

Accountability Systems in Education

Enrico Gori

Abstract Accountability systems in education has been longly debated and criticized. In this paper we face with some solution to the main problems of these systems, that can be overcome by using Rasch measures of students abilities in basic skills. Some new results on value-added analysis are reported based on models that take account of measurement errors in abilities. This results show that ideas regarding the effect of socio-economic status on learning must be reconsidered.

1 Quality in Education

The role of states and governments in Education has been widely debated (Glenn & de Groof, (2003)). Recently this debate has been focused on Quality. The term “Quality” may concern the production process, or the output of Education. Obviously quality of the process is related to quality of the output. But trying to define the first without paying attention to the latter maybe misleading. In Education, where Quality of the output is still a not well defined concept, this error is quite widespread in Italy (De Rozario (2002)) with special reference to professional training programs and University. In such an error did not fall the main international researches on Education which identified Quality of the output as the level of students in basic skills such as language, math and sciences, measured with methods which ensure comparability in space and time (OCSE-PISA & TIMMS). Results of this researches are quite interesting as shown, for example, by figures 1 and 2.

Studies in the field of economics of education and human capital, based on such data, showed that the levels in this basic skills are the most important indicators for educational systems (Hanushek & Kimko, (2000)). This implies that the construction of Accountability Systems focusing on basic skills (rather than quality of processes such as in ISO9000) is the most relevant problem (Gori, (2003)).

Figure 1 Country mean scores: PISA 2003 versus TIMSS 2003 for math and sciences (Wu, (2008))

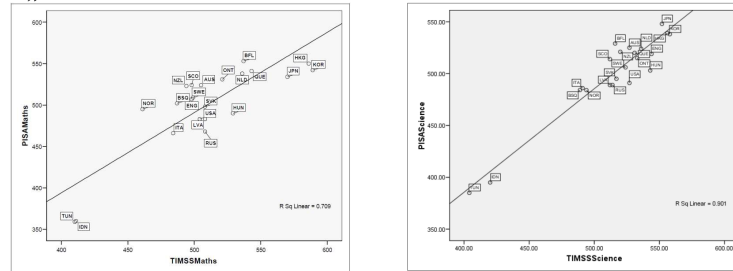
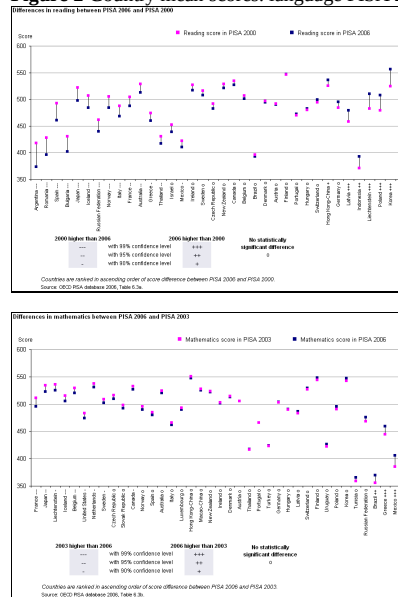


Figure 2 Country mean scores: language PISA 2000-2006, and math PISA 2003-2006



1.1 Which institutional framework for Education?

Relevant researches (Bishop & Woessmann, (2001)) shown that educational systems characterized by strong emphasis on accountability regarding achievement results and standards, are able to reach greater levels of quality, and this is true with special references to systems in which schools enjoy autonomy in resources management and teaching programs. These systems moreover show better results with respect to that characterized by indiscriminate grown of resources (Hanushek, (2003)), and/or pure

control on processes without references to achievement results (Bishop & Woessmann, (2001)). The ways in which systems of accountability can be structured and implemented, however, are very different. Comparative studies (Glenn & De Groof, (2003)) have shown that school systems differ from each other for the way they implement policies for the distribution of resources, school autonomy, and leverage freedom of choice for families: which leads directly to the problem of the institutional educational systems. Hence, systems of accountability may produce different effects in terms of quality, depending on the institutional framework (Gori and Vidoni, (2005); Gori and Vittadini (2005)). In the U.S., for example, near systems which lack of any accountability system, we find others which publish regularly statistics on the student performance in final examinations (report cards), or others where incentives are tied to these results. As shown by the researches of Hanushek and Raymond (2004) the response of school systems in terms of quality, is very different. So there is ample room for research to identify the ideal mix of elements:

- Assessment (or production of information),
- Resources (allocation rules and incentives)
- Autonomy (freedom to manage resources and processes)
- Freedom of choice by families

and interaction between them, in order to constantly improve the quality of schools and the education system as a whole. The discussion below will focuses on this aspect, seeking to identify some possible guidelines to direct the policies of the institutional reform of the Italian educational system. Although the above analysis are significant considerations quite in general.

1.2 Critical aspects of Accountability Systems

In the face of evidence and many voices in favor of implementing systems assessment, there are some critical issues (Dolton, (2002)) which you must clear the field before launching in the more proactive. These arguments must be faced first, before to deserve attention in order to build assessment systems more efficient and closer to the teachers and students. The most important are the following (Gori (2003)):

- (A) we often hear people arguing, negatively, that the levels of learning in basic skills are just one of several objectives of the school and the education system;
- (B) other critics to tests in basic skills are linked to the question of their weakness with regard to important aspects of measurement theory: the tests are constructed so as to favor certain students or groups and not would be "valid" and "reliable" because it is not possible to account for important information regarding the latent dimension that you want to measure (logical-mathematical skills, language, etc.), like other types of tests, written essays, oral tests, etc. ;
- (C) further criticism regards the negative consequences that may arise when these results are used to determine impact on schooling of the individual student and / or the teacher. The worst of these lies in the fact that students and / or teachers would focus on mere pass / teaching test at the expense of more comprehensive training (Learning / Teaching to the test).

2 Abilities in basic skills

Regarding point (A) there are at least two strong arguments in favor of an accountability system confined to monitoring of basic skills. The first dates back to John Stuart Mill and presents the impressive modernity in light of today's multi-racial and multicultural society: *"To Prevent the State from exercising through These Arrangements (exams), Improper Influence over an opinion, the knowledge required for passing an Examination (Beyond the Merely instrumental parts of knowledge, Such as languages and Their use) Should ... Be Confined Exclusively to facts and positive science"* (Mill (1806-73)). The second argument, more technical, sees two major theories in favor of the opportunity for evaluation systems to be limited, at least for now, to basic skills: the theory of measurement and economic theory. Regarding the theory of measurement, in order to produce measures that are "objective" and "valid", one of the most important requirement of this theory (Messick, 1989) must be met: namely that of "external validity". The satisfaction of this condition requires that the measure created must be related - statistically significant, justifiable in theory, and with a certain stability over time and space - with something that is widely regarded as objective and valid. With reference to economic theory, education is an investment to enhance human capital (Gori (2004)) and, as such, only results in tests in basic skills – between the different indicators that have been constructed in order to monitor the education system - have shown up to now to be related with phenomena (indeed objective and valid) such as employment, income, the opportunity to accumulate additional human capital (eg to successfully continue studies) and economic growth (Hanushek and Kimko, (2000)). These are some of the reasons why it is appropriate that the focus of the evaluation focuses mainly on basic skills of primary and secondary schools, without excluding the possibility that the measurement of learning may also extend to more specific skills and related to levels of more advanced studies, insofar as they relate to "facts and positive science only", according to Mill.

3 Validity and reliability of tests

With reference to point (B), it should be noted that there are different schools of thought on how to build valid and objective measures of latent dimensions, such as logical-mathematical skills, language etc. and, more generally, measures of latent dimensions in areas where measurements are not yet recognized as objective and reliable. Among these schools of thought the most relevant is that one from the Danish mathematician Georg Rasch, who based his theory and his work on the definition of *Specific Objectivity* (Rasch, (1968) (1977)). This concept represents a mathematically rigorous definition of the intuitive concept of "objectivity" in common sense. An interesting fact is that all measures of one-dimensional phenomena, such as weight, distance, temperature, etc., on which science has based its success in explaining physical phenomena, meet the definition introduced by Rasch. This means that any new measures on dimensions not yet "measured" - or believed "not measurable" - which should be regarded as objective as those used in the natural sciences, must meet the definition of *Specific Objectivity* (Gori et al. (2005)). From this perspective, the path

taken by Rasch is not "one of the options for analyzing the results of a test", but is the only scientifically valid way to create objective measures in those areas of science - typically the so-called human sciences - which have not yet gained the dignity and recognition of the physical and natural sciences and that, precisely on the possibility of using "real" objective measures have based their success. In this sense is often forgotten that these measures are the result of extensive research by humanity in phenomena (such as temperature, or weight) that have long been represented as something mysterious and "qualitative" (ie not measurable) such phenomena as the "motivation", the "proficiency" or other entity more or less abstract (since it is not objectively measurable). Rasch's legacy lies in the method for constructing objective measures in the humanities. For our purposes the definition of *Specific Objectivity* rule out the possibility that a test may be considered if it benefits certain students or groups. This aspect is just one of the major areas of research development in the construction of tests in the methodology of Rasch (Gori et al. (2005)). In addition, the Rasch methodology is the only one that can be used (according to the principle of *Specific Objectivity*), to analyze information like ratings achieved by a student in a written or oral, for the measurement of skills, along with the usual test Multiple choice (Linacre, (1989), Gori et al. (2005)).

3.1 External validity and measurement errors

With reference to the issue of external validity, the Rasch methodology moreover presents an additional advantage over other approaches (eg those based on the consideration of the percentage of correct answers) which is the ability to obtain measures of the degree of uncertainty in the estimated amount (called standard error of measurement). This information is crucial to correctly calculate the correlations and, through them, to judge the degree of external validity of the measures constructed. It is in fact well known (but often forgotten: cf. Rogosa (1995)) that the correlation between two quantity θ (for example the ability in math) and Y (for example income, or any other variable measured without error), when θ is measured with error, is underestimated (attenuated) so greater the greater the variability of measurement error. This could lead to the conclusion that the measure is not valid since the correlation is considered, erroneously, not statistically significant. From the formal point of view the situation is as follows: If you indicate with θ_n the latent ability of the student n in the skill-dimension of interest, and with $\hat{\theta}_n = \theta_n + \varepsilon_n$, its estimate, where ε_n is the measurement error, the Rasch model provides estimates of $\hat{\theta}_n$ and estimates of $\text{var}(\varepsilon_n | \theta_n)$. Using such quantities is possible to introduce the reliability index :

$$(1) \quad \rho(\hat{\theta}) = \frac{\text{var}(\theta_n)}{\text{var}(\theta_n) + \frac{1}{N} \sum_{n=1}^N \text{var}(\varepsilon_n | \theta_n)}.$$

This index is crucial to study the correlation between θ (the ability in this case) and other variables such as gender, socioeconomic status, context and so on. As Rogosa (1995) has shown, the observed correlation $\rho(W, \hat{\theta})$ between a random variable observed without error W , and $\hat{\theta}$, must be divided by the square root of the index of reliability in order to obtain an unbiased estimate of the true relationship sought

$$\rho(W, \theta) = \frac{\rho(W, \hat{\theta})}{\sqrt{\rho(\hat{\theta})}}.$$

Studies that do not take explicit account of the presence of measurement errors can therefore lead to completely wrong conclusions. Among these, for example, the ones that seek to identify the reasons for the different quality, measuring the role of school, class, and socioeconomic factors in the growth of skills, known as "value added" models (Aitkin and Longford (1986)). Typically these studies postulate that the level of skills at time 2 is function of the level of skills at time 1 and socio-economic variables and context, such

$$(2) \quad \theta_{ij}(t_2) = \alpha + \beta \cdot \theta_{ij}(t_1) + \gamma \cdot x_{ij} + u_j + e_{ij}.$$

Now, often has been forgotten that $\theta_{ij}(t_2), \theta_{ij}(t_1)$ are observable only with error.

This has two main consequences (Battaaz, Bellio, Gori, (2010)):

a) the effect of $\theta_{ij}(t_1)$ is systematically underestimated (as in the case of the attenuation of the correlation), which can also produce a bias in the estimation of the effect of other explanatory factors: the distortion typically implies an overestimation of the effect of factors positively correlated with $\theta_{ij}(t_1)$ (such as socio-economic conditions), and an underestimation of those negatively correlated with $\theta_{ij}(t_1)$; in this regard refer to the results obtained by a method of correction for the error in the explanatory variable $\theta_{ij}(t_1)$ (Battaaz, Bellio, Gori, (2010)) in multilevel models; such results show that the effect of socioeconomic factors, once account is taken of measurement errors, is much reduced. In particular (Table 1), passing from Naive model, in which no account is taken of measurement errors, to ML estimates on which account is taken of this aspect:

- the initial levels of skills have a much higher effect in explaining the next level;
- from 7 socioeconomic variables (SES) significant in the naïve model, just two remain significant once measurement error is considered. S.E.S. variables instead continues to have a relevant weight in explaining the initial level. T

Therefore the model with measurement error, besides providing the correct estimates, has also the advantage to clarify the dynamics of classroom learning that is little influenced by SES factors, and much more influenced by the baseline and the class. While baseline is obviously linked to socio-economic factors: it's worth to note that this is true because we often lack information regarding the school history of the students, but if we have such information probably the socio-economic factors effect could be much lower.

Table 1 – Comparisons between estimates Naive and ML

	Naive (*)		ML S.E.S effect estimates on dependent variables (*)		ML S.E.S. effect estimates on initial value (*)	
	Estimates	S.E	Estimates	S.E	Estimates	S.E
time 1	0.238	0.235	0.697	0.265	-	-
time 2	1.218	0.230	1.585	0.258	-	-
Female	0.078	0.049	0.085	0.055	-0.061	0.063
Mother education level (relative to primary school)						
middle school	0.219	0.152	0.179	0.169	0.120	0.192
high school	0.303	0.155	0.263	0.173	0.148	0.197
university	0.214	0.166	0.147	0.184	0.226	0.210
Father education level (relative to primary school)						
middle school	0.279	0.139	0.210	0.158	0.203	0.176
high school	0.422	0.142	0.293	0.162	0.418	0.179
university	0.514	0.151	0.388	0.171	0.393	0.190
Number of books at home (relative to 0-10)						
11-25	0.085	0.173	-0.062	0.189	0.500	0.216
26-100	0.277	0.162	0.053	0.178	0.819	0.202
101-200	0.347	0.165	0.074	0.182	0.988	0.204
> 200	0.566	0.166	0.268	0.183	1.000	0.204
Baseline level	0.574	0.032	0.836	0.044	-1.334	0.275

(*) in italic: estimates of S.E.S. significantly different from zero at 10%

b) the usual intraclass correlation index $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$, which measures the portion

of variability in growth (or value added) due to the class, contains in the denominator the variance of the estimate of $\theta_{ij}(t_2)$, for which this index tend to be lower than true (see Gori (2006)) and therefore all studies that do not take account of measurement errors tend to systematically underestimate the effect of class on the growth of abilities (value added). The underestimation of the intra-class correlation coefficient and overestimation of the effect of socio-economic factors are likely to divert attention from what intuitively (and in fact if you make the necessary corrections for measurement errors) is, perhaps, the most important determinant of the skills, the effect of class and teachers, which is much more important than socioeconomic status. The latter is often considered instead, in many scientific circles too "politicized", the "dominant" factor. The work by Gori (2006) and Battauz, Bellio, Gori (2010) shows instead that this result may arise from not considering measurement errors and past school history. Moreover here we must remind the distinction between two fundamental concepts in statistics: the significance of a parameter relative to an explanatory factor, present in a regression model, and the proportion of variability explained by this factor: the first term is wise if the parameter is statistically different from zero, with the second measure is rather important factor in explaining the observed differences. When a factor such as socioeconomic status, were statistically significant, but would explain only 10% of the variance of skills, this means that there are other factors that explain 90% of the variability observed. In this context, policies aimed at removing differences due to social factors would be able to reduce the variability by only 10% that would be fair if the other 90% was due to chance or to factors which can not be affected. Fortunately, in

the case of education, it is not. In this regard, see the results on the percentage of variability of the growth of logical-mathematical skills (obtained from a model type gain $\theta_{ij}(t_2) - \theta_{ij}(t_1) = \gamma \cdot x_{ij} + u_j + e_{ij}$) explained by economic factors and class, in the cited study (Gori (2006)): once you correct for the variance of the measurement error the proportion of variability explained by socio-economic factors, while significant, attesting levels around 3%, while that explained by the class averages 57%.

4 Using the results of evaluation

The question of the use of test results in order to determine the student's school career is certainly more complicated. But we can again appeal to the concept of external validity. Let's take an example, and referring to the case of its use to determine university admission. This case raises the question of standards of entry for the continuation of studies. In this respect, one can say that the risk of non-admission may lead the student to copy or study just for the test to overcome. However, the recent debate at the University of California regarding the "predictive validity" of the SAT I test (Saul & Studley, (2002)) brings evidence for that evaluation systems should produce externally valid measures, that is correlated with important facts as success in school career, the ability to acquire skills for the job, the probability of employment, income and so on. That means that if the evaluation system, instead of imposing consequences on students who did not pass a given test, should restrict the attention to the production and external validation of objective measures in basic skills, providing information to society and the different stakeholders about the correlation between these measures and important aspects of life, and the threshold levels beyond which or below which, it can be to determine a high risk of negative consequences. This information should not have negative consequences in terms of learning to the test. An example of this can be found in the case of information that the health care system provides us with regard to body weight or cholesterol levels, and their association with cardiovascular disease or other negative consequences: no one prevents us from taking food in an excessive (or students from cheating), but we must realize that there is a correlation between the level of food intake (entry level skills) and the probability of negative consequences for health (pass or be rejected for exams). Obviously here it is clear that the replacement of "consequences" (no admission if you do not pass the test) with the "advice" "Look, if you reach this score in this test, you can support in comfort at home as many times as you want, your likely to fail the exams will be 90% " should really be followed by failure, if is the case. Abolishing this "consequence" from the education system might create a system devoted only to create "diploma-papers", devoid of any real value added in terms of human capital. The effect is comparable in substance, in the long run, the creation of currency without compensation in terms of goods produced (Woessmann, (2003)). The measurement of ability in basic skills and its growth could also be extended to university, at least in disciplines of the first and second year, and could validly represent an alternative to the "customer satisfaction", which is more a satisfaction against a subject (influenced by self-selection bias), than a measure of teacher effectiveness. Another aspect of crucial importance for policies to improve learning, which derives from the Rasch model and its ability to produce comparable measures in time and linked together, being the construction of growth

Figure 3: Growth curve based on Rasch measurements



1. Aitkin M., Longford N. (1986) Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society, A*, 149
2. Battauz M., Bellio R., Gori E.:Covariate measurement error adjustment for multilevel models with application to educational data, *Journal of Educational and Behavioural Statistics*
3. Bishop J. H., Woessmann L. (2001), *Institutional Effects in a Simple Model of Educational Production*. Kiel Working Paper 1085
4. De Rozario P. (2002) Nota sintetica sulle «visite di studio» organizzate dal CEDEFOP sul tema della qualità dell'IFP nel quadro del FORUM sulla qualità dell'IFP, Gruppo tecnico, Febbraio 2002
5. Dolton P. (2002) Improving Education Quality: How Best to Evaluate Our Schools? A Discussion. Presented at "Taking Account of Accountability: Assessing Politics and Policy Program on Education Policy and Governance Education reform conference", Harvard 2002. <http://www.bos.frb.org/economic/conf/conf47/dolton.pdf>
6. Glenn C., de Groof J. (2003) Finding the right balance. Freedom, autonomy and accountability in education, Lemma, Utrecht

7. Gori, E.: Valutazione degli apprendimenti disciplinari nella scuola secondaria di primo grado, par. 2, <http://crell.jrc.ec.europa.eu/Publications/GuidoGay.pdf> (2006)
8. Gori E. (2003) Quali prospettive dalla ricerca sulla qualità e l'efficacia della scuola per la costruzione di sistemi di Accountability dell'istruzione. Non Profit, 1.2003, Maggioli ed., Rimini
9. Gori E. (2004) L'investimento in Capitale Umano attraverso l'istruzione, in G. Vittadini (a cura di) (2004) Capitale Umano. La ricchezza dell'Europa. Guerini e ass. ed.
10. Gori E., Vidoni D. (2005) Who watches the watchers? The limits of present accountability policies and a possible way out. To appear on the Boston Journal of Education
11. Gori E., Vittadini G. (2005) Sussidiarietà, Valutazione e Capitale Umano, in Non Profit, 2.2005, Maggioli ed., Rimini
12. Gori E., Plazzi G., Sanarico M. (2005) La valutazione e la misurazione nelle scienze sociali: oggettività specifica, statistiche sufficienti e modello di Rasch, in Non Profit, 2.2005, Maggioli ed., Rimini
13. Hanushek E. A.; Kimko D. D. (2000) Schooling, Labor-Force Quality, and the Growth of Nations. American Economic Review, Vol. 90, No. 5, December: 1184–208
14. Hanushek E. A. (2003) The Failure of Input-based Schooling Policies, Economic Journal, Vol. 113, No. 485, February: 64–98
15. Hanushek E. A., Raymond M. E. (2004) Does School Accountability Lead to Improved Student Performance? <http://www.nber.org/papers/w10591>, National Bureau of Economic Research
16. Linacre J. M. (1989) Many-facet Rasch measurement. Chicago: MESA Press
17. Margaret Wu, A Comparison of PISA and TIMSS 2003 achievement results in Mathematics and Science, http://www.iea.nl/fileadmin/user_upload/IRC2008/Papers/TIMSS_Mathematics/Wu.pdf
18. Messick S. (1989) Validity. In R. Linn (Ed.), Educational measurement (3rd ed., pp.13-100). Washington, DC: American Council on Education
19. Mill J. S. (1806–73) On Liberty, Chp V. The Harvard Classics. 1909–14
20. Rasch G. (1968) Mathematical theory of objectivity and its consequences for model construction. In "Report from the European Meeting on Statistics, Econometrics and Management Sciences", Amsterdam
21. Rasch G. (1977) On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements, Danish Yearbook of Philosophy, 14, 58-94
22. Rogosa D. (1995) Myths and Methods: "Myths About Longitudinal Research" plus Supplemental Questions, in Mordechai G. J. (1995) The Analysis of Change, Lawrence Erlbaum Ass., Pub. New Jersey
23. Saul G., Studley R. (2002) UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California, Educational Assessment, Vol. 8, No. 1, Pages 1-26 168
24. Woessmann L. (2003) Central Exams as the "Currency" of School Systems: International Evidence on the Complementarity of School Autonomy and Central Exams", CESifo DICE Report 1 (4), 46-56