seems an artistic production" (Kish, 1965: 563), and a overview is provided below.

Each research assistant received one set of questionnaires, either short or long, and was responsible for submitting it to students of one public high school of the city of São Paulo (SP, Brazil). Fourteen schools were chosen according to assistant's convenience, as access to schools is quite difficult, and test was performed by students within a specific week in mid September. Research assistants were not aware of the differences of the two sets of questionnaires. The invitation letter required by the Ethics Commission of our institution (FEUSP) was part of every questionnaire, and stated that students were invited to collaborate in a research about assessment; they would not be identified in the answer sheet, and the several participating schools in this piece of research would not be identified or ranked.

School validation relied on a two level process. Reports of how the questionnaire was presented to students and answered were analyzed, prior to the answer processing. Any kind of reported situations which were not exactly the ideal ones led to school exclusion. For instance, when different research assistants went to the same school, it was excluded from the sample, as students could have had notice of the different length of the questions. We could validate fourteen schools at this level. On another level of scrutiny, as part of the statistical analysis, school results were studied, a search for outliers was carried out (see below), and one case was found in the group of schools where abridged questions were presented, and the report of that specific school was reconsidered. The school has a long record of good performance in large scale evaluations, but students now had very low scores compared with the average of other schools. Score on the standard questions were 11%, which is surprisingly low for items with five options. The conclusion was that this specific school was close to the university campus and students were not motivated to perform the test, as they are quite used to similar "university experiments". As they could not recognize items as "ENEM questions" the task was probably seen as "a waste of time". Therefore, that school was considered an outlier, and the number of students aswering abridged questions was corrected to n=127 (889 items analyzed). Students which answered full ENEM questions was n=233 (1631 items analyzed), with a total sample size of 360 students belonging to 13 schools, and 2,520 ENEM items and 861 standard items analyzed.

ITEMS EXAMPLES

The following examples show the twofold forms of presentation of selected items. In the full version items were reproduced from the beginning, where the question number appears for the first

time. In the abridged version, stimulus was removed, and the version presented to students began where the question number appears for the second time, in the examples below. Colors will be discussed below.

> 7 (full) - The biogeochemical carbon cycle comprises various compartments, including Earth, the atmosphere and the oceans, and various processes allowing the transfer of compounds between these reservoirs. Carbon stocks stored in the form of non-renewable resources, such as oil, are limited, being of great importance to realize the importance of replacing fossil fuels by renewable fuels.
>
> 7 (abridged) - The use of fossil fuels affects the carbon cycle, as it causes:
>
> a) increase in the percentage of carbon on earth.
> b) reduction in the rate of photosynthesis of higher plants.
> c) increased production of carbohydrates produced by plants.
> d) increase in the amount of atmosphere's carbon.
> e) reduction of the overall amount of carbon stored in the oceans.

> 8 (full) - A new method for producing artificial insulin using recombinant DNA technology was developed by researchers at the Department of Cell Biology, University of Brasilia (UNB) in partnership with the private sector. Researchers have genetically modified *Escherichia coli* bacteria, which became able to synthesize the hormone. The process allowed the manufacture of insulin in larger quantities and in only 30 days, one third of the time required to obtain it by the traditional method, which consists in the extraction of the hormone from slaughtered animals' pancreas.
>
> Ciência Hoje 24 April 2001. Available at: http://cienciahoje.uol.com.br (adapted).
>
> 8 (abridged) - The production of insulin by recombinant DNA technique has, as a consequence :
>
> a) improvement of the process of extracting insulin from porcine pancreas .
> b) the selection of antibiotic-resistant microorganisms .
> c) progress in the technique of chemical synthesis of hormones.
> d) favorable impact on the health of diabetics .
> e) creation of transgenic animals.

Distractors' keywords appear in color, associated with related terms in stimulus. As Thiessen et al (1989) argued, they play an important role in item planning, and improve options' plausibility. As we will argue later, a long, but not relevant, stimulus may improve the effectiveness of distractors to the point of flawing the whole item.

RESULTS

The total number of questions focusing the national test was 2,520 (Table 1), other 861 standard questions were included in order to test sample homogeneity (Table 2), with a total number of 3,381 questions answered and processed.

Statistical analysis included parametric essays, and search for outliers. One school (EEI1FB, n=73) fell into this category, as previously mentioned, and was excluded from the sample. Fisher's Exact Test for ENEM questions (Table 1) reported p-value without statistically significant differences between the groups on four questions (Q1, p-value= 0.906; Q5, p-value=; 0.077; Q9, p-value = 0.901; Q10, p-value = 0.152), and statistically significant differences on three questions, in favor of abridged questions (Q03, p-value = 0.006; Q7, p-value < 0.001 e Q8, p-value < 0.001). Results of the same statistical analyses for the three standard questions (Table 2) confirmed the sample's homogeneity of the two groups (Q2, p-value = 0.787; Q4, p-value = 0.116 e Q6, p-value = 0.140).

Table 01

Right answers of the 2,520 ENEM questions (F.E.T= Fisher Exact Test)

			Full ENEM Questions						
N	School	n_0	Q1	Q3	Q5	Q7	Q8	Q9	Q10
1	EEA0HD	15	1	12	13	4	2	3	3
2	EEB0NL	52	18	30	32	11	18	10	12
3	EEC0SB	32	15	22	20	6	16	10	21
4	EED0XS	13	7	9	12	9	6	7	8
5	EME0EA	32	12	8	17	7	4	17	7
6	EEF0PM	34	21	27	28	23	19	3	17
7	EEG0GC	13	7	10	11	11	3	9	11
8	EEH0BT	42	27	32	34	26	9	21	20
	Total	233	108	150	167	97	77	80	99
			46%	64%	72%	42%	30%	34%	42%
			Abridged ENEM Questions						
		n_1	Q1	Q3	Q5	Q7	Q8	Q9	Q10
09	EEK1BM	43	14	35	36	25	29	13	15
10	ETL1HV	25	10	19	20	20	7	11	18
11	EEM1BC	19	14	17	13	9	5	14	6
12	EEN1MS	18	4	10	12	9	10	3	11
13	EEO1HF	22	20	19	21	16	18	4	14
	Total	127	62	100	102	79	69	45	64
			49%	79%	80%	62%	54%	35%	50%
	F.E.T	p.value	0.906	0.006	0.077	<0.001	<0.001	0.901	0.152

Table 02

Results of the 861 standard questions (F.E.T= Fisher Exact Test)

			Full ENEM Questions		
N	School	n_0	Q2	Q4	Q6
1	EEA0HD	15	4	10	2
2	EEB0NL	52	8	17	4
3	EEC0SB	32	8	13	7
4	EED0XS	13	4	7	5
5	EME0EA	32	3	11	7
6	EEF0PM	34	4	11	7
7	EEG0GC	13	3	11	4
8	EEH0BT	42	15	6	10
	Total	233	49	86	46
			21%	37%	20%
			Abridged ENEM Questions		
		n_1	Q2	Q4	Q6
09	EEK1BM	43	4	9	4
10	ETL1HV	25	10	16	10
11	EEM1BC	19	6	16	6
12	EEN1MS	18	1	1	1
13	EEO1HF	22	4	16	12
	Total	127	25	58	33
			20%	46%	26%
	F.E.T	p.value	0.787	0.116	0.140

Table 1 presents the results of the two groups of experimental questions. Considering this group of low contextualization items, the hypothesis that stimulus is relevant to student performance found no support, confirming previous categorization. An even more surprising result was found, as comparing the two groups of ENEM questions answers it is possible to state that questions Q3, Q7 and Q8 allowed a statistically significant higher student performance when they brought no stimulus, showing a phenomenon we named *reversed induced performance* ("rip"). In other words, jumping stimulus brought to students, in this group of questions, either the same or even better probability of a good performance.

A further analysis was performed with linguistic tools looking for causal explanations of these surprising results. The group of students which answered items with no stimulus, went directly to the stem line, and was not influenced by the text presented to the other group. These texts had keywords, such as oil and insulin, which were also inadvertently referred to by their superordinated words ("fossil fuels", and "hormones"), demanding previous knowledge for full understanding.

In the item examples given, question 7 brings a text with poor information on the topic of carbon cycle, and has lack of cohesion, comprising also the global warming issues. Item stem explores previous student knowledge on a specific topic (effect of fossil fuels on the atmosphere). Without previous knowledge, students, under pressure due to the tight time schedule, would read options directly looking for similarities between keywords found there and in the text. There are three "carbon reservoirs" mentioned in the text, and they appear on three different options. The stimulus would drive students' attention to these three options, whereas without it they would face a different situation, thus becoming weak distractors. "Fuel" is a keyword in the stem, which easily connects to the idea of combustion and smoke. The closest keyword is "atmosphere", which is found in the right answer. Therefore, lack of cohesion of the text could lead students to jump stimulus, and concentrate in the stem, rising the probability of success, including reasons other than those originally thought. This trajectory could explain the observed "rip".

The other example is even clearer, as question 8 was presented above so that keywords were colored, as their related terms, in the options and item stem. Apparently, students have to apply information given in the text, as stem is plenty of keywords such as "insulin". Stimulus brings keywords which appear (or have correlated ideas) in four distractors. The only option which has no connection with stimulus, as mentions "diabetics", is the right one. Students should recollect facts about hormones and insulin, and know something about the related diseases, as stimulus has evident lack of cohesion regarding the context of the right answer, related to diabetics' treatment. Students who read stimulus would be bound to focus attention on the four distractors. There is a clear lack of cohesion, as stimulus does not mention any disease; this strategy of diverting students' attention by changing subject, making stimulus not relevant for the answer, we called "bafflement", which tends to improve "rip". In fact, this was the question with the greatest difference between the two groups (Table 1). In real action, students could jump stimulus and would not be mislead to concentrate their attention in the wrong options; with previous knowledge about insulin and related disease, answer would be easily found. Therefore, it is possible to understand the observed "rip" as a consequence of this bafflement strategy.

The four items in which no statistical difference between the abridged and full questions was found also deserve analysis, as students who jumped stimulus, and went directly to options, were as successful as those who did all the reading. However, as they have only three minutes for each question, 'jumping" students would have saved precious time for other questions, rising the probability of a higher final score in real action.

These results show that low contextualization in ENEM questions with focus on Biology do rely on students' previous knowledge, and are not objective indicators of the alleged "structural modalities

of intelligence". Moreover, items actually favor students with better reading and time management skills than a balanced amount of biological knowledge and thinking skills.

DISCUSSION

Results show that low contextualization items (Nentwig et al, 2009) deserve more attention regarding future research. If presented with a stimulus, which demand a considerable length of time to be read and understood (score 1 in the cited article), they are actually "context deficient" ("cont-def"). These items allow at least two different paths for the right answer, what brings a considerable problem for the task of determining its degree of difficulty, with a serious implication for the Item Response Theory. Contrary to direct items with no context (score 0 in the cited article), or with actually relevant information in the stimulus (score 2 in the cited article), "cont-def" questions not only allow similar probability of success with stimulus or without it, as seen in questions 1, 5, 9 and 10 (Table 1), but also may turn the question even more difficult. As seen in questions 3, 7 and 8 (Table 1), scores of students that received stimulus were significantly lower, showing a new phenomenon, which we called *reverse induced performance* ("rip").

This new phenomenon should be focused carefully in items pre-testing, as it brings a profound effect to the determination of the degree of difficulty. The validity of Item Response Theory requires unidimentional items, therefore items must be rip-free. This piece of research offers a practical approach to perform such testing, with two randomized groups of students, one of them receiving abridged items, which are suspect of being cont-def.

This research brought a new light to a long known fact, related to the commercially successful "ENEM training courses", privately owned, which have been active at least since 2003 (Zákia and Oliveira, 2003: 885). There was suspicion that they were useless, as students would receive all, or almost all, information needed in items' context stimulus, therefore, training would be of no help to raise students' performance in ENEM. All seven experimental questions produced results that can explain the need of a specific student training to get higher scores in that exam. Within a very tight time schedule, students may be trained not only to extract and process information given in the stimulus, but also – and mainly - to select and discard information which is not relevant to assign the right option or even to lower distractors' efficiency.

The 2009 reform turned Novo ENEM not only into an instrument to select students for public universities, but also aiming at monitoring education quality in a nationwide basis. The democratization of higher education access provided by ENEM (if any) may be due to sudden changes and would tend to disappear as time management skills are differently apprehended by students in the socioeconomic spectrum.

The proposal of turning ENEM into a compulsory State Exam is currently under discussion. Our results suggest that assessment-based educational reform and education quality monitoring based on this instrument should be considered with caution. Further research is necessary, encompassing other content areas, in order to have a clearer idea of the real impact of low contextualization items in large scale exams such as ENEM.

ACKNOWLEDGEMENTS

Authors want to express their gratitude to the following persons: Alessandra Lupi, Alessandra Stranieri, Alessandra Ramin, Andréia Vieira, Ariana Carmona, Bianca Dazzani, Bruna Lourenço, Bruno Vieira, Carolina Bueno, Cristiano J. da Silva, Débora Brandt, Fernando Sábio, Giselle Armando, Guilherme Antar, Guilherme Stagni, Helenadja Mota, Henrique Neves, João Ferreira, Karina Tisovec, Laisa Lorenti, Mariana Rosim, Marina Medeiros, Natacha Loído, Pedro Machado,

Priscylla Arruda, Rafael Ogawa, Renato Rego, Rodrigo Gonçalves, Rodrigo Dioz, Samara Moreira, Talita Oliveira, Thaís de Melo, Thales Hurtado, Thiago Madrigrano, Vitor Lee. The following institutions provided funds for the several parts of the research: CNPq, FAPERGS, FAPESP, FEUSP and Pró-Reitoria de Pesquisa da USP.

REFERENCES

Franco, C., A. Bonamino (1999). O ENEM no contexto das políticas para o ensino médio. *Química Nova na Escola 10*, 26-31.

Kish. L. (1965). *Survey Sampling*. New York: Wiley & Sons, Inc.

Ministério da Educação (1999). *Exame Nacional do Ensino Médio – ENEM: documento básico 2000*. MEC/INEP.

_____(2013). *Sobre o ENEM*. Available at http://portal.inep.gov.br/web/enem/sobre-o-enem [access on Dic 15 2013].

Nentwig, P., Roennebeck, S., Schoeps, K., Rumann, S. and Carstensen, C. (2009). Performance and levels of contextualization in a selection of OECD countries in PISA 2006. *Journal of Research in Science Teaching, 46* (8), 897-908.

Orlandi, E. P. (2012). *Discurso e leitura*. São Paulo: Cortêz.

Thiessen, D. L. Steinbeck and A.R. Fitzpatrick (1989). Multiple-choice items: the distractors are also part of the item. *Journal of Educational Measurement 26* (2),161-176.

Zákia, S.; R. P. Oliveira (2003). Políticas de avaliação da educação e quase mercado no Brasil. *Educação e Sociedade 24* (84), 873-895.