and also used to estimate the level of impact of educational policies.

Absolutely. All these can be made possible through standard setting.

Anani Sarab: By the way you referred to the fact thaf students who are lagging behind can be identified through formative assessment. Are there any arrangements in schools to help teachers support this group of students?

Yes, formative assessment is the best way to identify students who need more attention academically. They can also be referred to the school counselors to determine their needs and then schools provide tutorial service for them. Through the counselor support the teacher herself knows that she should pay close attention to these students.

Anani Sarab: The procedures mentioned can change education to a very costly endeavor. Do you think of any strategy that can help education systems to go through these processes with less cost?

We can do it with the minimum level of expenses. It can be applied in a step-by-step manner to reduce the expenses. The minimum that can be done initially is that teachers can be helped and supported to do the formative assessment and send notes to parents that these students are not progressing as expected. The formative assessment can be arranged in such a way that students who are lagging behind can be identified and the cost for this step is quite affordable. The important point is that formative assessment should enable teachers to know where the students are failing and where they are making progress.

many failures he or she will be fired. So the teachers have lots of responsibilities in raising the standards of achievement. They have to understand the criteria and the policies of educational assessment. Making schools to try their best is dependent on meaningful assessment. Without meaningful assessment accountability makes no sense. The number of failures can signal the level of work done and the steps that should be taken during the teaching period to compensate for the shortcomings. The teachers are supposed to monitor progress through identifying which students are lagging behind. They bring attention to those students in time to help them bridge their learning gaps. I received a grant from the National Science Education foundation for 3.5 million to reassess formative assessment and educate teachers to know how to identify the students who are lagging behind before it is too late. We are supposed to report the changes made to the National Science Foundation. The formative assessment that teachers have to do is intended to identify the students who are lagging behind and provide support for them to become proficient.

Anani Sarab: Based on what you said we can conclude that formative assessment is aligned to summative assessment.

Exactly. The results of summative assessment are too little and too late. They come when classes are over. It is through formative assessment that teacher can identify the students who are lagging behind and they can do something for them before it is too late. The students who need more attention or their learning rate and should be accommodated are identified through formative assessment.

Anani Sarab: All this process that you described is formed in response to education policies. How do they make sure that the policies are established and how do they maintain them?

The Federal government set laws and regulations such as the No Child Left Behind policy and Every Students Succeed Act (ESSA). The states and schools have to obey and follow these policies which have changed into laws and regulations. For example, NCLB entitles students to take ELP tests. The most interesting thing in the educational system of US is that all the scores throughout the United States are roughly comparable. And this is made possible through standard setting, and standardization of assessment. With no standardization, a score of 20 in one school might mean 10 in another school. When assessment is based on raw scores, assessment lacks comparability. Here in the US, all schools have to follow the same assessment criteria.

Anani Sarab: Do you think that the university entrance exam can compensate for the lack of comparability of raw scores?

I have a lot of issues and problems related to the university entrance examinations in Iran and many other countries. They are not based any content standards. They are not based on any sound criteria of assessment. So as was mentioned before, assessment should be based on setting standards to identify those who are proficient and above proficient in order to make meaningful decisions about the students' future academic career.

Anani Sarab: This type of assessment can also be linked to teacher appraisal

are considered proficient. So the major question is: At what mark the students reach the level of proficient, at what they reach above proficient, etc. For example, the judge determines the student who reach to number 25 as below proficient, to number 35 as proficient and number 60 as above proficient based on the order of the items in the ordered test booklet. This procedure is called Bookmark which is a very commonly used approach in the United States and many other countries. Bookmark and Mapmark are the two most commonly-used approaches in setting standards. So setting standards is exactly this. A group of judges sit round 6 to 8 tables in groups of 10. They determine how many items with which level of difficulty the students have to know in order to labeled as proficient. So the judgment is not based on raw scores. It is based on item function and item content. So it is not a score. It is just achievement level. And the achievement level is one of five or six categories: well below proficient (1) below proficient (2) proficient (3) above proficient (4) and well above proficient (5). So the students are not assessed based on test scores but based on level of proficiency.

Anani Sarab:So in this way the scores are made meaningful because they can indicate the students' competency level.

Yes, they make the score meaningful and the judgment is based on some criteria not based on comparison. The assessment does not differ depending on which group of students in which school or class are assessed. The students' achievement is based on content. It is used for several years for a group of students.

Anani Sarab: How do they maintain the level of difficulty year by year?

Usually it is very expensive to create an assessment with the same level of difficulty across years. The assessment that I've been describing cost millions of dollars. They keep the produced assessment for at least two or three years. As soon as they see the necessity of changing the assessment they do standard setting again. Even if they make minor changes in the tests they repeat the standard setting. The assessment is expensive as test development and standard setting include complicated procedures run by advisory groups, item writers, supervisors and expert judges. Tests have to go through several stages before they are made ready for the students to take and judgments are made.

Anani Sarab: Spending that amount of money on assessment should have political implications. How can the political side of assessment be explained?

Accountability is the political aspect of test development. Schools have to produce certain number of students at or above proficient. If they don't then they cut their budgets. A certain percent of the students have to reach the level of proficient for a school to receive the full budget. Schools set some criteria for teachers to do their best in order to achieve the targets. If a teacher has too

based on content and psychometric properties of the items. If an item has psychometric problems, for example one distractor does not function well, they remove it. Or for example if an item has been responded by very few test takers , such an item does not have good discrimination power. The same is true for an item that has been responded almost by everyone. So the item difficulty index should be around 0.50 for an item to have good discriminatory power. The point biserial correlation should be above 0.30 and 0.40. IRT analysis be should accessible. As you can see a lot of analyses are done to make sure that the items are free of defects. They revise some of the items that are repairable and drop the items that cannot be repaired.

When the item writer has completed the job of item writing, the items are subjected to field testing. For each subject area, at least four to five thousand subjects take the test and then the results are analyzed based on content and psychometric properties of the items. If an item has psychometric problems, for example one distractor does not function well, they remove it

Anani Sarab: How is the cut-off score determined for a test?

We call it standard-setting and it comes after filed-testing. As you well know, there are two ways of assessment: the normative-based scoring and the criterion-based scoring. In normative scoring, it does not matter what the student get; how many questions he/she has answered. They compare the result with the norm. Let me give an example. For example, we have a test of English as a foreign language. The maximum score is 100. A group of students get the maximum score of 20 and the minimum score of 2. The student who get the score of 20 gets an A and the student whose score is 2 gets an F, etc. They don't pay attention to whether the students have answered the majority of the items or not. But in criterion-referenced scoring, there is a criterion-based score that the students should get. They should for example get 80% or 90% of the questions correct to be accepted at the cut-off point. They set these based on the standard setting approaches. The Angoff, modified Angoff, bookmark, map mark methods, etc. are used for standard setting. In the Bookmark approach which is based on IRT model, if they have 100 items they rank order them based on the difficulty of the items. The difficulty level is determined empirically using the IRT 3-parameter logistic model. Then they invite a group of 60 to 80 judges in groups of 10 round table. Each judge first ask the question: If I wanted to label the student as proficient what items with what level of difficulty should he/she be able to answer correctly? If I wanted to label the student as pre-proficient or below proficient how many items has he/she answer correctly. The judges must be very experienced with content standards and the student performance to be able to make these judgments. The judges are given what is called the ordered test booklet in which the items are ordered from the very easy to the very difficult ones. The judges start with the very easy item and continue to the items below which the students who answered all the items correctly

and train them. There is extensive training for them. As the items should be sound and free of unnecessary linguistic complexity, content ambiguities, and cultural biases; the item writers should receive extensive training to do their job effectively.

Anani Sarab: How is the process monitored?

There are advisory boards in the state education department which oversee the process. I've been the member of ten states' advisory boards. We oversee all the activities and finally we have to approve it. Otherwise they cannot send the test to be printed and made functional.

Anani Sarab: How many members are in the advisory boards?

Advisory boards have five to eight members. They are invited three or four times a year and each time for three to five days for each subject. During their stay, they go through all stages including training, item writing, moderating, and alignment. They check the quality and if they find any issues they ask the team to repeat the procedure. So all states have advisory boards.
The members are experts in the area of psychometrics and have experience in test construction. They themselves should be involved in test construction from the very beginning to the end.

Anani Sarab: What kind of training and in what form is it provided to the team members?

The training is provided through class sessions and workshops in which a number of themes related to item development like alignment are discussed and practice runs are provided. They provide a lot of feedback to the team members.

When content standards have been operationally defined and expert have agreed that these are the ones that have to be measured they create a test blueprint for item writers in which the experts in the educational boards decide about the details. The details of the test construction are then provided to the item developers. In other words, very detailed and precise guidelines are provided for the item writers

Anani Sarab: What qualifications should the item writers have to be selected as team members?

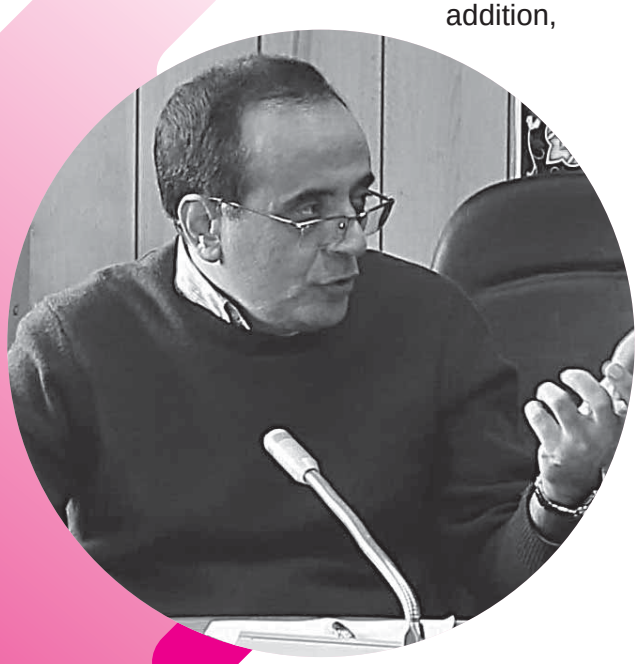
They have to have teaching experience and testing experience. They should have experience in classroom assessment and teacher-made tests. In general, they should be familiar with psychometrics. These are the qualifications that item writers should have. They should reach to a certain level of test development knowledge and experience to be considered as item writers. They are paid good amounts of money to develop items. In other words, in addition to qualifications the incentive and motivation is there. They spend as much time as needed for item writing. When they are in the group, they have supervisors. They are constantly checked and if they have any questions they are attended to by the supervisors. In general, they are heavily and extensively supervised.

When the item writer has completed the job of item writing, the items are subjected to field testing. For each subject area, at least four to five thousand subjects take the test and then the results are analyzed

based on content standards. They have to write an item to address the content standard. So if you refer to the common core standards of kindergarten, you will see the list of the content items that kindergarten children should know. The item writers take the list and develop the test blueprint. Test blueprints are created based on content standards. Based on the importance of each content standard item writers write less or more items.

Anani Sarab: To what extent are these standards consistent across states?

States have their own content standards. But when you put all 50 states together you don't see major differences. They are almost the same, but the politics of the states dictate that they have to have their own content standards. Since the common core standards were initiated, the states decided to have their own common core content standards. So in answering to your question, states have their own content standards but when you look at them altogether they are very similar. For example, the math standards of year two include addition, subtraction, multiplication and division, but different states might have different subscales for this common core content standard.



Anani Sarab: Do you make any distinction between standards and goals and objectives? They seem to be used interchangeably.

They are usually used interchangeably; however, we should remember that standards are supposed to be operationally defined to be measurable. The standards that I provided earlier are all measurable in an objective way. That is why the expectation is that if a large number of individual item writers write items for the same standards they will write similar items. The statements are so transparent that they need minimum levels of interpretation. There is no need for one item writer to write all the items. By the way all the standards have to be approved by the state education boards so there is a political aspect attached to it. When content standards have been operationally defined and expert have agreed that these are the ones that have to be measured they create a test blueprint for item writers in which the experts in the educational boards decide about the details. The details of the test construction are then provided to the item developers. In other words, very detailed and precise guidelines are provided for them. When items are developed based on the guidelines they are aligned with the content standards.

Anani Sarab: What qualifications should item writers have before they are recruited as item writers. Do they have to be teachers of the same content materials?

They normally select a group of teachers

content standards?

They use the Norman Webb's Alignment procedure. In this alignment procedure, they align test items with content areas in the following categories: (A) **Categorical Concurrence:** correspondence between the standards and assessment results are reported (**B) Depth of knowledge Consistency:** ratings of most cognitively demanding assessment (**C) Range of Knowledge Correspondence**: standards and assessments cover a comparable span of knowledge within topics **(D) Structure of Knowledge Comparability**: the relationships among ideas (**E) Balance of Representation**: the weight by topic or subtopics in the standards corresponds with their weight on the assessments (**F) Dispositional Consonance**: the desired dispositions toward the content area students are to develop.

They include a group of 10 to 20 judges. The judges are each individually asked to make a judgment whether a specific item is aligned on the above six categories, on a Likert scale of 5, with the content standards. So Norman Webb's procedure is a commonly used approach for aligning tests to content standards or the language of the four subjects, that is math, science, ELA and social sciences. So alignment has to be done in two areas; one is the academic content standards and the other one is the language proficiency standards.

Anani Sarab: So the alignment which is done in two different fields should normally be done with different experts; content experts and language experts. Is that right?

Yes, exactly. With ELP assessments, the alignment should be done with ELP standards and the language of math, science, social science and language arts. So in the domain of content the content experts do the job while in the second field the experts should do the alignment with the ELP standards. All states must have content standards in K-12 starting from kindergarten to year 12 in high school. The states have made their standards public. If you go to the department of education in any state you would find the standards in all levels with all the details. I have an example for the kindergarten. Standard 1.1 says that they are able to identify the front cover, the back cover and the title page of a book. Standards 1.2 says: they are able to follow words from left to right and from top to bottom on the principal printed page of a book. Standard 1.3: they understand that printed material provides information. When you give these content standards to test item writers, they know exactly how to write the items. It is not based on their opinions but

The English language proficiency has to be aligned with the language of content; that is, the language of mathematics, science, ELA and social sciences. So all language proficiencies have to be aligned to the language of school subjects. Therefore, we call them academic language proficiency

The English language proficiency has to be aligned with the language of content; that is, the language of mathematics, science, ELA and social sciences. So all language proficiencies have to be aligned to the language of school subjects. Therefore, we call them academic language proficiency. English language proficiency has four different subscales or sections; reading, writing, speaking, and listening. The combination of reading, writing, speaking and listening subscales were first introduced when NCLB and ESSA were introduced. Reading and writing can be more academic while speaking and listening are more related to social and conversational language. The focus is mainly on academic language skills. This means that ELP is aligned with the language of content subjects. When ELL students pass the language proficiency test and considered proficient they are ready to join the mainstream classrooms.

Anani Sarab: How is this focus on academic skills maintained in language proficiency assessment?

There are two different ways of making judgment about whether the students are English language proficient or not. They create a compensatory model or a conjunctive model. Based on the compensatory model they put all the components together to create a total score. The problem with this model is that normally the students are more proficient in listening and speaking rather than reading and writing. A student might be considered as proficient based on his/her very high scores in conversational language, while the same individual may have low levels of proficiency in reading and writing. So this compensatory model does not really work. Some students when they enter this country might be very fluent in listening and speaking skills but they may not be that much fluent in reading and writing skills. The conjunctive approach assumes that the students should develop all four skills to a proficient level. So even if a student is proficient based on the total score but has lower than desired proficiency in one or more skill he or she has to continue with the English language services. The implication is that native speakers, English or Farsi speaker, have to be proficient in academic and non-academic language; that is, in all domains of proficiency. Some ELL students who have been in this country for some time might have a lot of family and friends with whom to speak English. Through oral communication, they have become proficient in listening and speaking but not proficient in reading and writing. So many states do not use the conjunctive model but they use weights instead. For example, a consortium of 37 states put weights on the scores. They weigh reading and writing at 35%, listening and speaking at 15%. In this way, they compensate for this issue.

Anani Sarab: How do they align English language proficiency with

Educational Research Association, the 2013 National Association of Test Directors: Outstanding Contribution to Educational Assessment, the 2014 University of California, Davis: Distinguished Scholarly Public Service Award, the 2015 UC Davis School of Education Outstanding Faculty award and the 2016 national AERA E.F. Lindquist Award. He holds a Master's degree in psychology and a PhD degree in psychometrics from Vanderbilt University.

Anani Sarab: Through your research, you've made a strong case for the link between language and content. Would you please elaborate on the link between language and content in relation to English as L1 and L2?

There are two acronyms; English Language Proficiency (ELP) and English Language Arts (ELA). ELA is content assessment and is based on state standards. Most of the states use Common Core State Standards (CCSS). But ELP is based on English language proficiency standards originated by TESOL. So there are two completely different sets of standards. For native English speakers, we don't measure ELP at all. They don't need it. Based on the No Child Left Behind (NCLB) initiative, English language learners should do both ELP and ELA. They have to use ELP in order to make sure that English Language Learner (ELL) students are ready to participate in mainstream classrooms. If they are not ready or if they do not have the right level of English proficiency they have to receive more English training in order to be able to participate in the mainstream classes. So when students enter schools, they complete a survey called Home Language Survey (HLS) to check whether they speak a language other than English at home, if they do, then there are tested for their level of English proficiency using a simple English proficiency test called Screener. Based on the results of this test, the incoming students are categorized into proficient, and non-proficient in English. English proficient students will join the mainstream classrooms. The non-proficient students; however, are provided with English Language Development (ELD) services as long as they need the service to become proficient enough in English to participate in mainstream content classes. But ELA (English language arts) includes content standards which are based on state standards developed by the states. The Common Core State Standards may be followed by all states if they choose to do so. The states develop assessment based on these standards. They receive each some 25 million dollars to develop the assessment. They make the assessment based on those standards and they try to make it as accessible as they can. By accessibility, I mean they take the linguistic and cultural biases out of the tests to make sure that all sub-groups of students have the same level of access to the tests. They provide accommodation to make them accessible for ELL. Most states are members of one of the two common core assessment consortia (SBAC & PARCC). The consortia are supposed to develop standardized tests for the member states. They started the development of standardized assessment in 2010 and it took them five years to create these assessments. As we know, there are two different sets of language proficiency; academic and social conversational language. In No Child Left Behind (NCLB) and Every Student Succeeds Act (ESSA), they specifically refer to these specific proficiency types.

پلکت ویژه

Assessing Language and Content: An Interview with Professor Abedi

به کوشش محمدرضا عنانی‌سراب
دانشگاه شهید بهشتی
reza_ananisarab@yahoo.co.uk

پروفسور جمال عابدی، استاد برجستهٔ دانشگاه کالیفرنیا، دیویس، متخصص در زمینهٔ سنجش و ارزیابی است. علایق پژوهشی وی را مطالعه در زمینه‌های روان‌سنجی و تولید آزمون‌های استاندارد پیشرفت تحصیلی تشکیل می‌دهند. پروفسور عابدی در کارهای پژوهشی اخیر خود به مطالعهٔ اعتبار سنجش و ارزیابی و طبقه‌بندی و تعدیل آزمون‌های ویژهٔ زبان‌آموزان و زبان‌آموزان استثنایی پرداخته است. جوایز متعددی از جمله جایزه مشارکت برجسته در مرتبط کردن پژوهش و عمل توسط انجمن پژوهش‌های آموزشی آمریکا در سال ۲۰۰۳، جایزهٔ حاصل عمر در سال ۲۰۰۸ توسط انجمن تحقیقات آموزشی کالیفرنیا، جایزهٔ انجمن ملی مدیران آزمون تحت عنوان مشارکت برجسته در سنجش و ارزیابی آموزشی در سال ۲۰۱۳، جایزهٔ دانشگاه کالیفرنیا دیویس تحت عنوان خدمات عمومی برجسته دانشگاهی در سال ۲۰۱٤، جایزهٔ فعالیت‌های برجسته مدرسه علوم تربیتی دانشگاه کالیفرنیا دیویس در سال ۲۰۱۵ و جایزه لیندکویست در سال ۲۰۱٦ به وی اعطا شده است. پروفسور عابدی دارای درجهٔ کارشناسی ارشد روان‌شناسی و درجه دکتری روان‌سنجی از دانشگاه وندربیلت می‌باشد.

ضمن تشکر از پروفسور عابدی که دعوت سردبیر مجله را برای گفت‌وگو در خصوص ارزیابی پیشرفت تحصیلی در دروس محتوایی برنامهٔ درسی و ارتباط آن با توانش زبانی زبان اول و دوم و تولید آزمون‌های استاندارد پذیرفته و به تفصیل به این مسائل پرداختند حاصل این گفت‌وگو در زیر به خوانندگان مجله تقدیم می‌گردد. در این گفت‌وگو مسائل زیر مورد بحث قرار گرفته است:

- ارتباط زبان انگلیسی به‌عنوان زبان اول و دوم با یادگیری دروس محتوایی،
- استانداردهای محتوا و استانداردهای زبانی،
- مراحل تولید آزمون‌های استاندارد شامل تهیه مشخصات آزمون، تولید سؤالات آزمون، تعیین استانداردهای پیشرفت تحصیلی،
- نقش سیاست‌گذاری‌های آموزشی در تهیه آزمون‌های استاندارد.

Jamal Abedi is a Professor of educational measurement at the University of California, Davis. Abedi's research interests include studies in the areas of psychometrics and test development. His recent works include studies on the validity of assessment, accommodation, and classification for English language learners (ELLs) and ELLs with disabilities. Abedi serves on assessment advisory boards for a number of states and assessment consortia as an expert in testing ELLs. Abedi is the recipient of the 2003 Outstanding Contribution Relating Research to Practice award by the American Educational Research Association (AERA), the 2008 Lifetime Achievement Award by the California