Can We Get Better Assessment From A Tutoring System Compared to Traditional Paper Testing? Can We Have Our Cake (Better Assessment) and Eat It too (Student Learning During the Test)?

Mingyu Feng[1], Neil Heffernan[2]
{mingyu.feng@sri.com, nth@wpi.edu}
1 SRI International, Menlo Park, CA 94025
2 Worcester Polytechnic Institute, Worcester, MA 01609

Abstract. Dynamic assessment (DA) has been advocated as an interactive approach to conduct assessments to students in the learning systems as it can differentiate student proficiency at a finer grained level. Sternberg and others have been pursuing an alternative to IQ tests. They proposed to give students tests to see how much assistance it takes a student to learn a topic; and to use as a measure of their learning gain. They referred to this as dynamic assessment. It was suggested that this assisting-while-testing procedure could be done well by computer. To researchers in the intelligent tutoring system community, it comes as no surprise that measuring how much assistance a student needs to complete a task successfully is probably a good indicator of this lack of knowledge. However, a cautionary note is that conducting DA takes more time than simply administering regular test items to students. In this paper, we report a study analyzing 40-minutes data of 1,392 students from two school years using educational data mining techniques. We compare two conditions: one contains only practice items without intervention while the other condition allows students to seek for help when they encounter difficulties. The result suggests that for the purpose of assessing student performance, it is more efficient for students to take DA than just having practice items.

Keywords: Dynamic assessment, assessment in learning system.

1 Introduction

In the past twenty years, much attention from the Intelligent Tutoring System (ITS) community has been paid to improve the quality of student learning while the topic of improving the quality of assessment has not been emphasized as much. However, student assessment is very important. In the US, state tests mandated by "No Child Left Behind" are causing many schools to give extra tests to see if they can group students together to get special help. Of course, giving tests for this practices is not meant to help students learn, but is mainly focused on being able to tell teachers and principals about who needs help on what. It would be great if intelligent tutoring systems could be used to do the tests, so that no time from instruction is "stolen" to do extra assessments. Many psychometricians would argue that let students learn while being tested will make the assessment harder since you are trying to measure a moving target. Can an ITS, if given the same amount of time, be a better assessor of students (while also of course providing the benefit of helping students learn during that time period. Is it possible to have our cake (better assessment) and eat it too (also let student learn)?

As an intelligent tutoring system adapts the educational interaction to the specific needs of the individual student, student modeling is an essential component in an ITS as well. The learning effectiveness depends heavily on the understanding of student knowledge, difficulties, and misconceptions. Yet, assessing students automatically, continuously and accurately without interfering with student learning is an appealing but also a challenging task.

Dynamic assessment (DA, or sometimes called dynamic testing, Grigorenko & Sternberg, 1998) has been advocated as an interactive approach to conducting assessments to students in the learning systems as it can differentiate student proficiency at the finer grained level. Different from traditional assessment, DA uses the amount and nature of the assistance that students receive which is normally not available in traditional practice test situations as a way to judge the extent of student knowledge limitations. Even before the computer supported systems become popular, much work has been done on developing "testing metrics" for dynamic testing (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2001, 2002) to supplement accuracy data (wrong/right scores) from a single setting. Researchers have been interested in trying to get more assessment value by comparing traditional assessment (static testing; students getting an item marked wrong or even getting partial credit) with a measure that shows how much help they needed. Grigorenko and Sternberg (1998) reviewed relevant literature on this topic and expressed enthusiasm for the idea. Sternberg & Grigorenko (2001, 2002) argued that dynamic tests not only serve to enhance students' learning of cognitive skills, but also provide more accurate measures of ability to learn than traditional static tests. Campione and colleagues (Bryant, Brown & Campione, 1983; Campione & Brown, 1985) took a graduated prompting procedure to compare traditional testing paradigms against a dynamic testing paradigm. In the dynamic testing paradigm, learners are offered increasingly more explicit prewritten hints in response to incorrect responses. In this study they wanted to predict learning gains between pretest and posttest. They found that student learning gains were not as well correlated ($R = 0.45$) with static ability score as with their "dynamic testing" ($R = 0.60$) score. They also suggested that this dynamic method could be effectively done by computer, but never pushed toward to conduct such studies using a computer system.

Intelligent Tutoring Systems are perfect test beds for DA as they naturally lead students into a tutoring process to help students with the difficulties they have encountered. Traditional paper and pencil or even some online assessment usually focuses on students' responses to test items and whether they are answered correctly or incorrectly. It ignores all other student behaviors during the test (e.g., response time). However, the most unique information from DA is information about the learner's responsiveness to intervention (Fuches et al. 2007) in the tutoring system. There have been a few studies that pay attention to such unique information. For instance, recently Fuches and colleagues (Fuches et al., 2008) employed DA in predicting third graders' development of mathematical problem solving. We (Feng, Heffernan & Koedinger, 2006, 2009) have also taken advantage of a computer-based tutoring system (ASSISTments, www.assistment.org, Razzaq et al., 2005), to collect extensive information while students interact with the system. Our results showed that the assistance model that includes no assessment result on the main problems leads to significantly better predictions than the lean model that is based on the assessment results alone. This relative success of the assistance model over the lean model highlights the power of the assistance measures, which suggests not only is it possible to get reliable information during "teaching on the test", but also data from the teaching process actually improves reliability.

Although DA has been shown to be effective predicting student performance, yet there is a cautionary note about DA since students are allowed to request assistance: it generally takes longer for students to finish a test using the DA approach than using a traditional test. For instance, in Feng et al. (2009) we reported that we could do a better job predicting student state test score using DA than a contrast case, the traditional testing situation. However, there is a caveat that the DA condition has included more time than the contrast case, which seems unfair for the contrast case. Although this sort of contrast leaves out the instructional benefit (e.g., Razzaq & Heffernan, 2006, 2007; Feng, Heffernan, Beck & Koedinger, 2008) of the tutoring system and, moreover, may not be well received by teachers and students, whether or not the system using DA would yield a better prediction of state scores or learning is still worth of further research. In this paper, we report a study that aims to answer this question.

2 Methods

2.1 ASSISTments, the test bed

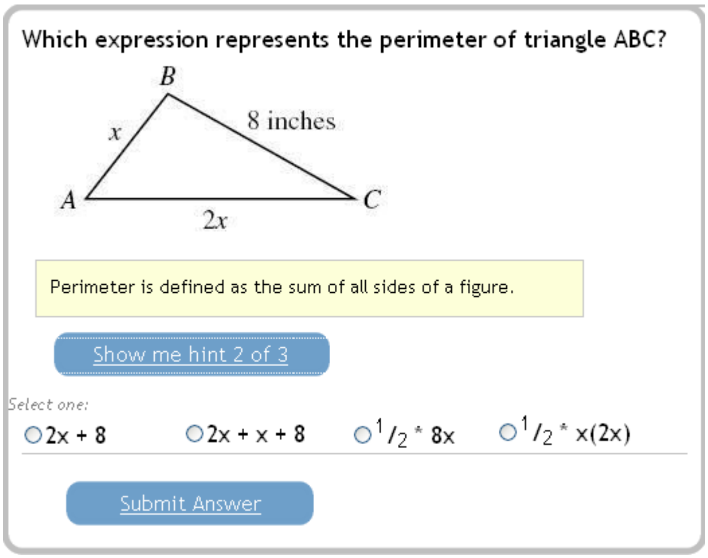

Fig.1. A screenshot showing student requested a hint for one scaffolding question in ASSISTments

Traditionally, the areas of testing (i.e. psychometrics) and instruction (i.e., math educational research and instructional technology research) were separated fields of research with their own goals. The ASSISTments system is an attempt to blend the positive features of both computer-based tutoring and benchmark testing. The online system presents math problems to students of approximately 13 to 16 years old in middle school or high school to solve. If a student gets an item (the **main** item) right, they will get a new item. If a student has trouble solving a problem, the system provides instructional assistance to lead the student through by breaking the problem into a few **scaffolding** steps (typically 3~5 per problem), or displaying **hint** messages on the screen (usually 2~4 per question), upon student request as shown in Fig.1. Although the system is web-based hence accessible in principle anywhere/anytime, students typically interact with the system during one class period in the schools' computer labs every three or four weeks. As students interact with the system, time-stamped student answers and student actions are logged into the background database. The hypothesis is that ASSISTments can do a better job of assessing student knowledge limitations than practice tests or other online testing approaches by using the DA approach based on the data collected online.

2.2 Approach

Fundamentally, in order to find out whether DA was worth the time, we would want to run a study comparing the assessment value of the following two different conditions:

- Static assessment condition (A): students were presented with one static (as opposed to dynamic) test item and were requested to submit an answer. Once they had done that, more static items followed.
- Dynamic assessment condition (B): students were presented with one static test item followed by a DA portion where they could request help.

Then the question was: Is condition B better, or at least as good considering the learning effect, at assessing students after we control the time?

We could have conducted a randomized controlled experiment with the two conditions. But, since the logging system of ASSISTments had collected data with the information needed by DA, we chose to compare predictions made based on log data from 40 minutes of time across *simulated* conditions that were similar but not exactly the same as above:

- **Simulated static assessment condition (A'):** 40 minutes of student work selected from existing log data on only main items
- **Dynamic assessment condition (B'):** 40 minutes of work selected from existing log data on both main items and the scaffolding steps and hints

Such a simulation study using educational data mining techniques not only saved time from setting up and carrying out classroom experiments, but also allowed us to compare the same student's work in two different conditions, which naturally rules out the subject effect. There would be no threat to validity of the comparison as both A' and B' allow learning on the test so there was a general trend up that you would expect[1]. So we will not devote much attention to the learning value of these conditions. We will refer interested readers to our previous publications where there were evidences showing that the system led to better student learning (e.g. Razzaq & Heffernan, 2006, 2007; Feng, Heffernan, Beck & Koedinger, 2008).

We chose to use student's end of year state accountability test score as the measure of student achievement, and we used data from conditions A' and B' to predict state test scores and compare the predictive accuracy of the two conditions.

[1] Students in the static condition A' potentially could have spent more time in the system considering the tutoring portion following main items.

2.3 Data

The first raw data set we considered came from the 2004 – 2005 school year, the first full year in which the ASSISTment system was used in classes in 2 middle schools in Massachusetts. 912 8th grade students' logs were maintained in the system over the time period from September to May. Among these students, we were able to obtain complete data for 628. The data set contained online interaction data from the ASSISTment system and the results of 8th grade state tests taken in May 2005. Students whose state test scores were not available and those who had done less than 40 minutes of work were excluded.

The second raw data set we used was from the 2005-2006 school year. We collected a full data set for 764 students from Worcester Public Schools, including the online data from ASSISTments and their 8th grade state test raw scores. We applied the same filter to exclude students who had not done enough work. In both years, the items involved in the data sets were given to students in a random order.

For each of the two raw data sets, we prepared two data sets for analysis, one for simulated static assessment condition (A') and one for dynamic assessment condition (B'). The data for condition A' included student response data during the first 40 minutes of work on only main problems; all responses and other actions during the DA portion were ignored. On the contrary, the data for condition B included all the responses for main questions and scaffoldings, as well as hint requests. For instance, consider the following scenario:

> *Chris spent one minute trying to answer a main question in ASSISTments but failed, and was forced into the tutoring session. Chris then spent four minutes working through the three scaffolding questions. Chris answered one scaffolding question correctly and requested hints for the other two.*

This scenario counted as 1 minute of static work among the 40 minutes of data we prepared for condition A' with a response to the main question being recorded as zero. Yet it counted as 5 minutes of dynamic work in the data for condition B', including 1 correct response to scaffolding, 2 incorrect responses to scaffolding and 2 hint requests.

2.4 Metrics

We followed our work in Feng, Heffernan & Koedinger (2006, 2009) of developing online metrics for dynamic testing that measures student accuracy, speed, attempts, and help-seeking behaviors. Simply, the metrics we picked were

- Main_Percent_Correct[2] – students' percent correct on main questions, which we often referred to as the "static metric".
- Main_Count - the number of main items students completed. This measures students' attendance and how on-task they were. This measure also reflects students' knowledge since better students have a higher potential to finish more items in the same amount of time. This is especially true for condition B'

[2] Student requesting for help before attempting a problem are considered as making an incorrect response to the main question.

where students' work on scaffolding also counted as part of the 40 minute work. While in condition A', low performing kids could go through many items but give wrong answers since their time consumed during the tutoring session is disregarded.

- Scaffold_Percent_Correct - students' percent correct on scaffolding questions. In addition to original items, students' performance on scaffolding questions was also a reasonable reflection of their knowledge. For instance, two students who get the same original item wrong may, in fact, have different knowledge levels and this may be reflected in that one may do better on scaffolding questions than the other.
- Avg_Hint_Request - the average number of hint requests per question.
- Avg_Attempt - the average number of attempts students made for each question.
- Avg_Question_Time - on average, how long it takes for a student to answer a question, whether original or scaffolding, measured in seconds.

The last five metrics are DA style metrics and were not measured in traditional tests. They indicate the amount of assistance students needed to finish problems and the amount of time they needed to finish the questions. Our hypothesis is that the last three metrics will be negatively correlated with students' performance. Thereby, the more hints they request, the more attempts they make on a question and the longer they need to go through a question, the worse their performance.

Among the above six metrics, condition A' used only the first one as predictor to simulate paper practice tests by scoring students either correct or incorrect on each main problem while condition B' used all the metrics.

2.5 Modeling

We ran stepwise linear regression[3] to use the metrics described above to predict student state test scores. The same process was repeated on the second year's data. For all the models, the dependent variable is the state test score but the independent variables differ. Specifically, for condition A', the independent variable of the simple linear regression model is *Main_Percent_Correct;* while for condition B', it changed to be a collection of metrics: *Main_Percent_Correct, Main_Count, Scaffold_Percent_Correct, Avg_Hint_Request, Avg_Attempt, Avg_Question_Time.*

2.6 Results

First, we noticed that in both years, students finished more test items in the 40 minutes in static condition than in dynamic condition, which is not surprising considering the tutoring portion in the DA condition. Particularly, in year 2004-2005, the average number of main items finished was 22 in the simulated static assessment condition while it was only 11 in the dynamic condition; in year 2005-2006, the number was 31 in the static condition but it was only 13 in the dynamic condition.

[3] probability of F-to-enter <= .05, probability of F-to-remove >= .10

Then, we examined the parameters and associated coefficients in the linear regression models of both conditions.

Table 1. Parameters of simple regression models for simulated static assessment condition (A')

Condition A'	Parameter	Coefficient
2004-2005	Intercept	16.383
	Main_Percent_Correct	24.690
2005-2006	Intercept	13.993
	Main_Percent_Correct	40.479

As shown in Table 2, the first three parameters entered the models were the same in both years (with the order changed a little bit). Scaffold_Percent_Correct was the most significant predictor in the first year while in the second year, it changed to be Main_Percent_Correct. Also, in the later year 2005-2006, Avg_Attempt was considered as a significant predictor while in the first year it was Avg_Hint. Yet, it was consistent with our hypothesis that more attempts or more hints on a question will end up with a lower estimated score.

Table 2. Parameters entered regression models of dynamic condition (B')

Condition B' (2004-2005)	Parameter	Coefficient
0	Intercept	17.090
1	Scaffold_Percent_Correct	16.311
2	Main_Percent_Correct	7.107
3	Main_Count	0.179
4	Avg_Hint	-2.580

Condition B' (2005-2006)	Parameter	Coefficient
0	Intercept	16.061
1	Main_Percent_Correct	21.331
2	Scaffold_Percent_Correct	16.242
3	Main_Count	0.172
4	Avg_Attempt	-2.543

Now that we had looked at the parameters in the regression models, we would examine which condition does a better job predicting state test score. The R square's of all models were summarized in Table 3. Additionally, because the models in different conditions always had different numbers of parameters, we also chose to use Bayesian Information Criterion (BIC) to compare the generalization quality of the models. We applied the formula for linear regression models introduced by Raftery (1995, p135), which was different from what is typical used for calculating BIC but most convenient for linear regression models:

$$BIC = n*ln(1-R^2) + p*ln(n)$$

where

> *n*: the sample size (for the 2004-2005 data case, n = 628; for the 2005-2006 data, n=764)
> *ln*: natural logarithm

p: the number of independent variables included in each model (not including intercept)

Table 3. Summary of models

	R^2		BIC	
	2004-2005	2005-2006	2004-2005	2005-2006
Simulated static condition	0.174	0.377	-114	-354
Dynamic condition	0.240	0.426	-147	-398

As we can see from Table 3, in both years, the R square of the model from the dynamic condition was always higher than that of the simulated static condition. Raftery (1995) discussed a Bayesian model selection procedure, in which the author proposed the heuristic of a BIC difference of 10 was about the same as getting a p-value of 0.05. And the lower BIC indicated a better fitted model. Thereby, we can see, in both years, the dynamic assessment condition did a significantly better job at predicting state test scores than the control condition which is static.

2.7 Validation

Before jumping into the conclusion saying dynamic assessment is more efficient than just giving practice test items, we performed 5-fold cross validation on the 2004-2005 data set. For the testing data, we calculated mean absolute difference (MAD) as a measure of prediction accuracy, which was computed as the average of the absolute difference between students' real state test scores and the predicted scores across all students included in the testing set.

Table 4. Results of cross validation

	MAD					
Fold	**1**	**2**	**3**	**4**	**5**	**Avg.**
Simulated static condition	9.44	9.05	8.67	9.01	9.13	9.01
Dynamic condition	9.02	8.65	8.74	9.04	8.57	8.7
p-value (95%) from two-sided paired t-test comparing absolute difference of two conditions	0.35	0.36	0.88	0.94	0.13	**0.10**

As illustrated in Table 4, out of the 5 folds, DA condition ended up with a lower MAD in 3 folds. On average, DA condition did a better job predicting state test scores in the testing set: The difference between MADs of the DA condition and simulated static condition was bigger in these 3 DA-winning folds, and it was much smaller in the other 2 folds (folds 3 &4). Even though, the results from two-sided paired t-test indicated none of the difference was statistically significant.

Then we took a closer look to see whether the trained regression models of the DA condition were consistent across the 5 folds validation. We found out that the trained models were fairly stable. The four variables as shown in Table 2 (2004-2005 portion), Scaffold_Percent_Correct, Main_Percent_Correct, Main_Count, Avg_Hint entered all five trained models while no other variables have been selected. Scaffold_Percent_Correct was always the most significant predictor across all folds while the entering order of the other variables varied during the stepwise variable selection process. The associated coefficients of the selected variables differed across folds with variance ranging between 0.0 (Main_Count) and 2.4 (Scaffold_Percent_Correct).

As the last step, we took the average of coefficients from the five trained models and applied the model on the full data set of year 2004-2005. The average model from the simulated static condition and the DA condition produced MAD of 9.01 and 8.7 respectively. The paired t-test suggested that there was a marginally significant difference (p=0.10)

All in all, based on the results, we conclude that dynamic assessment is more efficient than just giving practice test items. So, not only that students are learning during DA but also DA can produce at least as accurate assessment of student math performance as traditional practice test, even limited by using the same amount of testing time.

This is surprising as students in the dynamic assessment do few problems and yet we get better assessment results. Of course, DA has another major advantage in that kids are learning during the test and therefore are not wasting their time just testing, while the practices tests are not likely to lead to much learning.

3 Conclusion

Dynamic assessment (DA) has been advocated as an interactive approach to conducting assessments to students in the learning systems as it can differentiate student proficiency at the finer grained level. In this paper, we compare dynamic assessment against a tough contrast case where students are doing assessment all the time in order to evaluate efficiency and accuracy of dynamic assessment in a tutoring system.

Contribution: The contribution of this paper lies in that it eliminates the cautionary note about dynamic assessment that says DA will always need a longer time to do as well at assessing students, which further validates the usage of tutoring systems for assessment. ITS researchers have showed the effectiveness of such systems at promoting learning (e.g. Koedinger et al., 1997). This paper adds to that fact and presents a nice result suggesting that maybe, students should take their tests in an ITS as well!

General implication: Combining with our previous findings (Feng, Heffernan & Koedinger, 2006, 2009), this paper tells us that not only we can better assess students while teaching them, but also the assessment can be done efficiently. Our results are important because they provide evidence that reliable and efficient assessment and instructional assistance can be effectively blended. At the *Race to The Top Assessment Competition* public input meetings, experts advocated for computer-based state assessments and argued the tests should be taken more than once a year (U.S. Dept of Ed., 2009). The general implication from this series research suggests that such computer-based, continuous assessment systems are possible to build and that they can be quite accurate and efficient at helping schools get information on their students while allowing student learning at the same time.

Acknowledgements

We would like to acknowledge funding from the Institute of Education Sciences of US Department of Education, the National Science Foundation, the Office of Naval

Research and the Spencer Foundation. All of the opinions expressed in this paper are those solely of the authors and not those of our funding organizations.

References

1. Brown, A. L., Bryant, N.R., & Campione, J. C. (1983). Preschool children's learning and transfer of matrices problems: Potential for improvement. Paper presented at the Society for Research in Child Development meetings, Detroit.
2. Campione, J.C., Brown, A.L., & Bryant, N.R. (1985). Individual differences in learning and memory. In R.J. Sternberg (Ed.). *Human abilities: An information-processing approach*, 103–126. New York: W.H. Freeman.
3. Campione, J.C.& Brown, A.L. (1985). Dynamic assessment: One approach and some initial data. Technical Report No. 361. Cambridge, MA: Illinois University, Urbana. Center for the Study of Reading. ED269735
4. Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*. 19(3), 2009.
5. Feng, M., Heffernan, N., Beck, J, & Koedinger, K. (2008). Can we predict which groups of questions students will learn from? In Beck & Baker (Eds.). *Proceedings of the 1st International Conference on Education Data Mining.* Montreal, 2008.
6. Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Addressing the testing challenge with a web based E-assessment system that tutors as it assesses. *Proceedings of the 15th Annual World Wide Web Conference.* ACM Press: New York.
7. Fuchs, L.S., Compton, D.L., Fuchs, D., Hollenbeck, K.N., Craddock, C.F., & Hamlett, C.L (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. Journal of Educational Psychology, 100(4), 829-250.
8. Fuchs, D., Fuchs, L.S., Compton, D.L., Bouton, B., Caffrey, E., & Hill, L. (2007). Dynamic assessment as responsiveness to intervention. Teaching Exceptional Children, 39 (5), 58-63.
9. Grigorenko, E. L. and Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124,* 75–111.
10. Koedinger, K. R., Anderson, J. R.., Hadley, W. H.., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30-43.
11. Raftery, A. E. (1995). Bayesian model selection in social research. In *Sociological Methodology*, 25, 111-163.
12. Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., & Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G.McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th Artificial Intelligence in Education.* Amsterdam: ISO Press. pp. 555–562.
13. U.S. Dept of Ed. (2009). Transcripts from U.S. Department of Education, Race to the Top Assessment Competition public and expert input meetings. Retrieved Jan, 2010 from http://www.ed.gov/programs/racetothetop-assessment/boston-transcript-2.pdf.
14. Razzaq, L., Heffernan, N.T. (2006). Scaffolding vs. hints in the Assistment System. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 635-644. 2006.
15. Razzaq, L., Heffernan, N.T., Lindeman, R.W. (2007). What level of tutor interaction is best?. In Luckin & Koedinger (Eds.). *Proceedings of the 13th Conference on Artificial Intelligence in Education*. Amsterdam, Netherlands: IOS Press.
16. Sternburg, R.J., & Grigorenko, E.L. (2001). All testing is dynamic testing. *Issues in Education*, 7, 137-170.
17. Sternburg, R.J., & Grigorenko, E.L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, England: Cambridge University Press.