[WORK IN PROGRES] Apr. 13 2013,

Does Testing Improve Learning?

Jakob Wandall, JW@NordicMetrics.Com

In the literature there are evidence that evaluation feedback can be a powerful performance improving mechanism (e.g. Visscher 2003). But a question that is frequently raised, yet to be answered unambiguously is if/when testing leads to better results.

A test that delivers valid, reliable and understandable measures on the individual student's academic skills and progress provides information that will improve the teacher's possibility to target instruction to this student. It is a fair assumption that better targeted instruction will lead to improved learning, more academic growth and better student achievement. Just as precise information of the patient's age, weight, temperature and blood pressure can improve the quality of the doctor's diagnosis, subsequently the treatment and hopefully lead to better health. This theory about the causality between testing and achievement is easy to understand.

Usually there is more than one student in the classroom and therefore a series of complex issues need to be discussed: Do all student benefit equally from the same instruction? What does it take for at teacher to target instruction to 20 individual students at the same time? Which factors – apart from the teachers' knowledge and skills – are influencing learning in school (e.g. school leadership, educational policy, educational materials and curriculum)? How can teachers use information from aggregated test-scores for the class to improve the instruction? – Just to mention a few.

It is not my ambition to cover all these subjects. It is more modest: To try to shed some light over the distribution and development of ability of the students in Denmark and the US. My hypothesis is

that these factors are significant in this matter: 1) If the academic level is fairly homogeneous it must be easier to provide instruction that is relevant for all students than if there are big differences between top and bottom. 2) If there is a certain level of knowledge about the deviation between top and bottom in the class, this can be taken into consideration when planning the instruction.

Reading Development in Denmark and the U.S. based on National Achievement Tests

In the following deviation of ability and the progress in reading in schools, Denmark and U.S. is compared, knowing that the differences of these countries are enormous – size, population, equality, diversity in culture and so on. Having said that, there is a specifically interesting difference between the two countries: Until 2010 there did not exist a test (apart from ILSA) in reading in Danish reporting results on an interval scale. So the Danish schools have virtually not been influenced by testing in reading. It is almost opposite in the U.S. as reading is the most tested construct, and students' reading ability probably is measured more often than their temperature, weight and height (Stenner et.al. 2006). Even though there is not sufficient foundation for a hypothesis that the difference in testing is the main explanation of the difference in reading, a comparison might give some interesting information.

In U.S., frequent NAEP surveys have been conducted since the start of the 1990's in order to monitor the achievement levels and growth in several subject areas such as reading. A lot of effort has been made to improve the U.S. reading results in the last 20 years, but the results show very limited change: 4th and 8th grade students have improved 4 and 5 scale points on the 500 point NAEP-scale from 1992 to 2011 (NAEP, 2011). The deviation in the students' performance has been measured, and it turns out that there is a large difference between the top and bottom in any school, state, and likely within any class: The best 10% of the 4th grade students obtain better reading scores than the bottom

25% students in 12th grade (see Figure 1, right side). US-schools have focus on both the high and low performing students, in Denmark the main focus is on improving the low performing students and schools. And PISA-data indicate that the distance between top and bottom in reading in U.S. is larger than in Denmark, see Table 3.

Table 3: US and Denmark, variations in PISA scores 2009

PISA score 2009	Percentiles					
	10th	25th	Mean score	75th	90th	10th-90th
Denmark	383	440	495	554	599	216
United States	372	433	500	569	625	253

Source: www.oecd .org

It would therefore be expected to find a higher degree of inequality between individual schools and students in the US than in Denmark. There has been no systematic retrieval of data on reading ability in Denmark prior to 2010; the best guess on the development in the reading level in Denmark comes from PISA. Denmark has participated though all the years, with practically no change in results (2000-03-06-09: 497-492-496-495).

In 2010 came the first results from the national tests which made it possible to get a picture of growth and deviation between top and bottom in Denmark.

Danish and US student reading performance on the same scale

The scale of the Danish National test (Rasch logits) are not directly comparable to the scale of NAEP, but since they both – like the PISA-scale - are interval scales, and the PISA-scores for US and Denmark are known (table 3), the PISA scale can be used as a common denominator.

This way both Rasch Logits and NAEP score points are converted to a common PISA-scale, see Table 4.

Figure 3. Growth and deviation. Danish national testing (DNT) 2010, reading (text comprehension) and NAEP reading 2009 (compared by PISA 2009)

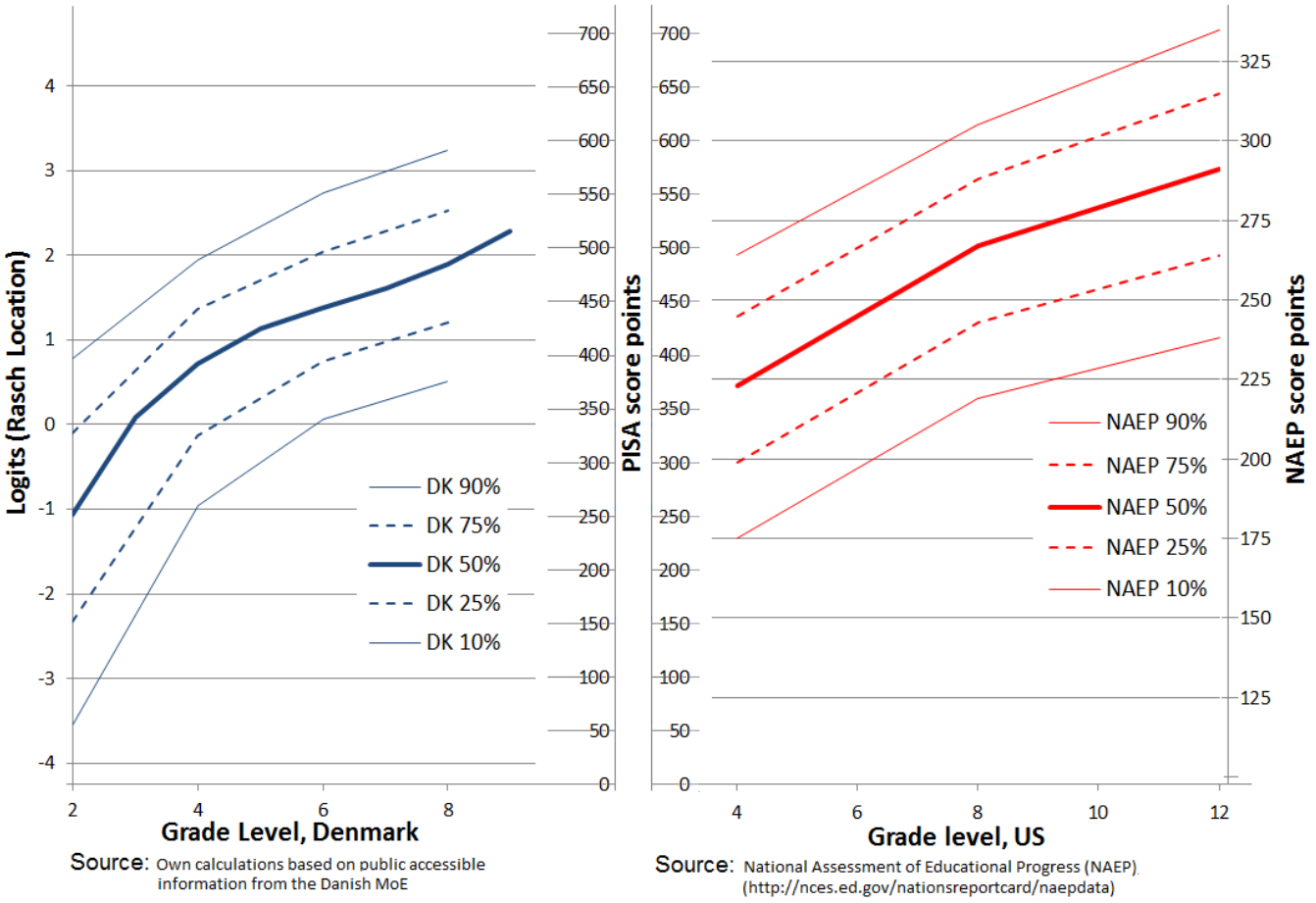

Table 4 – Test scores from Danish National Tests and NAEP transposed into a common PISA-scale (2009)

	10'th Percentile	25'th Percentile	50'th Percentile	75'th Percentile	90'th Percentile
DK, grade	Rasch-logits converted PISA-scale scores				
2	57	155	257	335	406
4	266	333	400	453	500
6	348	403	454	508	564
8	384	440	496	547	604
US, grade	NAEP-scores converted PISA-scale scores				
4	237	307	378	443	499
8	366	437	507	569	619
12	422	499	578	649	708

(Source: Same as Figure 3 combined with Table 3)

There is a general problem in of comparing reading scores in different languages (Wainer 2011) and more specifically, in the way it is done in PISA (Kreiner 2012). Apart from these reservations this simple equating/comparison could correctly be criticized for a several things: US children start earlier in school than Danish students, the test results for NAEP and the Danish national tests are for 8'th graders, the PISA results are for 15 year old people. But it provides an overview, and it looks as though the reading development in Denmark and the US is fairly similar (confirming hypothesis on comparability by Ralph, Keller & Crouse 1994).

In order to make the reading development comparable between the two countries, the difference between the 10'th and the 90'th percentile are compared to annual average growth1 in reading ability – see Table 5.

Table 5: Measuring reading ability variation in grade levels

		DK-Logits	U.S. NAEP-score
A.	Difference from 4'th to 8'th grade (50'th percentile)	1,19	44
B.	Difference 10'th-90'th percentile, 4'th grade	2,91	89
C.	Difference 10'th-90'th percentile, 8'th grade	2,73	86
		Grade levels (years)	
D.=((B+C)/2)/(A/4)	Difference from 10th to 90'th percentile	9,5	8,0

Source: (Same as Figure 3)

The last line in table 5 (the difference from 10th to 90'th percentile measured in Grade levels) is to be read like this: With the annual growth in students' ability pr. year registered between 4'th and 8'th grade, how many years should a students at the 10'th percentile go to school before he reach the reading ability of their classmates at the 90'th percentile: A limited share of difference in student performance is related to states, local communities, schools or classes – the major part of the deviation

[1] To be precise it not really growth, but an approximation to growth as the analysis is a comparison of cohorts, not a longitudinal study. Longitudinal data is not yet available in Denmark. Analysis of NAEP longitudinal shows the same tendencies as the present finding (Ralph, Keller & Crouse 1994)

is found between students in the class. In every "normal" school and class there are students both far above and far below national average. And there is evidence, that the difference between the students' growth in the best and the worst performing schools is next to nothing compared to the variance between student performance in the average class (Ralph, Keller & Crouse 1994).

The Effect of Testing on Teaching And Learning

It appears that the distance between high and low performing students is larger in Denmark (9,5 yrs.) than in the U.S. (8 yrs.) even though the distance between the 10'th and 90'th percentile for 8 grade is 15% larger in US than in Denmark. The explanation is that the annual average rate of growth in reading ability between 4th and 8th grade for U.S. students (130 PISA-score-points) is 35% higher than the similar of the Danes (96 PISA-score-points).

Local reading experts in Denmark (the reading consultants in the municipalities) have pointed out what the explanation could be: When the initial reading development (the decoding skills) is in place, the students are able to read. Then it seems as though reading activity generally declines and so does the development in reading ability. At Danish schools, teachers and administrators have not had any tests or other tools to monitor the development, so there has been a very limited knowledge on this issue until quite recently.

This data on progress (measured in grade levels) have a series of (at least from a policymaker's point of view) rather depressing implications:

- The students that starts in school below average will most likely end up below average – statistically only few crosses the 50-percentile during K12 education.

- The teacher/school has have very limited, but measurable, influence on the rate of academic growth among the students.
- State/National policy has yet to prove a measurable impact on performance. Maybe it is there, but what I have found until now is hypotheses, correlations and assumptions on causality – not evidence.
- The teacher responsible for reading instruction in practically any class has to deal with students performing far above and far below the average. The "grade level" textbook in the class is targeted to a minority in the class – maybe 1/3. The other 2/3 of the students are either not challenged or under water. Teachers, who have never tested the students (like in Denmark 2010), would not know how well the students read[2]; and without testing, the possibility to target the instruction is limited.

Getting back to the initial question: **Does Testing Improve Learning?** The analysis above indicates that between 4th and 8th grade, the average growth of students' reading achievement is 35% higher in U.S. than in Denmark. It is plausible that educational systems using scientific methods (like most states in U.S.) to keep track and to measure effect, *all other things being equal*, would outperform educational systems that do not collect data (like Denmark until recently). But all other things are far from equal. We have a hypothesis supported by data – but no nearly strong enough to call it evidence. So the answer is "definitely maybe": it's plausible that there is a connection between testing and high

[2] Experience from informal studies I performed (as employed in the Danish National School Aagency) on teachers knowledge) in 2010 demonstrated that the teachers perception of the students ability is based on assessment of the observable, productive skills (like speech and writing) and to a certain extend Decoding (by monitoring reading aloud). As comprehension is a receptive and not observable skill the natural assumption – if no other data is available - is that the student comprehend at the same level as they decode/speak/write. This assumption is often erroneous even though these skills usually are correlated to some extent.

academic performance; but the causal structure is not clear (if one is causing the other or if they are both effect of other common traits).

It is, however, evident that knowledge on how the individual student is doing can be beneficial for the work of the teacher. And a school leader monitoring student results and giving constructive feedback to the teacher can have a strong positive effect. At the same time there have been examples on misuse of testing, creating mistrust that could work counterproductive, depending on the culture and values.

Concluding remarks

In Denmark formal testing is a fairly new experience to a lot of teachers. There was a lot of anxiety when national testing was introduced in 2010. The Netherlands have much of the same culture as Danes, but the Dutch schools are used to testing and tests are used for both pedagogical and accountability purposes. In the U.S. there is a stronger focus on accountability and formal testing is high stakes. How well a student does in the test means a lot to the future possibilities. And as a teacher, your student's results could mean the difference between a bonus and being fired.

Looking at regulations there is a big difference in the reason on why you go to school. In the US it is mainly for cognitive achievement (Heckman 2008), but there is a trend I research pointing towards more focus on non-cognitive outcomes (Levin 2012, Heckman 2011). This has only had a minor impact on the school systems in U.S. yet. The tools for measuring non-cognitive abilities and personal traits in education are not that well developed yet. And in a society and an educational system where accountability are the top priority, it seems hard to imagine that non-measurable abilities will play a big role if they can be measured on a scale.

[WORK IN PROGRES] Apr. 13 2013,

In Denmark you are also supposed to learn how to read and write, but for a purpose: To develop creative, democratic, imaginative and critical citizens; Non-cognitive skills that always have been considered for important outcome of school the last 200 years. The problem in Denmark is that the basic cognitive skills (math, reading etc.) have not been in focus until 1994 (publications of the first ILSA where Denmark performed below expected). When the student can read reading is developing thru practice and Danish students are reading less and less year after year (according to PISA surveys on reading habits). This could be the explanation low academic growth in Denmark compared to U.S.

We have different cultures, values, strengths and weaknesses: Even that we cannot copy each other's solutions, we can learn a lot from each other, international cooperation and a comparative approach.

References

Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New York: Harper and Row.

Esping-Andersen, G. (1990). The three worlds of welfare capitalism. Princeton, NJ: Princeton University Press.

Feuer, M. J. (in press). International large-scale assessment: Validity in presentation and use. In M. Chatterji (Ed.), Validity and test use: An international dialogue on educational assessment, accountability, and equity (pp. xxx-xxx). Bingley, UK: Emerald Group Publishing Limited.

Feuer, M. J. (2012). No country left behind: Rhetoric and reality of international large-scale assessment. (The 13th William H. Angoff Memorial Lecture), Princeton, NJ: ETS.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), Reading comprehension difficulties: Processes and intervention (pp. 1–13). Mahwah, NJ: Erlbaum.

Hatch, Thomas 2012, RESPONSIBILITY AND ACCOUNTABILITY IN CONTEXT, Teachers College, Columbia University (in print) 2012

Hattie, J. (2003). Teachers make a difference: What is the research evidence? Retrieved from http://www.acer.edu.au/documents/RC2003_Hattie_TeachersMakeADifference.pdf

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. London: Routledge.

Heckman, J. J. (2008). Schools, skills and synapses. Economic Inquiry, 46 (3), 289-324.

Heckman, J. J. (2011). Integrating personality psychology into economics. (Working Paper No. 17378). Retrieved from http://www.nber.org/papers/w17378

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. (2nd ed.) Thousand Oaks, CA: Sage Publications.

Hofstede, G. (2002). The pitfalls of cross-national survey research: A reply to the article by Spector et al. on the

psychometric properties of the Hofstede values survey module, Applied Psychology, 51 (1), 170–173.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and Writing, 2(2), 127-160.

Inglehart, R. (1990). Culture shift in advanced industrial society. Princeton, NJ: Princeton University Press.

Inglehart, R. (1997). Modernization and postmodernization: Cultural, economic and political change in 43 societies. Princeton, NJ: Princeton University Press.

Inglehart, R., & Basanez, M., & Moreno, A. (1998). Human values and beliefs: A cross-cultural sourcebook. Ann Arbor, MI: University of Michigan Press.

Inglehart, R., & Welzel, C. (2005). Modernization, cultural change and democracy. New York: Cambridge University Press.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. Quarterly Journal of Economics, 118, 843-878.

Kreiner, S. (2011). Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. Copenhagen: University of Copenhagen, Department of Biostatistics.

Levin, H. M. (2012). More than just test scores. Quarterly Review of Comparative Education, 42(3), 269-284.

Lundahl, C. (2009). Varför nationelle prov? – framväxt, dilemman, möjligheter. Sweden, Lund: Studenterlitteratur AB

Maseland, R., & van Hoorn, A. (2009). Measuring values for cross-cultural research (NiCE Working Paper 09-107) Retrieved from www.ru.nl/publish/pages/516298/nice_09107.pdf

NAEP. (2011). NAEP reading assessment 2011. Retrieved from http://nces.ed.gov/nationsreportcard/reading/

OECD. (2010). PISA 2009 results: Learning trends - Changes in student performance since 2000. Volume V. Paris: Organization for Economic Cooperation and Development

OECD. (2004). OECD-rapport om grundskolen i Danmark - 2004 - Uddannelsesstyrelsens temahæfteserie nr. 5 – 2004. Retrieved from http://pub.uvm.dk/2004/oecd/

OECD-PISA (2013). Databases for PISA 2000-2009. Retrieved from http://www.oecd.org

Ralph, J., Keller, D., & Crouse, J. (1994). How effective are American schools? The Phi Delta Kappan, 76(2), 144-150.

Rambøll. (2011, November). Evaluering af mere frit skolevalg (2.0). Report for the Ministry of Education. Retrieved from http://uvm.dk/Om-os/Ministeriet/Kontakt/Presserum/~/UVM-DK/Content/News/Udd/Folke/2012/Feb/~/media/UVM/Filer/Udd/Folke/PDF12/120221%20Frit%20skolevalg%20rapport.ashx

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: DPI.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2009). ICCS 2009 international report: Civic knowledge, attitudes and engagement among lower secondary school students in thirty-eight countries. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Shulman, L.S. (2007). Counting and recounting: Assessment and the quest for accountability. Change, 39(1), 20-25

Snook, I., Clark, J., Harker, R., O'Neill, A., & O'Neil, J. (2009). Invisible learnings? A commentary on John Hattie's Book: Visible learning: A synthesis of over 800 Meta-analyses relating to achievement. New Zealand Journal of Educational Studies, 44(1), 93-106

Spector, P. E., Cooper, C. L., & Sparks, K. (2001). An international study of the psychometric properties of the Hofstede Values Survey Module 1994: A comparison of individual and country/province level results. Applied

Psychology, 50(2), 269-281.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are lexile text measures? Journal of Applied Measurement, 7(3), 307-322.

Suen, H. K., & French, J. L. (2003). A history of the development of psychological and educational testing. In C. R. Reynolds & R. Kamphaus (Eds.) Handbook of Psychological and Educational Assessment of Children: Intelligence and Achievement. New York: Guilford.

UK Department for Education. (2011). Independent Review of Key Stage 2 testing, assessment and accountability (Government Response). London, UK: Department for Education.

Visscher, A. J. (in press). Evaluation-centered school improvement: Potential, prerequisites, and validity considerations. In M. Chatterji (Ed.), Validity and test use: An international dialogue on educational assessment, accountability, and equity (pp. xxx-xxx). Bingley, UK: Emerald Group Publishing Limited.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. School Effectiveness and School Improvement, 14, 321–349.

Wainer, H. (2011). Uneducated guesses: Using evidence to uncover misguided education policies. Princeton, NJ: Princeton University Press.

Wandall, J. (2009). CAT as a Pedagogic Tool. In *The transition to computer-based assessment - New approaches to skills assessment and implications for large-scale testing*. (Eds. Scheuermann & Björnsson) Luxembourg: Office for Official Publications of the European Communities, EUR – Scientific and Technical Research series, 2009: 45-50

Wandall, J. (2011). National tests in Denmark – CAT as a pedagogic tool. Journal of Applied Testing Technology, vol.12, May 2011

Zuckerman, P. (2009). Why are Danes and Swedes so irreligious? Nordic Journal of Religion and Society, 22(1): 55–69