

独立行政法人大学入試センター
入学者選抜研究機構国際シンポジウム報告書

2011 国際シンポジウム

教育テストの可能性

— 21世紀型能力の育成と高大接続 —

主催 独立行政法人 大学入試センター
協賛 UCLA/CRESST (National Center for Research on Evaluation,
Standards & Student Testing)
ACT (American College Testing)
KICE (Korean Institute for Curriculum and Evaluation
韓国教育課程評価院)
後援 朝日新聞社／文部科学省

平成 24 年 (2012 年) 3 月

独立行政法人大学入試センター入学者選抜研究機構

独立行政法人大学入試センター
入学者選抜研究機構国際シンポジウム報告書

教育テストの可能性

— 21世紀型能力の育成と高大接続 —

目 次

プログラム	1
登壇者紹介	2
開会挨拶	6
吉本高志（独立行政法人 大学入試センター理事長）	
第1部 基調講演	
“New Approaches to University Entrance Examinations in Korea – NEAT and CSAT (College Scholastic Ability Test)”	9
「韓国における大学入試（CSAT 大学修学能力試験）の現状と英語能力試験改革の動向」 Kyung-Ae Jin (KICE 韓国教育課程評価院, 英語能力試験担当部長)	
“New Approaches to Educational Testing and ACT”	30
「ACT(American College Testing)の現状と教育テスト開発の新たな展開」 Deborah Harris (ACT, Inc., Measurement and Reporting Service 部長)	
“New Approaches to Measuring 21 st Learning”	68
「21世紀能力の育成とその成果を測定する新たな方法の展望」 Eva L. Baker (UCLA=CRESST 所長, 前 AERA 会長)	
「我が国の初中等教育政策と大学入試」	118
銭谷眞美（東京国立博物館長, 元文部科学事務次官）	
第2部 指定討論	
“Validity Issues in Moving Ahead”	129
「前進による妥当性問題」 Joan Herman (UCLA=CRESST 副所長)	
「パネリスト講演からの示唆」	156
荒井克弘（大学入試センター 入学者選抜研究機構長）	
エバ・ベーカーによる総括資料	164

プログラム

日時：2011（平成23）年11月18日（金）

場所：有楽町朝日ホール

総合司会 田中義郎（大学入試センター入学者選抜研究機構 客員教授）

13:00 — 13:10 開会挨拶 大学入試センター理事長 吉本 高志

◆第1部 基調講演（13:10 — 16:00）

13:10 — 13:50

講演（1）“New Approaches to University Entrance Examinations in Korea - NEAT and CSAT(College Scholastic Ability Test)”

（韓国における大学入試(CSAT 大学修学能力試験)の現状と英語能力試験改革の動向）

Kyung-Ae Jin(KICE 韓国教育課程評価院, 英語能力試験担当部長)

13:50 — 14:30

講演（2）“New Approaches to Educational Testing and ACT”

（ACT(American College Testing)の現状と教育テスト開発の新たな展開）

Deborah Harris(ACT, Inc., Measurement and Reporting Service 部長)

14:30 — 14:40 休憩

14:40 — 15:20

講演（3）“New Approaches to Measuring 21st Learning”

（21世紀能力の育成とその成果を測定する新たな方法の展望）

Eva L. Baker (UCLA=CRESST 所長、前 AERA 会長)

15:20 — 16:00

講演（4）「我が国の初中等教育政策と大学入試」

銭谷眞美（東京国立博物館長、元文部科学省事務次官）

16:00 — 16:15 休憩

◆第2部 指定討論 / 総括（16:15 — 17:25）

16:15 — 16:40 （1）Joan Herman (UCLA=CRESST 副所長)

16:40 — 17:05 （2）荒井克弘（大学入試センター入学者選抜研究機構長）

17:05 — 17:25 総括

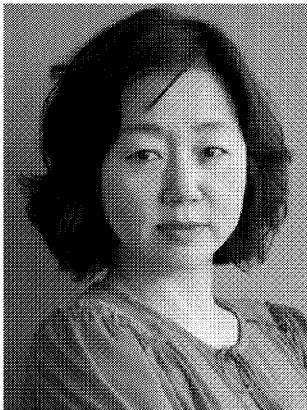
17:25 閉会挨拶 大学入試センター理事 惣脇 宏

17:30 終了

登壇者紹介

第1部 基調講演

講演（1）Kyung-Ae Jin(KICE 韓国教育課程評価院, NEAT (National English Ability Test) 部長) ギョン・エー・ジン



<プロフィール>

米国ピッツバーグ大学より博士号取得。専門は、外国語教育学。1994年から1998年まで、LG アカデミーにて英語教育プログラムの開発責任者を務める。1998年より、KICE 韓国教育課程評価院に移り、英語能力試験担当部長。現在進行中の韓国大学入試における英語能力試験改革の担当者でもある。その他、韓国外国語教育学会副会長、韓国英語能力テスト学会副会長、等を務めている。

講演（2）Deborah Harris(ACT, Inc., Measurement and Reporting Service 部長)
デボラ・ハリス



<プロフィール>

米国ウイスコンシン大学より博士号を取得。専門は、教育心理学。1984年にACT, Inc.に勤めて以来、25年以上に渡って、学力テスト、職業能力テスト、資格試験に関わり、多くの人々の教育達成や職業上の成功を支援すべく、それらの開発や評価に従事して来た。現在は、大凡、50人のACT, Inc.の心理測定専門家およびコンピュータ・プログラミングの専門家を統括し、テストの有効性の検証を行っている。ACT, Inc.のテスト開発の未来を最も良く知る専門家である。米国教育測定委員会(NCME)理事を務めている。

講演(3) Eva L. Baker (UCLA 卓越終身教授、UCLA=CRESST 所長/ WERA(the World Education Research Association)会長、前 AERA (American Educational Research Association 会長、米国科学アカデミー会員) エヴァ・ベーカー

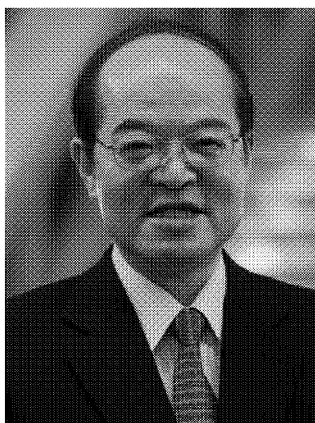


<プロフィール>

米国 UCLA より博士号取得。専門は、教育研究方法学および教育心理学。UCLA の卓越終身教授であり、1985年より CRESST の所長を務めている。複雑な学習の新たな評価方法の開発や有効性の検証、テクノロジーを活用した学習評価システムの開発や評価、等の領域で世界的リーダーである。前米国教育研究学会(AERA)会長、世界教育研究学会(WERA)会長、米国科学アカデミー会員等である。米国政府の教育開発および評価に関する様々な委員会で要職を、世界各国政府の教育評価デザインのアドバイザーを担ってきた。また、「21世紀スキル」の獲得を支援するために、子どもたちはもとより、医者、軍人といった専門職に対する学習支援ソフトの開発や評価での活躍も知られている。

講演(4) 「我が国の初中等教育政策と大学入試」

銭谷真美 (東京国立博物館長、元文部科学省事務次官)



<プロフィール>

東北大学教育学部卒、1973年文部省入省(大学学術局国際学術課)。文部大臣官房審議官(初等中等教育局担当)、内閣審議官(内閣官房内閣内政審議室教育改革国民会議担当室長)、文化庁次長、文部科学省生涯学習政策局長、文部科学省事務次官を歴任。2009年より現職。

第2部 指定討論

1. Joan Herman (UCLA=CRESST 副所長) ジョアン・ハーマン



<プロフィール>

米国 UCLA より博士号取得。専門は、教育心理学、学習および教育方法学。UCLA=CRESST の副所長。教育効果を高め、かつ、学力を高めるための教育システムのデザインの開発および評価に高い関心を持ち、ベーカー所長共々、新たな社会のニーズを分析し、「21世紀スキル」の獲得を支援する研究活動を推進している。

2. 荒井克弘 (大学入試センター・入学者選抜研究機構長)



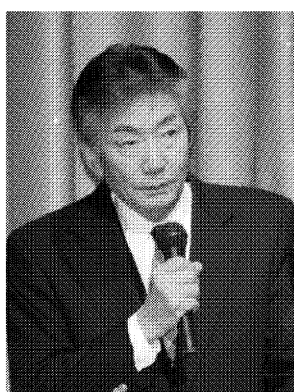
<プロフィール>

独立行政法人大学入試センター入学者選抜研究機構長、教授。専門は高等教育研究、教育計画論。東京工業大学大学院理工学研究科博士課程修了、工学博士。国立教育研究所室長、広島大学大学教育研究センター教授、東北大学大学院教育学研究科教授、同大教育学部長・研究科長、教務担当副学長などを歴任。2009年より現職。

開会の挨拶

吉本高志

(独立行政法人 大学入試センター理事長)



本日はご多忙のところ、またご遠方から多数の方にお集まりいただきありがとうございます。大学入試センターの2011年の国際シンポジウムを開会いたします。

今年は、日本にとりまして未曾有の大震災の年でありました。被災された方々に心からお見舞い申し上げるとともに、被災地の一日も早い復興を願っております。来年1月の大学入試センター試験におきましても、罹災した志願者の受験料免除をはじめ、被災地の受験者が不安なく試験を受けられるよう、例年にもまして試験場の交通アクセス、環境の安全・整備に努めております。本日

の会場にも、高校・大学関係者が多数おられると存じますが、その折にはご協力のほどをよろしくお願いいたします。

さて、今回のシンポジウムのテーマですが、昨年度は大学、高等教育の側から「これからの大学入学者選抜を考える」と題して、議論をいたしました。今回は視点をかえて、初中等教育の側から「高校と大学の接続」を議論したいと考えました。タイトルの「教育テストの可能性」という言葉にはこの趣旨が込められております。

これまで、「大学進学のためのテスト」というのもっぱら”選抜”が目的でした。しかしいまや「テスト」は選抜だけでなく、教育をいかに支援するか、学びをいかにゆたかにするか、という課題が重要になっています。21世紀には、これまでより多くの人々が高等教育を受ける時代がやってきます。大学入試センター試験も「選抜テスト」としてだけでなく、高校と大学をつなぐ仕組みとして新しい教育的役割を期待されることになるだろうと考えます。

本日は、このテーマをより広い視野から議論していただくため、国内外から著名な講演者、コメンテーターの方々をお招きしました。これだけの専門家が一堂に会して「教育テスト」を論じるのはめずらしく、貴重な機会であると存じます。実り多いシンポジウムとなることを心から期待しております。

また、今回のシンポジウムでは文部科学省、朝日新聞社からの後援をいただきました。この場を借りて感謝を申し上げ、開会の挨拶とさせていただきます。

第 1 部
基 調 講 演

New Approaches to University Entrance Examinations in Korea - NEAT and CSAT (College Scholastic Ability Test)

Kyung-Ae Jin

Korea Institute for Curriculum and Evaluation

Good afternoon. My name is Kyung Ay Jing, I am from Korea, the Korean Institute for Curriculum and Evaluation. I would like to express special gratitude to the president of NCUEE, professor Dr. Yoshimoto, Professor Tanaka, and Professor Arai for inviting me to this presentation and for showing me such hospitality. I feel honored to present with such distinguished scholars, such as professor Eva Baker, Deborah Harris, and John Herman.

Today, I would like to present about the new English test in Korea that our institute is developing. The name of the new English test is the National English Ability Test. The background of developing this new English test is, the Korean government and specialist realized that Korean students lacked practical English skills even with more than 10 years of formal English education in school. Also, we found that there is a need to enhance practical English education, especially students' speaking and writing skills in English. The current CSAT test, which is College Scholastic Ability Test, is such an important high-stake test to go to University in Korea, so CSAT influences many aspects of education in Korea. The English section of CSAT does not assess speaking and writing skills. So we came up with the idea that we have to include the assessment of speaking and writing skills in the CSAT test. So this is the background of why we decided to develop this new English test.

The differences between the CSAT test and NEAT test are the following: In the CSAT test currently being delivered in Korea, the English section measures just listening and reading. In the new test, we measured four skills – listening, reading, speaking, and writing. All of the CSAT items are multiple choice, but in the new test reading and listening are still multiple choice, but speaking and writing will be measured through performance assessment, which means actual performance in speaking and writing. The current CSAT test is non-referenced but the new test is criterion-referenced, so we set up the criteria first and we see whether the students are attaining or achieving the reference or objectives. In the current test, nine levels of percentile scores are provided to the students, therefore it is just relative ability that is expressed through the school report. But in the new test we, since this is criterion-referenced, we set up the achievement standards first and we measure whether the students are achieving to excellent, average, basic or below basic with these achievement standards. So that is the big difference from the previous test.

And the current CSAT test is a paper and pencil test. For example, more than 600 thousand students are taking the test in one day to get the scores to apply to colleges. So it is paper and pencil

test, but we found that with this paper and pencil test, we cannot measure abilities like speaking and writing, and we cannot score these skills with paper and pencil tests. So what we have established is an internet-based test. So the new test is delivered through the internet, especially VDI, which is like a cloud-computing system, so it is virtual desktop infrastructure. Therefore, students are taking the test at the test center through the computers, and the test items and testing programs are implemented in the main center, so the students' personal computer just works as a monitor. So, we tried to introduce this internet-based test.

Also, the current CSAT test, students are given just one chance to take the test. So for students, that one day is such an important day in Korea that will decide the students' future with that one CSAT test. So with the new test, we are designing the test and delivering the test at least two opportunities for each student, and there will be more practice tests. However, we decided to give at least two opportunities for the students to take the test. Current CSAT tests are just one test paper. But in the NEAT test, we are developing two different test papers, so students can choose either the level two test, which is a bit higher level, but the main difference is the content – it is an academic English test. But the level three test is more oriented towards practical English. So we provide two different versions of the NEAT test, and students can choose either the level two test or the level three test. And the college or universities will provide guidelines on whether they will require either level two or three test based on the characteristics and contents of the courses of the university.

More background. Actually, this idea to develop a new test was raised in 2006, so it is from the previous government. The previous government and the current government both put a lot of emphasis on English education. They believe that English skills will be one of the competitive power and skills in this global era. So in 2006, the previous government announced several English reform policies, and the objective is to improve students' English ability in speaking and writing. Since many industries feel that even though graduation of high school and college, the employees are not good at communicating in English. That is why the government has put more emphasis on English education. In 2007, they announced a new plan, especially they announced this new test.

The government is providing many new policies and support for English education, but the reason that is not changing satisfactorily is because of the university entrance test. So we thought that without changing the format of the entrance test, we cannot change the content of education in schools. So they have announced the plan to develop a new English test in 2007. And in 2008, the current government's President emphasized that since 2006, KICE and Ministry of Education have developed the idea and have prepared for this new test, and in 2008, the President and the Minister of Education announced that the new test will replace the current CSAT test. So the idea is that we first started to develop the new test, but in 2008 they made a plan to replace the current CSAT test with this new test.

Since 2009, we have done altogether seven field tests with almost around 50 thousand students. So after we have designed this test and developed items for this test, we have done field tests. In the first public hearing for this new test was in 2010. After the first public hearing, we have implemented several additional field tests, and the final public hearing was May this year. We have a public hearing from teachers and specialists in the field about this new test.

The goals of this new test is to let the students acquire basic communicative skills through school education so students will select either NEAT level two or three based on aptitude and future career. And directions of developing this new test is to align the contents of the new test with the curriculum. The English national curriculum in Korea, we have introduced this communicative-competence oriented English curriculum since the early 1980s. But since there is a gap between this curriculum and this CSAT test, that is why the government realized that there is no effectiveness in English education. So the direction is to align the NEAT test with the national curriculum. Also, another direction or strategy is to provide teacher training programs with this new test. What we do is, one of the big tasks in developing this new test was to train teachers how to write items, how to score, or instruction related to the new test in the classroom. We tried to develop this test with this new curriculum and teacher training efforts.

The target system of this new test, will be, as I said, if we give two opportunities to the students, the students will be 1200 thousand students taking 24 tests. The reason for 24 tests is since this an internet-based test, we cannot have all the students taking the test at the same time because of the facility. So 50 thousand students take the test, and we need 12 test papers to assess the 600 thousand students. So if we are giving two opportunities, that is how this number came up. So we keep training item writers, or the secondary school teachers. We are building up a system headquarters and 1700 test centers throughout the country.

The timetable of the development is as follows. We have developed and done several pre-tests, and since next year, 2012, general implementation of the test will begin. So, around seven colleges and universities have a plan to use the scores of this new test next year. The general implementation will start next May. At the end of next year, the government will make a decision whether this new test will replace the current CSAT test, and if this decision is made, the English section of the CSAT will be abolished and the new NEAT test will begin in 2015. We need at least three years of notice period before the entrance test changes. The new English curriculum is a 2009 revision, and we periodically revise our national curriculum. The new national curriculum has been developed aligned with this new test. It will be applied in 2013 for first-year high school students who will take this NEAT test in 2015. So, we are trying to prepare the schools for this new test.

The level two test is for academic English skills, as I mentioned, and the level three test is basic practical English test. The level two test will be 3000 words and 2000 words for the level three test.

The level two test is aligned with some of the curriculum and the level three test is also aligned with the curriculum. This test is measuring four domains or skills: listening, reading, speaking, and writing. These are the number of items and test time. In each test paper, there will be some anchor items for testing, since 50 thousand students are taking the test but the total population of the students will be 600 thousand, so each set of 50 thousand will have different test papers, so the students' scores will be equated later with common items in order to prevent inherent item difficulty.

So these are some of the topics for the level three test – practical English. These topics have been used to write test items for the level three test. The basic academic English test we also have some topics. We try to differentiate these two tests in terms of contents, which are designed by the topics of each test item. I may not go too much into detail into all of the test specifications but briefly, in listening we have items for literal understanding and inferential understanding. These are the numbers of words for the level two test and the level three test. The level two test includes two items per each text. The current CSAT test basic policy for the English test is one question for one text. We tried to add multiple items for each text, and also multiple paragraphs. So as you see the number of words for each item in the level two and three tests differ.

Also, the characteristics of the new item types, since this is an internet-based test, and since we tried to focus on more practical English, we tried to include test items including charts, pictures, graphs, and in general more authentic texts, rather than just texts that can be used in the classroom. When we developed this new test, we also tried to align it more closely with the textbooks. Before, the item types of the CSAT test were different from what the students are actually learning in the classroom through the textbook. But in this new test, we tried to develop items that would be similar to activities in the classroom and information presented in the textbooks. Textbooks were good resources when writing items or studying for the test.

I will not go over each item type, but I will show you the items that have distinctive characteristics of this new test. The most important aspect of this new test is it has a speaking and writing component in the national test. It requires practical issues of implementing, scoring, speaking, and writing items, but since the will to reform English education has decided to introduce speaking and writing tests. This is a new listening test item. For example, the test item on the right-hand side is a schedule for a cinema. We tried to introduce authentic materials – the real materials for the test. The students are listening to the script and see which picture the speaker decided to watch. The item on the right-hand side is also a listening test. That is a map that visitors can have when visiting sight-seeing places, so students are listening to the monologue or announcements and seeing the exact place on the map. So that is a new test item that we introduced.

Also, the test on the right-hand side is a new type. So that is an actual advertisement. We tried to understand the students' literacy skills with advertisements, also. Students are supposed to see this

advertisement and understand what it says. That is also a new item type in the level three test – the practical test. This test, a level two test, has a longer text. We tried to get the students to understand the longer text as well. This is the speaking test.

Computer: Part 1: you will see three pictures. Then you will be asked one question. Answer the question based on the picture. You should answer with one or two full sentences. After each beep, record your answer for 15 second. After you hear two short beeps, stop recording. Question number one: Is the game exciting?

Students are just answering the questions and they are supposed to record their speaking into the microphone that is attached to the headset on the computer. This is a picture description task for the speaking test.

Computer: Part 2: You will see six pictures and tell a story based on the picture. You have one minute to review your answer. After the beep, you will have one minute to record your answer. The story should start with: “One day a man was...” After you hear two short beeps, stop recording. Now let's begin.

Computer: Part 3: You will hear a story describing a problem. You will have one minute to think about your advice on how to solve the problem. After the beep you will have one minute to record the answer. After you hear two short beeps, stop recording. Now let us begin.

So students are actually recording their answers through the microphone. This is an example for the test for writing. We provide some prompts like a place, time of visit, and the reason for choosing it, so the students are supposed to write about a place they have traveled to recently. That is one of the item types for writing.

So I have introduced some of the item types for this new test and implementation as I presented. This is the cloud-computing system. The system is now being built at KICE (Korea Institute for Curriculum Evaluation) and the setting up of the Virtual Desktop Infrastructure (VDI) at the center students login on the monitor and take the test at the test center, and the client at the test center downloads the test items and uploads the answers. The reason that we introduced this cloud-computing system is to minimize the “if” factor of the PC environment at the test center, and therefore provide a more equivalent testing environment. So during the implementation since 2009, we realized that first it was the client type – that is there is a server at each test center and they download the test. But we realized that the environment of students’ PCs differ. So we do not want that environment, the PC environment will affect the students’ performance. So that is why we tried to incorporate this new technique. This year, we are building 500 testing centers or school computer labs. Schools are applying for these new test centers and KICE is examining the test centers. We allow the

center to be enrolled for the NEAT test. By 2013, 1700 test centers will be built throughout the country.

One of the issues for this test is also scoring. KICE is providing a structured rater training process. Each rater has practiced rating at least 240 samples, and they are certified as raters. This is the scoring method. Each rater is scoring two items, but it will be linked with other raters in an attempt to control the rater effect through the writing design. So for example, up till now, we have quite reliable inter-rater scores so far, so we tried to train the raters for speaking and writing specifically to a 0.8 reliability. We cannot secure so far, but we are trying to increase the inter-rater reliability. We are also thinking of applying automated scoring for writing. As I mentioned before, the scoring is criterion-referenced. The item is analyzed after the students' performance on the test. It is also analyzed with item response theory. The reason that we are applying this item response theory is because the scores of these items are later used for standard-setting, which means we have to actually finally provide an excellent, average, basic, or below basic. And the students' scores will be used as a cut score to distinguish these four levels. Also, equating is implemented, which will be done by more specialists.

We try to use the methods that are already used in the United States and the best testing companies throughout the world. The idea is that we have common items that are provided for each test paper. Therefore, we have introduced these common items, and we equate the student scores later. In order to get this common item information, we have constantly provided field tests. The item scores and item information are stored in the item bank so that they will be used later as this common item for creating. We also have a standard-setting procedure. For example, if you compare across 20 faculties, they are participating in standard setting, which means that scores are given for four levels. This is the one sample for scores for listening and reading. Finally, students get scores for each skill – speaking, reading, listening, and writing – and this performance level description, which is related to their achievement level, is also provided.

In conclusion, we have done some surveys during the development of the test. People think that this new test will enhance the students' speaking and writing ability. Also, a lot of research was done last year that evaluated some pauses for English education and most of them agree to adopt new tests. 72.5% of parents expect this innovation of speaking and writing testing and instruction in school will enhance the speaking and writing instruction in school. The parents believe that this new test will allow students to speak and write English better. We believe that this new test will enhance speaking and writing skills in schools, and will have a positive influence on English education in schools for more practical and effective skill adoption. This is the end of my presentation. Thank you.

韓国における大学入試（CSAT:大学修学能力試験）の現状と

英語能力試験改革の動向

Kyung-Ae Jin

韓国教育課程評価院

こんにちは。韓国教育課程評価院（Korea Institute for Curriculum and Evaluation : KICE）から参りましたジン・キョン・エと申します。

大学入試センター理事長の吉本先生をはじめ、田中先生、荒井先生には、今回のプレゼンテーションにご招待いただきましたことを感謝申し上げます。また、大変なご歓待をいただきましたことを感謝いたします。さらに、エバ・ベイカー先生、デボラ・ハリス先生、ジョアン・ハーマン先生といった素晴らしい先生方とともに登壇できますことを大変光栄に思います。

本日、私は、私ども韓国教育課程評価院が開発しております韓国における新しい英語の試験についてお話しさせていただきます。この新たな英語の試験の名称は、全国英語能力試験（National English Ability Test : NEAT）と申します。この新しい英語能力試験が開発された背景といたしましては、韓国の生徒は10年間も学校で英語教育を受けているのに、実践的な英語能力が不足しているということを韓国政府や専門家が認識したことにあります。また、私どもは実践的な英語教育、特に話す・書くというスキルを強化する必要があると感じました。現在、韓国では大学修学能力試験（College Scholastic Ability Test : CSAT）という試験を行っております。韓国においてこの試験は、大学進学のために重要でハイスティックな試験という位置づけとなっております。そのため、CSATは、韓国における教育の多くの面に影響いたします。しかし、そのCSATにおいて英語について行われる試験は、話す・書くというスキルについての評価をしておりません。そこで私どもはCSATに話すスキル・書くスキルのアセスメントを組み込まなくてはならないという考えを示しました。これが、新しい英語試験を開発することに決めた理由の背景です。

このCSATとNEATの違いというのは以下ようになっております。現在韓国で行われておりますCSATにおいて、英語の試験は単に聞くこと、読むことを測定します。しかし、新しい試験では4つのスキル、すなわち聞く、話す、読む、そして書くというスキルを測定します。また、CSATの全て項目は多肢選択式となっております。新しい試験においても、読むこと、聞くことに関しては多肢選択式です。しかし、話すことと書くことについてはパフォーマンスアセスメントを通して測定されます。つまり、話す、書く際の実際のパフォーマンスを評価することを意味します。現在のCSATは、非準拠（non-referenced）ですが、新しい英語の試験は、基準準拠（criterion-referenced）となっております。したがって、私どもは最初に目標を設定し、そして生徒が、基準または目標に到達しているか、達成しているか達しているかということを確かめます。さらに、現在の試験では、パーセントイル得点を9つ

のレベルに分けたものを生徒に提供します。したがって、成績通知によって提供されるのはあくまでも相対的なものです。しかし、新しい試験のほうは目標基準準拠ですので、私どもは最初にアチーブメントスタンダードを作成し、生徒が優秀であるのか（excellent）、平均的なのか（average）、基礎的なのか（basic）、基礎より低いのか（below basic）ということ、このアチーブメントスタンダードをもとに測定します。これがCSATとの大きな違いです。

そして、現在行われているCSATは、紙と鉛筆を用いた試験になっております。例えば、60万人以上の生徒が、大学へ出願するための得点を得るには、試験を1日で受けます。紙と鉛筆の試験となっておりますが、それでは話すことや書くことといったスキルを得点化することができません。そのため、インターネットベースの試験を新たに作成しようということになりました。新しい試験に関してはインターネットを通じて受験することになります。そこでは、VDI（Virtual Desktop Infrastructure）を用います。これはデスクトップ環境を仮想化してサーバ上に集約したものです。したがって受験生は、テストセンターのパソコンからネットワークを通じてサーバ上の仮想マシンに接続し、試験を受けます。試験の項目やプログラムはメインセンターで実行されるので、受験生のパソコンはちょうどモニター画面として働きます。このように、私どもはインターネットベースの試験を導入しようとしています。

さらに、現在行われているCSATにおいては、受験する機会が1回しかありません。そのため、韓国の受験生にとっては、1回のCSATによって将来が決まってしまうので、その1日がとても重要な日になるのです。そこで私どもは、新しい試験において、受験生が少なくとも2回は受験できる機会を提供できるように試験を設計しております。そして、練習のための試験も用意しております。さらに、現在のCSATにおいて行われる試験は1つです。しかし、NEATに関しましては、2つの異なる試験を開発しております。そのため、受験生はレベル2という試験かもう一つの試験（レベル3）という2つのレベルの試験のどちらかを選ぶことができます。この2つの試験の主な違いはその内容です。レベル2というのはレベル3よりも少し難易度が高く、学術的な英語の試験となっています。レベル3に関してはより実践的な英語を指向しております。このように2つのバージョンの試験を私どもは提供します。そして、受験生はレベル2かレベル3のテストを選ぶことができます。さらに、大学のほうでも大学のコースの特徴や内容に基づいて、レベル2とレベル3のどちらのレベルの試験が必要なのかということガイドラインとして提供するようになっております。

また背景的なことですが、実は新しいテストの開発という考えが出てきたのが2006年のことで、それは前政権のときでした。前政権も今の政権も両方とも英語教育に重点を置くという点で同じです。彼らは、英語のスキルがこのグローバルな時代のスキルや競争力の一つであると考えています。したがって、2006年に前政権がいくつかの英語改革政策を打ち出しました。その目的は、生徒の話すあるいは書くといった英語能力を高めることです。多くの産業では、高校や大学を卒業しても、従業員は英語で情報交換をすることが得意ではないと感じています。そのため、政府は英語教育を強化しようと、2007年に新しい計画を打ち出しました。特にこの新しい試験について発表いたしました。

政府は、英語教育のための多くの新しい政策や支援を提供しています。しかし、満足のいく変化とな

っていません。その理由は、大学入学試験にあります。そのため、この大学入学試験のフォーマットを変えなければ、学校教育の内容も変えられないと考えました。したがって、2007年に新しい英語試験の開発する計画を発表しました。2006年以来、韓国教育課程評価院と教育科学技術部（日本でいう文部科学省）がそのアイデアを発展させて、この新しい試験の準備をしたということを、2008年に現政権の大統領は力説しました。そして2008年、大統領と教育科学技術部長官（日本でいう文部科学大臣）は、現在行われているCSATを新しい試験と取り替えると発表いたしました。したがって、そのアイデアに基づき、この新しいテストの開発が始められました。そして、2008年に現在のCSATから新しい試験と入れ替える計画を立てました。

2009年から5万人近い生徒を対象にした試行テストを7回行いました。つまり私どもは、試験をデザインし、項目を開発した後に、試行テストを行ったのです。この新しい試験のための最初の公聴会が2010年に持たれました。最初の公聴会を行った後、何回かやはり追加的な試行テストを行いました。そして、最後の公聴会は今年の5月に行われました。私どもは、この新しい試験に関して教師あるいはこの分野の専門家からの公聴会を行っているのです。

この新しい試験の目標としては、将来のキャリアや適性に基づいて、生徒がNEATのレベル2とレベル3のいずれかを選択するように、基本的な伝達スキルを学校教育の中で身につけることです。また、新しい試験の開発の方向としましては、新しい試験内容とカリキュラムを整合させていくことです。韓国における英語のナショナル・カリキュラムでは、1980年代前半から伝達能力に関心を向けた英語カリキュラムを導入しました。しかし、実はこのカリキュラムとCSATのテストの間にギャップがありまして、そのために政府は英語教育が有効ではないと考えました。したがって、向かう方向といたしましては、NEATテストをナショナル・カリキュラムと整合させることです。さらに、別の方向あるいは戦略としましては、教員養成プログラムにこの新しい試験を提供することです。つまり、この新しい試験の開発における大きな課題の一つは、項目を作成する方法、得点をつける方法、あるいは教室において新しい試験と関連づけられた教師を養成することでした。私どもは、新しいカリキュラムと教員養成の取り組みを備えた試験を開発しようと努めました。

先ほど言いましたように60万人の受験生に2回の受験機会を与えた場合には、この新しいテストのターゲットシステムでは、24回の試験を受ける120万人の受験生になるでしょう。この24回の試験になった理由は、インターネットを使ったテストですので、設備的に全員が同時に受けることはできないためです。そこで、1回の試験について5万人が受験をするということになり、そして60万人の受験生を評価するために、12回の試験、12の試験問題を必要とします。さらに、1人に2回の機会を与えるということで120万人分、24回の試験ということになるわけです。そのため、私どもは、項目作成者または中等教育の教員を養成し続けているのです。私どもは、このインターネットを用いたシステムのために本部と全国に1,700のテストセンターを設けています。

この後の開発の予定につきましては以下の通りとなっております。これまで予備試験ということで何回か試験を開発し実行しました。そして、来年、2012年以降は、全体的な実施を行っていきます。7つ

の大学が来年、この新しい試験の得点を使用することを計画しています。このような全体的な実施は来年の5月から始まります。来年末には政府が、この新しい試験を現在のCSATの替わりにするかどうかの決定を行います。もしその決定がなされた場合、CSATの英語は廃止され、そして新しいNEAT試験が2015年をもって開始されます。この入学試験が変更する前に、少なくとも3年間の通知期間を必要としています。さらに、新しい英語のナショナル・カリキュラムは2009年に改訂が行われました。私どもは、ナショナル・カリキュラムの改訂を定期的に行いますが、新しいナショナル・カリキュラムは、この新しい試験と連携して開発されています。2015年にこのNEATを受けるのは2013年に高校1年生となる生徒ですので、それは2013年から適用されます。したがって、私どもは高校にこの新しい試験に向けての準備しようとしております。

先ほど、NEATにはレベル2、レベル3があるという話をいたしました。レベル2の試験は学術的な英語スキルの試験であり、レベル3の試験というのは基礎的な実践英語スキルの試験ということになります。レベル2のほうは3,000単語、レベル3は2,000単語の語数となるでしょう。レベル2の試験はカリキュラムの一部と連携し、レベル3の試験もカリキュラムに合わせたものになっております。この試験ですが、4つの領域あるいはスキルを測定しております。それは、聞くこと、読むこと、話すこと、そして書くことです。項目数と試験時間が、資料に載っております。各試験問題を5万人ずつ受験しますが、全ての受験生の数は60万人ですので、テストのためのアンカー項目が入っています。そのため、受験生の得点は、異なるテストを受けることによる固有の項目困難度の違いを防ぐために、いくつかの共通項目を用いて、後で等化を行います。

次に、資料を見ていただきたいのですが、これらは、実践的な英語であるレベル3の試験のトピックのうちのいくつかです。これらのトピックは、レベル3の試験のための試験項目を作成するために使用されました。さらに、学術的な英語の試験についてもいくつかトピックがあります。そのため、私どもは内容の点から、この2つの試験を区別しようとしています。そして、それは各試験項目のトピックによってデザインされております。今回、ここでテスト仕様書の中についてそれほど詳細に立ち入るわけではありませんが、手短かに申しますと、聞く試験に関しましては、文字通りの理解、そして推論による理解のための項目を持っております。そして、これらは、レベル2の試験およびレベル3の試験に対する言葉の数です。レベル2の試験については、各テキストにつき項目が2つ含まれています。英語の試験としての現在のCSATでは、1テキストにつき1項目が基本的な方針となっております。私どもは、各テキストに多数の項目、さらに多数のパラグラフを加えようとしていました。そのため、各項目について言葉の数を見ても、レベル2とレベル3では異なっています。

さらに新しい項目タイプの特徴ですが、これはインターネットベースの試験であり、そしてより実践的な英語に注目しようとしたので、その項目は、教室で使用されるものよりもより実践的で、絵やグラフ、図表といったものを含んでおります。それから、この新しい試験を開発したとき、より密接に試験を教科書に合わせようとしていました。CSATの項目タイプは、実際に教室で教科書を用いて習うものとは異なっていました。しかし、この新しい試験では、教科書に示される情報や教室での活動と類似する項目を開発しようとしていました。項目を作成することや、あるいは試験のために勉強をする素材と

して、教科書は良い資源となりました。

それぞれの項目タイプについては一つ一つ申し上げませんが、この新しい試験の際立った特徴のある項目をお見せしたいと思います。この新しい試験の最も重要な側面は、全国試験が、話す・書くという構成要素を持っているということです。そうしますと、話す・書くという項目についてどのように実行し、得点化するかという現実的な問題が出てきます。しかしながら、英語教育を改革するという決意以降、話す・書く試験を導入することが決定されています。さて、資料には新しいリスニング試験の項目が記されております。たとえば、右側のこの試験項目は映画館のスケジュールです。このように私どもは現実的な材料、実際の素材を試験に使おうとしています。受験生は、スクリプトを聞き、そして話し手がどの映画を鑑賞することに決めたのかのかについて答えます。また、右側の項目ですが、これもリスニングの問題です。それは地図になっておりまして、観光地へ訪れた時に持つような地図です。受験生は、その地図についてのモノログや発表を聞き、地図上の正確な場所を確認します。これが私どもが導入いたしました新しい試験の項目です。

また、右側の上の方の試験も、新しいタイプの試験問題です。これも実際の広告です。私どもは、広告を用いて受験生のリテラシースキルを判断しようとしてしました。受験生は、この広告を見て広告が何を言わんとしているかを理解するでしょう。これはレベル3の実践的な試験の新しい項目タイプでもあります。レベル2の試験ではかなり文面が長くなっております。私どもは受験生に長文理解を促そうとしているのです。

次は実際のスピーキングのテストをお聞きください。

コンピューター：

「パート1 3つの絵を見てください。そして、一つ質問をされますので、その絵に基づいて質問に答えてください。そのとき、1つか2つの完全文で答えてください。ピーと音がなりますので、その後、15秒の間であなたの答えを録音してください。ピッピッと短く2回なりましたら録音は終わります。第一問 そのゲームは面白いですか。」

受験生は問題に答えますが、彼らはその答えをコンピューターのヘッドセットに付けられているマイクへ話すことを録音します。これはスピーキングの試験のための絵に含まれる情報を説明するタスクです。

コンピューター：

「パート2 6枚の絵を見て、その絵に基づく物語を作ってください。答えを考える時間が1分間あります。その後、ピーと音が鳴りますので1分間に回答を録音します。答えるときは、『ある日、男は・・・』から始めてください。ピッピッと短く2回なりましたら録音は終わります。では始めます。」

「パート3 解決すべき問題についてのお話が聞こえます。そして、その問題を解決するためにどのような助言をしたらいいのか1分間で考えてください。その後、ピーと鳴りますので1分間に回答を

録音します。ピッピッと短く2回なりましたら録音は終わります。では始めます。」

受験生は自分の答えをマイクを通して録音することになります。さて、次は記述に関する試験の例です。私どもは、受験生が答えやすいように、場所、いつそこに到着したのか、そして、なぜそこに行くことを決めたのかなどについて、いくつかのプロンプトを提供します。したがって、受験生はそれをヒントに最近行ったことのある場所について書くでしょう。これが記述に関する項目タイプの1つです。

このように、新しい試験やその実施のために項目のさまざまなタイプを導入しております。それはここまでお示ししました通りです。これは、クラウドコンピューティングシステムとなっております。今現在、そのシステムを韓国教育課程評価院で構築しています。それは、仮想デスクトップ・インフラストラクチャ（VDI）を用いて、受験生がテストセンターのモニター上でログインをし、試験を受けるといったものです。テストセンターのクライアントが試験項目をダウンロードして、そして受験生の回答をアップロードします。私どもがこのクラウドコンピューティングシステムを導入したのは、テストセンターのPC環境の影響をなるべく最小限にするためです。したがって、より等しい試験環境を提供していると思います。これらのシステムを2009年から実施しておりますが、最初はクライアントタイプでした。それは、各テストセンターにサーバがあり、そこで試験をダウンロードします。しかし、私どもは、各受験生のPCの環境が異なっていることを知りました。PC環境は、受験生のパフォーマンスに影響を及ぼします。そのような環境を望みません。それで私どもは、この新しい技術を取り入れようと考えたのです。私どもは今年、500のテストセンターあるいは学校にコンピューター室を設置しています。各学校はこれらのテストセンターの設置について申請し、韓国教育課程評価院はそのテストセンターの審査を行います。私どもは、各テストセンターがNEATテストのために登録されることを許可します。2013年までに1,700のテストセンターを全国に設置する予定であります。

それから、この試験の課題の一つは、スコアリングです。韓国教育課程評価院は、構造化された評価者トレーニングプロセスを提供しております。各評価者が少なくとも240のサンプルを評価する練習をすることによって、彼らは評価者として保証されます。それから、私どもが行うスコアリング法ですが、各評価者が2つの項目をスコアリングします。これはライティング・デザインによって評価者の影響をコントロールし、そして、ほかの評価者のスコアとリンクされます。私どもはこれまで評価者間の得点の高い信頼性を確保していました。0.8程度の信頼性が得られるよう、各評価者にスピーキング、ライティングについてトレーニングしようと努めました。私どもは評価者間の信頼性を高めようとしております。また、ライティングを自動採点化させることなども考えております。既にお話し申し上げましたとおり、そのスコアリングは基準準拠となっております。その項目は受験生のパフォーマンスの後に分析されます。さらにそれは項目反応理論で分析されます。私どもが項目反応理論を用いる理由は、これらの項目の得点がのちに分割点・基準の設定（standard setting）に使用されるからです。すなわちということかといいますと、私どもは実際に優秀（excellent）、平均（average）、基本（basic）、基本以下（below basic）ということを最後に提供しなければならないため、それを基準にするのです。そして、受験生の得点は、これらの4つのレベルを識別するためのカット・スコアとして使用されます。さらに、多くの専門家によって等化が行われます。

私どもは、アメリカや世界中の優良テスト会社で既に用いられている方法を使用しようとしています。それは、各試験問題に共通項目を設けているということです。したがって、私どももこの共通項目を取り入れ、そして、受験生の得点を等化します。この共通項目の情報を入手するために、私どもは常にフィールドテストを実施しております。後で共通項目として使用されるように、項目得点と項目情報は、項目バンクに格納されます。これに加えて私どもは、基準設定手順 (standard-setting procedure) を行っております。たとえば、20の学部を比較する場合には、それらの学部は、4つのレベルに分けられる得点である標準的な設定を共有します。これが聞くこと・読むことについての得点における一つの例です。受験生は最終的に話す、読む、聞く、書くについてのスキルごとに得点を得ます。そして、このパフォーマンスレベルつまり到達度のレベルについての記述されたものを提供されます。

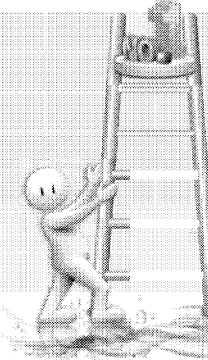
まとめになりますが、私どもはこのテストの開発に当たっていくつか調査を行いました。人々は、この新しいテストが生徒の話す・書く能力を高めるだろうと考えます。さらに、昨年多くの研究が行われました。そこでは、英語教育についての見直しが評価され、その大半が、新しいテストを採用することに同意しています。72.5%の保護者が、試験において話す・書くに焦点を向けたこの改革に期待しています。学校での指導において、話す・書くための指導にポイントが置かれるでしょう。また、保護者は、この新しい試験により、子どもが英語をよりよく話し・書けるようになると思っています。私どもは、この新しい試験によって学校における会話および記述のためのスキルが高まり、そして、より実践的で効果的な技術を採り入れて、学校における英語教育へポジティブな影響を与えるだろうと考えております。私の発表は以上となります。ご清聴どうもありがとうございました。

National English Ability Test

Jin, Kyung-Ae
(Korea Institute for Curriculum & Evaluation)

1. Backgrounds

A. Backgrounds



- Students lack practical English skills
- Need to enhance practical English education and students' speaking and writing skills in English
- Current CSAT (College Scholastic Ability Test) does not assess speaking and writing skill
- Need to develop National English Ability Test (NEAT) including speaking and writing component

1. Backgrounds

CSAT & NEAT

	CSAT	NEAT
Skill Assessed	Listening, Reading	Listening, Reading, Speaking, Writing
Item Type	Multiple Choice	Listening, Reading: Multiple Choice Speaking, Writing: Performance Assessment
Scoring	Norm-Referenced	Criterion-Referenced
Score Report	1-9 Stanine Score Standard Score Percentile Score	Achievement Standards: Excellent, Average, Basic, Below Basic
Implementation	Paper & Pencil One opportunity for each student	Internet-Based (VDI) Two opportunities for each student
Test Paper	One	Level 2(Academic) Level 3(Practical) (Students can choose)

1. Backgrounds

- Establishment and announcement of innovative plans for English education: Nov.
- Announces a new plan to develop a new national English ability test as an assessment reform effort to improve student English ability in speaking and writing for an "enhancement of practical English education."
- Establishment and announcement of a basic plan to develop NEAT: July 2007
- (For student user: Level 2, 3) Provide national educational standards for English education aligned with the curriculum.
- (For general user: Level 1) Substitute foreign tests being used for domestic purposes, assess four modalities in balance, and attempt to obtain international comparability.
- President's election promise and a core government's Feb.
- Introduce a new national English ability assessment emphasizing practical English skills as a substitute for the College Scholastic Ability Test.

1. Backgrounds

B. Field-tests

- NEAT field-tests ('09.5~'12)
 - * 1st (33 schools, 4,300 students) → 2nd(43 schools, 5,300 students) → 3rd(78schools, 10,600 students)
- Public Hearing for NEAT ('10.6)
- NEAT field-tests('10.9~'11.3)
 - * 4th(34 schools, 2,900 students) → 5th(169schools, 20,000 students) → 6th (16 schools, 1,900 students) → 7th (79 schools, 10,000 students)
- Final Public Hearing ('11.5)

2. Goals and Development directions

Goals

- Students will acquire basic communicative skills through school education
- Students will select either NEAT Level 2 or Level 3 based on their aptitude and future career

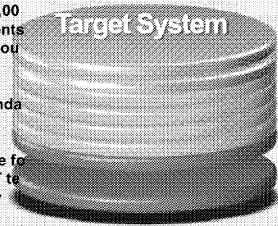
Directions

- Level 2: basic academic English
Level 3: practical English
- Align the contents of NEAT with the National curriculum to reform school education
- Provide teacher training programs focuses on item development, scoring and instruction with regard to the new test

3. Test Development

A. Target System

- Number of examinees and the tests : 1,200,000 and 24 tests(two opportunities per students , 50,000 students taking the test simultaneously)
- Secure item writers and raters: 5,000 secondary teachers
- IBT system: Headquarter(At Korea Institute for Curriculum and Evaluation)and 1,700 IBT test center('11~'13) throughout the country



3. Test Development

B. Timetable

Year	2009	2010	2011	2012	2013	2014	2015
Basic plan	Development and Pre-tests			General implementation			
NEAT Level 2 & Level 3	Pre-tests (3)	Pre-tests (2)	Pre-tests (3)	General Implementation (Pilot application for College entrance)	General Implementation	General Implementation	Implementation (24 times, 120 examinees)
(College Scholastic Ability Test)		Notice two versions of the test: A, B type	<Notice	Policy making on replacing CSAT	<Notice	Period>	Abolition of CSAT (Upon decision)
				Period>	A, B type, Increase listening items		

3. Test Development

C. Development Policy

- Develop two versions of NEAT: test Level 2(Basic Academic English)and Level 3(Practical English)
- New English Curriculum (2009 revision) has been currently being developed aligned with NEAT
 - ※ New curriculum : '11, 8 notification → '11, 9 textbook development → '13 Applied in high school 1st year → '15 NEAT application for college entrance

3. Test Development

<Key Characteristics>

Level 2		Level 3
• Require the degree of English proficiency necessary for college or university education	Objective	• Focus on basic and practical English ability.
• No grammar items • Focus on communicative skills • Topics on Basic Academic English	Characteristics	• No grammar items • Focus on communicative skills • Topics on Practical English: everyday life, work
•3000 words	Vocabulary	•2000 words
•English I & II, English reading comprehension & writing, Advanced English conversation	Agreement with curriculum	•General English, and Practical English conversation

3. Test Development

A. Format

In listening and reading domains, anchor items for test equating are included

Test Domains	Number of items & Test period		
	Level 2	Level 3	Test period
Listening▷	35▷	35▷	35min▷
Reading▷	35▷	35▷	60min▷
Speaking▷	4▷	4▷	15min▷
Writing▷	2▷	4▷	35min▷
Total▷	76▷	78▷	145min▷

3. Test Development

B. Topics : Practical English

Topics	Subtopics
Practical English	1 Everyday life: transportation, telecommunication, shopping, housing, restaurants, hospitals, beauty shops and etc.
	2 Travel & leisure: reservation transportation, concerts, exhibition, sports, hobbies, cooking, hotels, public places
	3 Family & school life: class, friends, birthday party, graduation, course learning, homework, class schedule, examination, library, test score
	4 Work related: forms, documents, employment, salary, marketing, announcement, notice, advertisement, manual,

3. Test Development

B. Topics : Basic Academic English

Topics	subtopics	
Basic Academic English	1	Humanities, sociology, politics, economy, history, education
	2	Science, technology, computer, information & communication, space, ocean, environment, expedition
	3	Art, literature, anthropology, philosophy
	4	Labor, job, career, gender equality, aging society, social welfare, population, juvenile problem
	5	Culture, public morality, public order, civil life, voluntary service, cooperation

3. Test Development

A. Behavioral Domain: Reading & Listening

Section	Behavioral skill	Level 2	Level 3
Listening	Literal understanding	60%	68%
	Inferential understanding	40%	32%
Reading	Literal understanding	32%	44%
	Inferential understanding	56%	48%
	Comprehensive Understanding	12%	8%

- More inferential and comprehensive understanding items in Level 2 than Level 3
- More literal understanding items in Level 3

3. Test Development

B. Number of Words (Listening)

Text	Number of items per text	Number of words per text	
		Level 2	Level 3
Dialogue	1	110 – 130	80 – 100
	2	160 – 180	–
Monologue	1	90 – 110	70 – 90
	2	140 – 160	90 – 110

3. Test Development

C. Number of Words (Reading)

Text	Number of items per text	Number of words per text	
		Level 2	Level 3
Single paragraph	1	130 – 150	110 – 135
Multiple paragraphs	2	230 – 250	170 – 200
	3	320 – 350	210 – 240

3. Test Development

D. Item Types

- Charts, pictures, graphs and authentic texts are adopted to enhance authenticity of the test and align with the curriculum and textbooks

(1) Listening

Level 2		Level 3	
Item Type	%	Item Type	%
Appropriate answer	10–15%	Appropriate answer	15–20%
Theme, Title, Main idea, Purpose, Opinion	30–35%	Theme, Title, Main idea, Purpose, Opinion	25–30%
Contents agreement, Request, Things to do	35–40%	Contents agreement, Request, Things to do, Find reason	35–40%
Find location, tables and information	15–20%	Find appropriate picture, Find location, tables and information	15–20%

3. Test Development

(2) Reading

Level 2		Level 3	
Item Type	%	Item Type	%
Theme, Title, Main idea, Purpose, Assertion	30–35	Theme, Title, Main idea, Purpose, Assertion	30–35
Literal Information(Contents Agreement)	30–35	Literal Information(Contents Agreement)	35–40
Infer blanks	15–20	Infer blanks	5–10
Sentence insertion	5–10	Find order of pictures or sentences	10–15
Find the order of contents	5–10	Infer meaning of phrases and pronouns	10–15
Infer meaning of phrases and pronouns	5–10		

3. Test Development

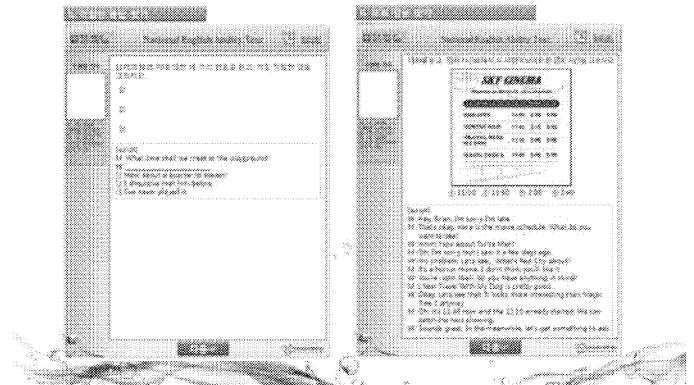
(3) Speaking

Level 2		Level 3	
Item Type	Number of Items	Item Type	Number of Items
Related Interviews	1(4)	Related Interviews	1(4)
Problem Solving	1	Problem Solving	1
Picture Description	1	Picture Description	1
Presentation	1	Speak based on pictures	1(3)
Total	4	Total	4

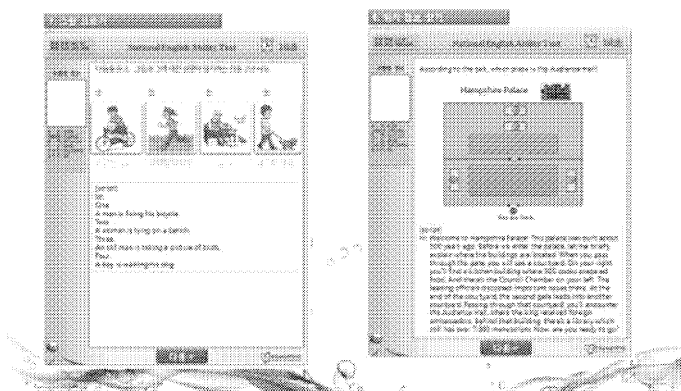
(4) Writing

Level 2		Level 3	
Item Type	Number of Items	Item Type	Number of Items
Guided Writing	1	Selective Picture Description	1
		Complete Detailed Picture description	1
Short Essay	1	E-mail Writing	1
		Writing based on pictures	1
Total	2	Total	4

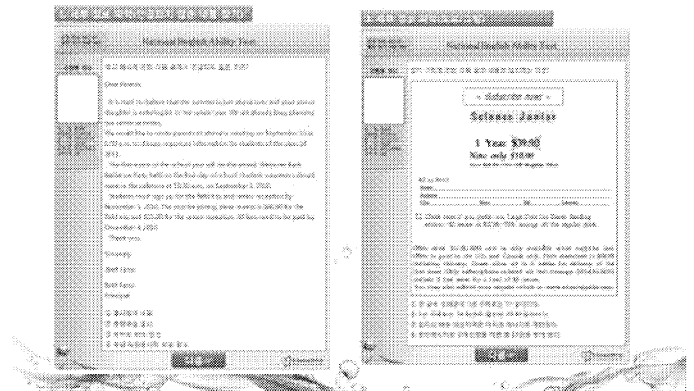
Listening



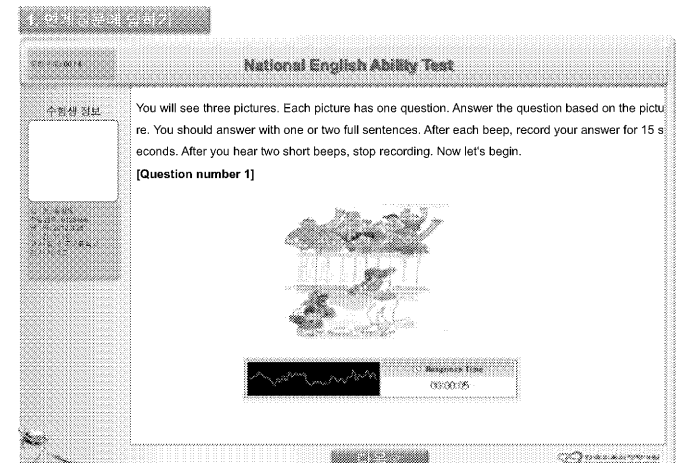
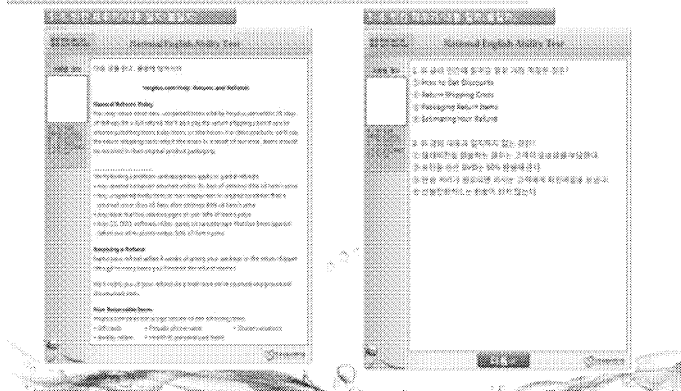
Listening



Reading



Reading



2. 그림 묘사하기

National English Ability Test

수험생 정보

You will see six pictures and tell a story based on the pictures. You have 1 minute to prepare your answer. After the beep, you will have 1 minute to record your answer. The story should start with "One day, a man was ..." After you hear two short beeps, stop recording. Now let's begin.

Response Time: 00:00:05

3. 문제 해결하기

National English Ability Test

수험생 정보

You will hear a story describing a problem. You will have 1 minute to think about how you would solve the problem. After the beep, you will have 1 minute to record your answer. After you hear two short beeps, stop recording. Now let's begin.

One of your friends asks you if you can see a movie on Saturday. You want to go but you can't because you have to go to your grandparent's house that day. However, you can go see a movie with your friend on Sunday. In this situation, what would you say to your friend?

Response Time: 00:00:05

3. 문제 해결하기

National English Ability Test

수험생 정보

You will hear a story describing a problem. You will have 1 minute to think about how you would solve the problem. After the beep, you will have 1 minute to record your answer. After you hear two short beeps, stop recording. Now let's begin.

One of your friends asks you if you can see a movie on Saturday. You want to go but you can't because you have to go to your grandparent's house that day. However, you can go see a movie with your friend on Sunday. In this situation, what would you say to your friend?

Response Time: 00:00:05

Writing

National English Ability Test

작성할 내용을 입력하십시오. (00:00:00)

당신의 문장의 주제를
당신의 문장의 주제
당신의 주제를 설명하십시오

Response Time: 00:00:05

4. IBT System

A. Cloud Computing IBT System

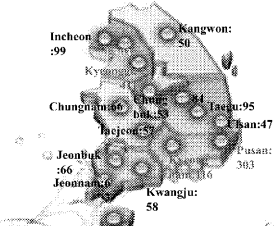
- 1) Setting up VDI (Virtual Desktop Infrastructure) at the Central Center
- 2) The students log in on their monitor and take the test at the test center
- 3) The client in the test center download the test items and uploads the answers
- 4) Minimize effect of PC environment for providing equivalent test environment

4. IBT System

II. 시험 시행 방침

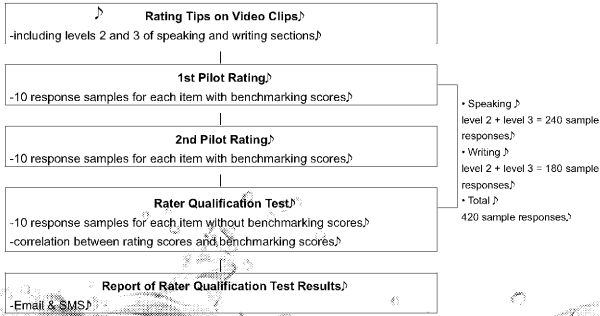
B. Test Center

Year	2011	2012	2013
Number of test sites (for simultaneous administration)	500 labs (21,000 students)	1,200 labs (36,000 students)	1,700 labs (51,000 students)



5. Scoring

Rater Training Process



5. Scoring

Scoring allotment of writing level 2

level2	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
Examinee	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1-260																				
261-520																				
521-780																				
781-1040																				
1041-1300																				
1301-1560																				
1561-1820																				
1821-2080																				
2081-2340																				
2341-2600																				

5. Scoring

* Inter-rater reliability(5th Field Test, *10.12)

	Form A	Form B	Form C	Form D	Form E
Speaking Level 2	0.859	0.785	0.890	0.840	0.843
Speaking Level 3	0.856	0.794	0.882	0.873	0.872
Writing Level 2	0.895	0.841	0.891	0.847	0.874
Writing Level 3	0.832	0.862	0.830	0.857	0.864

6. Score Report

Score Report: Provide performance level for listening, reading, speaking and writing

1) Criterion-referenced test

- Excellent, Average, Basic and Below Basic
- Standard Setting (Bookmark Method)

2) Test equating

- Students' scores of different tests are equated
- non-equivalent group design with external anchor items

5. Score Report

Item Analysis

- Items are psychometrically analyzed to determine possible issues with the items.
- Item statistics are used as feedback during standard-setting studies, can be used during the test form assembly process later, and both item and test statistics are relied on for equating.

5. Score Report

Item Analysis

Test items were analyzed with procedures from Classical Test Theory (CTT) and Item Response Theory (IRT) approaches.

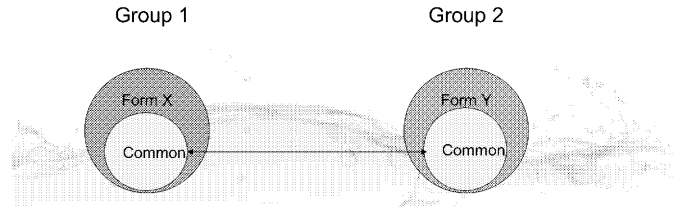
		Classical Test Theory				Item Response Theory					
		difficulty		discrimination		difficulty		discrimination		guessing	
		L-2	L-3	L-2	L-3	L-2	L-3	L-2	L-3	L-2	L-3
Listening	A	55.60	69.34	.38	.42	.15	-.36	1.08	1.33	.17	.23
	B	55.95	68.91	.43	.43	.14	-.37	1.17	1.36	.18	.23
Reading	A	49.90	68.55	.43	.48	.52	-.15	1.41	1.62	.21	.26
	B	48.57	64.99	.42	.49	.68	-.18	1.41	1.61	.22	.23

Equating

- Number of examinees and the tests :
 - 1,200,000 and 24 tests (two opportunities per student s, 50,000 students taking the test simultaneously)
- Multiple forms will be administered and need to be equated to each other.
- "Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content." (Kolen & Brennan, 2004, p.2)

Common-Item Nonequivalent Groups Design for Equating

- Score on common items indicate how performance of Group 1 and Group 2 differ
- The common items must be the same in Form X and Form Y
 - "mini version" of the test form (proportionally represent test content)
 - a similar location (item number)
 - exactly the same (no wording changes or rearranging of alternatives)



Standard Setting

- The process by which performance cut points (cut scores) are established on an assessment
- Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001)
 - Standard setters (panels) evaluated specially formatted test booklets and placed bookmarks at points where the difficulty of items appeared to change in ways that differentiated between adjacent performance levels.

Panel Participants

- A panel of 20 faculty, administrators, and teachers participated for each area.
- Panels select the most difficult item a borderline student would be likely to answer correctly and place a 'bookmark' at that location.

Score Reporting

- Based on test results, students are placed into one of the following four proficiency levels: A, B, C, and F, F being the lowest performance level and A being the highest.
- Scores are provided in the four sections: Listening, Reading, Speaking, and Writing. Separate proficiency levels and PLDs are reported for each section.

Score Reporting

□ Listening Cut Score

		Fail/C	C/B	B/A
Level 2	Round 1	-0.364	0.118	0.939
	Round 2	-1.584	0.099	0.796
	Round 3	-1.584	0.038	0.796
Level 3	Round 1	-1.448	-0.395	0.174
	Round 2	-1.448	-0.348	0.153
	Round 3	-1.448	-0.348	0.153

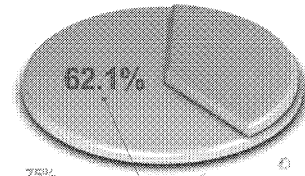
Report Card (Sample)

学期总结
 学期评语
 家长寄语
 教师寄语
 学生寄语
 家长寄语
 教师寄语
 学生寄语

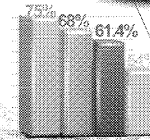
姓名	学号	性别	出生年月	民族	籍贯	联系电话	电子邮箱
张明	1001	男	2000-01-01	汉族	江苏南京	13800138000	zhangming@163.com
李华	1002	女	2000-02-02	汉族	浙江杭州	13900139000	lihua@163.com
王强	1003	男	2000-03-03	汉族	山东青岛	13700137000	wangqiang@163.com
陈静	1004	女	2000-04-04	汉族	广东广州	13600136000	chenjing@163.com
赵伟	1005	男	2000-05-05	汉族	河南郑州	13500135000	zhaoweili@163.com
孙丽	1006	女	2000-06-06	汉族	四川成都	13400134000	sunli@163.com
周涛	1007	男	2000-07-07	汉族	湖北武汉	13300133000	zhoutao@163.com
吴敏	1008	女	2000-08-08	汉族	湖南长沙	13200132000	wumin@163.com
郑凯	1009	男	2000-09-09	汉族	福建厦门	13100131000	zhengkai@163.com
徐悦	1010	女	2000-10-10	汉族	江西九江	13000130000	xuyue@163.com

2. Conclusions

A. Survey (2009, Jin et al.)



With 500 respondents, 62.1% said NEAT will enhance students' speaking and writing ability

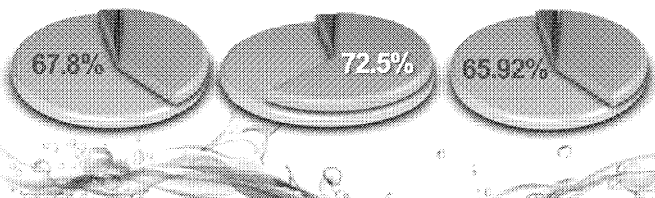


(English Education Specialists: 75%, Non-English Teachers: 68%, Parents: 61.4%, English Teachers: 54%)

2. Conclusions

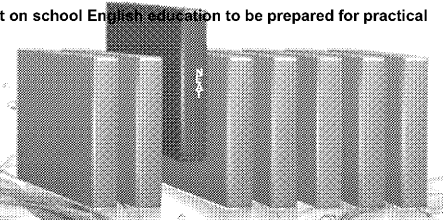
B. Evaluation of English Education Policy(2010, Kim, Jeon, Woo & Jin)

- 67.8% agree to adopt NEAT
- 72.5% of parents expect innovation of speaking and writing instruction at school
- 65.92% of parents expect that students will speak and write English better



6. Conclusions

- Enhance students' speaking and writing skills in addition to listening and reading through adopting NEAT
- Students can select different tests (Level 2 or Level 3) based on their future career and aptitude
- Through criterion-referenced test, the students can have clear performance standards
- Positive wash back effect on school English education to be prepared for practical English Education



Thank you

New Approaches to Educational Testing And ACT

Deborah J. Harris

ACT, Inc.

The intent of this paper, and the accompanying talk, is to provide information about some of the products and services ACT, Inc. provides. The focus will be primarily on educational testing, and how ACT is continuing to provide cutting edge information and guidance for students and educators, both in cognitive and noncognitive areas. ACT's expertise in the workforce area is also discussed briefly. The information provided is a compilation of research, products, and services ACT provides; additional information on the topics discussed may be found through the links listed at the end of the paper.

Overview

ACT, Inc., an independent, not-for-profit organization, was founded in 1959 around a single assessment program. Today, ACT, Inc. has offices across the country and internationally, and serves millions of people in high schools, colleges and universities, professional associations, businesses, and government agencies.

ACT, Inc.'s international involvement began decades ago with the administration of the ACT Assessment outside the United States. The ACT is now administered in over 135 countries; the number of colleges and universities outside the United States that use the ACT as part of their admission process has more than doubled in the past two years.

The growth of ACT's services specifically designed to serve international markets has been dramatic. ACT Education Solutions, Limited, (AES) now has offices in Sydney, Jakarta, Shanghai, and Singapore. AES offers several training programs, including the Global Assessment Certificate (GAC), which is aimed at helping students in non-English-speaking countries prepare for college and university undergraduate study in English-speaking countries, becoming the world's most widely recognized university preparation program for students whose first language is not English. There are over 90 GAC Teaching Centers in 12 countries, including China. More than a hundred Pathway Universities, located in the United States, Canada, United Kingdom, Ireland, Australia, Singapore, New Zealand, and Mexico, admit students who have earned the Global Assessment Certificate.

This paper will provide an overview of a select few of ACT, Inc.'s products and services in the education and workforce areas, beginning with

Note: In the U.S., colleges and universities set their own admissions criteria. High school students who wish to attend college usually take an assessment, such as the ACT, as part of their application process. However, some students may choose to take the SAT, or both the ACT and the SAT. Individual colleges and universities usually do not have additional institution-specific entrance exams. In addition, the ACT and the SAT are given multiple times throughout a year, and an individual student may take one or both exams multiple times.

The ACT

The ACT Assessment program is a comprehensive system designed to help high school

students develop postsecondary educational plans and to help postsecondary educational institutions meet the needs of their students. The ACT battery includes four multiple choice tests of educational achievement—English, Mathematics, Reading, and Science—and an optional Writing Test. The ACT also collects self-reported information about students' high school courses and grades, educational and career aspirations, extracurricular activities, and special educational needs.

ACT data are used for many purposes, by many users. High schools use ACT data in academic advising and counseling, evaluation studies, and accreditation documentation. Colleges and universities use ACT results for admissions and course placement. States may use the ACT as part of their statewide assessment. Many of the agencies that provide scholarships, loans, and other types of financial assistance to students use the ACT as a measure of student qualifications.

The ACT functions both as a stand-alone program and as part of the secondary school level of ACT's Educational Planning and Assessment System (EPAS.)

EPAS[®]

The ACT Educational Planning and Assessment System (EPAS[®]) is an integrated series of assessment and career planning programs – EXPLORE (grades 8 and 9), PLAN (grade 10), and the ACT (grade 11 and 12) – that is designed to help students increase their academic readiness for college and post secondary training. The system provides a longitudinal, systematic approach to educational and career planning, assessment, instructional support, and evaluation.

All three programs, EXPLORE, PLAN, and ACT, include curriculum-based assessments in English, Mathematics, Reading, and Science that have been empirically tied to postsecondary success in the U.S. These content tests measure what students are able to do with what they have learned in school and what they need in order to be college- and work-ready when they graduate from high school. In addition, scores on the content tests are reported on the same score scale across the batteries: ACT scores are reported on a scale from 1 to 36, while the maximum scores are 32 for PLAN and 25 for EXPLORE. Through the EPAS system, ACT has established the nation's largest longitudinal data system that monitors student progress from the 8th grade through college, so that the level of achievement and readiness students attain in K–12 can be compared and evaluated against their actual success in postsecondary education.

The U.S. College Readiness Benchmarks

ACT's College Readiness Benchmarks are the scores required on the ACT subject tests for high school students to have approximately a 75 percent chance of earning a grade of C or better, or a 50 percent chance of earning a grade of B or better, in selected credit-bearing courses commonly taken by first-year college students: English Composition; College Algebra; Biology; and social sciences courses such as History, Psychology, Sociology, Political Science, or Economics. Benchmarks for EXPLORE and PLAN were established as indicators of a students' progress toward meeting the Benchmarks on the ACT. The College Readiness Benchmarks were empirically derived based on the actual performance of students in first-year credit-bearing college courses, using data from 98 two-

and four-year institutions from all over the country and over 90,000 students. The Benchmark scores are provided in Table 1.

Table 1. ACT's College Readiness Benchmarks

TEST	EXPLORE	PLAN	ACT
English	13	15	18
Mathematics	17	19	22
Reading	15	17	21
Science	20	21	24

Forty-seven percent of all 2010 high school graduates in the United States—nearly 1.6 million— took the ACT during their high school career. Of these ACT-tested graduates, 24 percent met or surpassed all four of the ACT College Readiness Benchmarks, up from 21 percent in 2006 and from 23 percent in 2009. The percent of graduates ready to succeed in college coursework remains highest in English (66 percent), followed by reading (52 percent), mathematics (43 percent), and science (29 percent), indicating there is substantial room for improvement in college and career readiness. An important question in preparing all students for college and career by the time they graduate from high school is that of determining how much growth in academic achievement typically occurs during high school, and whether such growth can be accelerated to ensure that more students are ready for college and career when they graduate.

Growth

A sample of approximately 150,000 students who were administered EXPLORE, PLAN, and ACT was used for an initial look at growth issues. The average scores across the three assessments for each subject test are shown in Figure 1 at the top of the bars. Because the scores are reported on the same scale for each subject, across batteries, the average growth for these students between 8th grade and 12th grade is easily seen, ranging from 3.3 score points on the Science Test to 5.6 score points on the Reading Test.

The total sample of students was separated into whether they were on target for becoming college and career ready (i.e., those who met or exceeded the College Readiness Benchmarks for EXPLORE in the 8th grade), whether they were close to being on target (i.e., those who were within 2 or fewer score points of meeting each EXPLORE Benchmark), and those who were off target (i.e., those who were more than 2 score points from meeting each EXPLORE Benchmark). The average scores for these three groups in math are shown in Figure 2.

Figure 1: Average Growth in Achievement between Eighth and Twelfth Grades

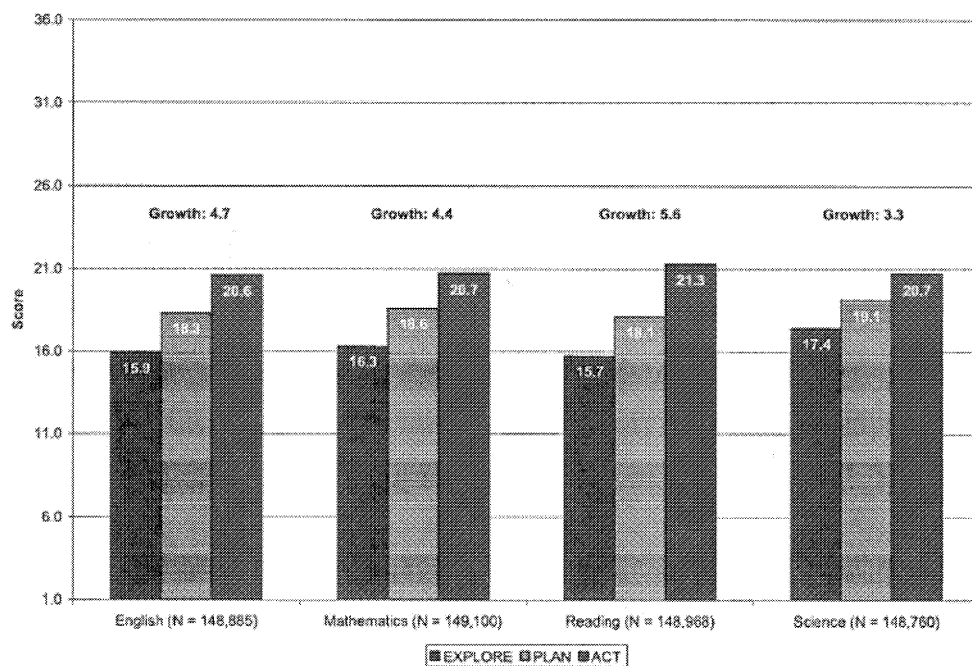
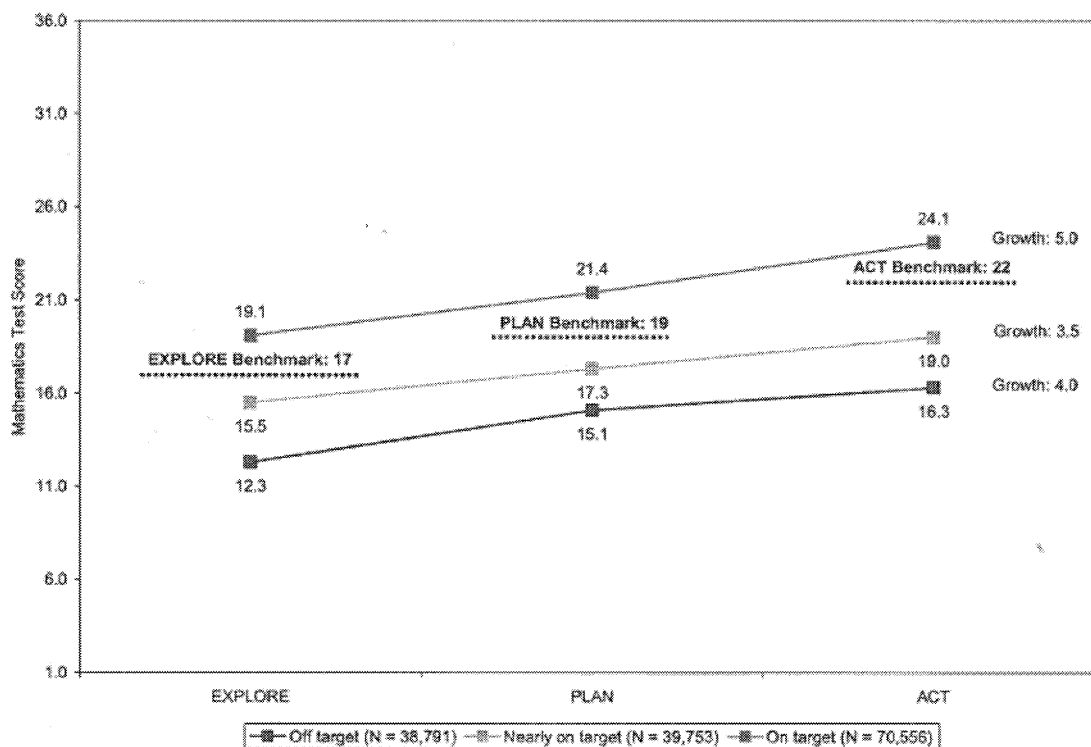


Figure 2: Average Growth in Achievement between Eighth and Twelfth Grades, by Degree of College Readiness



The average growth was greatest for the group of students who were on target for college and career readiness in 8th grade. Individual students' growth goals can be set using the

College Readiness Benchmarks and the growth trajectories as a yardstick. For students who are off target in 8th grade, a challenging yet reasonable goal on successive tests would be to reduce by half the difference between the student's score in a given subject and the corresponding College Readiness Benchmark. A second approach to set growth goals is by first measuring the average growth at high-performing high schools (i.e., schools showing the greatest growth) and then setting goals for students at lower- and average-performing high schools according to what is considered normal growth at the high-performing high schools.

Students who are significantly off target for college and career readiness in 8th grade are far less likely to become ready for college-level work during high school; therefore, academic interventions will be necessary for these students in order to help them attain the foundational academic skills that are necessary for college and career readiness.

ACT can also look at setting readiness targets below EXPLORE, by linking to state grades 3-7 assessments. This is typically done by calculating a comparable score on the 7th grade assessment to the College Readiness Benchmark on EXPLORE. Backmapping is then used to find the 3rd through 6th grade scores that indicate a student is on track. Table 2 displays what a College and Career Readiness (CCR) Ramp might look like. In addition, Table 3 illustrates what yearly goals, called NCEA Growth Goals, might look like -- defining a path for a student to follow to reach the CCR ramp.

Table 2. Example of College and Career Readiness Targets

Backwards-Mapping the College and Career Readiness Targets

ACT/NCEA continues this process down to the lowest test grade—typically Grade 3. The trajectory defined by College and Career Readiness Targets from Grades 3-7 and ACT's College Readiness Benchmarks from Grades 8-12 creates what is known as the College and Career Readiness Ramp.

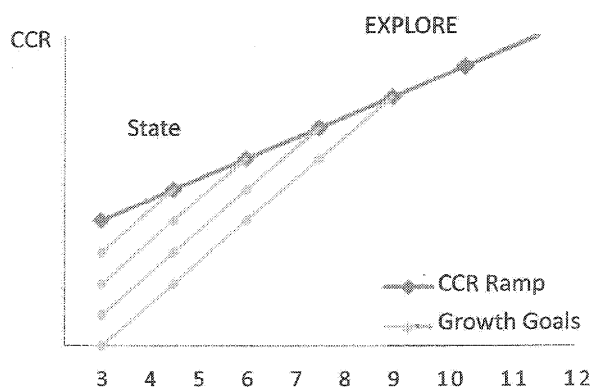
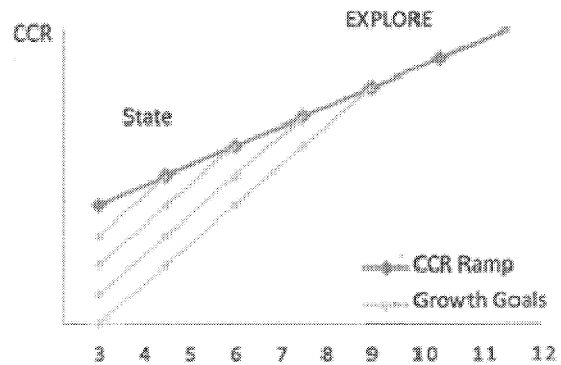


Table 3. Example of Yearly Growth Goals

Establishing Students' Yearly Growth Goals

ACT/NCEA then identifies yearly growth goals that a student must achieve in order to get themselves onto the College and Career Readiness Ramp. These yearly goals are known as NCEA's Growth Goals, and they define a path for a student to reach the College and Career Readiness Ramp in no more than four years.



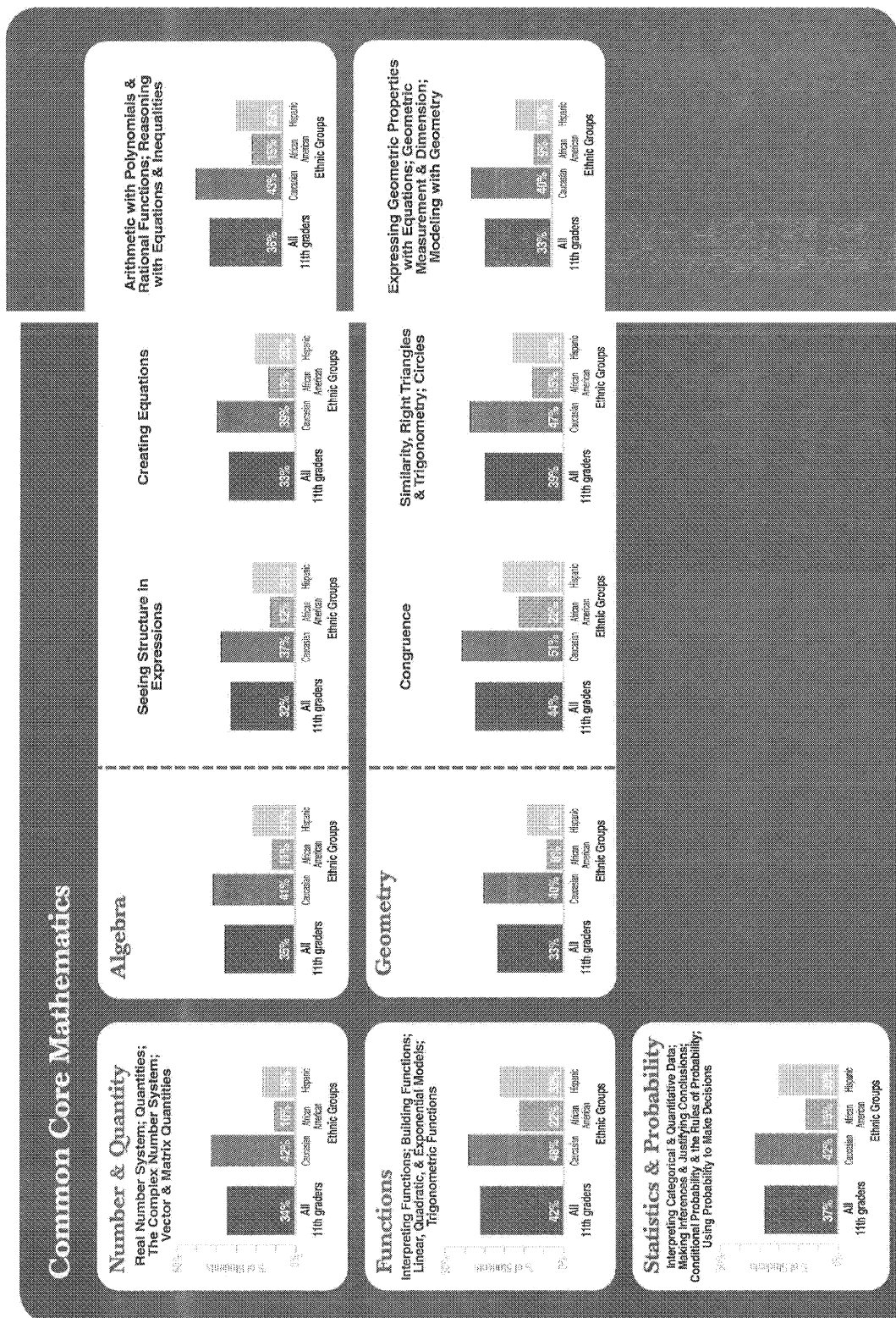
Common Core

The Common Core State Standards Initiative represents a significant reform to U.S. education, leading to consensus in the majority of states and territories on the essential knowledge and skills necessary for the college and career readiness of all students. ACT is pleased to have played a leading role in the development of the Common Core State Standards, with both ACT's longitudinal research identifying the knowledge and skills essential for success in postsecondary education and workforce training, and ACT's College Readiness Standards™ being among the resources used in the creation of the Common Core State Standards.

To provide an initial look at how high school students might perform on the Common Core, ACT used a sample of a quarter-million typical 11th grade high school students who had taken the ACT, and coded the items they were administered into Common Core clusters. Since performance standards have not yet been established for the Common Core State Standards, ACT used its research-based College Readiness Benchmarks to estimate college- and career-ready performance levels. For each of the clusters of Common Core State Standards for which ACT has data (i.e., all but Speaking & Listening and Research), the percentage of students in the sample who met or exceeded the performance level of college ready was computed. These analyses serve as a starting point for assessing achievement relative to the Common Core in advance of full state implementation efforts. Figure 3 shows the overall percentage of students in the sample who met ACT's College Readiness Benchmarks in mathematics, as well as the percentage of selected subgroups, indicating that across all Common Core domains, strands, and clusters, only one third to one-half of the high school students are reaching a college and career readiness level of achievement.

The period of time between Common Core adoption and Common Core implementation offers an important opportunity to evaluate and reframe education policy and practice at all levels. ACT believes these analyses results provide information that stakeholders can use to understand the current state of college and career readiness of students and to begin implementing programs and policies that best support the Common Core.

Figure 3: A First Look at Common Core Mathematics



International Benchmarking

ACT conducted a study using PLAN and Programme for International Student Assessment (PISA) to determine the relationship of the ACT College Readiness Benchmarks in reading and mathematics to the performance of participating countries from the Organisation for Economic Co-operation and Development (OECD), for the purposes of improving the comparison of U.S. student performance with global student performance. Although used with different populations and administered under different administration conditions, there are many striking similarities between PLAN and PISA. Both programs address the content areas of reading, mathematics, and science. Both PLAN and PISA assess what students can do with what they have learned in school, and they each focus on measuring higher-order critical thinking skills that are important for life after compulsory education. A more detailed comparison can be found in <http://www.act.org/research/policymakers/pdf/AffirmingtheGoal.pdf>.

The U.S. faces the challenge of adapting to the demands of a globalized economy, where jobs have become more specialized and more driven by technology, requiring higher levels of education and training—especially in mathematics and science. The U.S. workforce now competes internationally to a far greater extent than in the past, and international comparisons of academic achievement show U.S. students at a deficit when compared to students in many other nations. In order to determine if current standards are rigorous enough to result in preparation and achievement of students at the level necessary for international competitiveness, the relationship of international performance to U.S. college readiness performance needs to be examined. It is only when one can compare international student performance to a relevant performance standard like college readiness that one can understand whether the U.S. standard of performance for entry into credit-bearing, first-year college courses is a globally competitive performance standard. By comparing the performance standard of U.S. college readiness to international student performance, one is able to anchor actual student performance in credit-bearing college courses to international performance, allowing one to know how much improvement in student preparation is needed to be globally competitive.

ACT was uniquely positioned to conduct this research: by using the ACT College Readiness Benchmarks, it is possible to determine how the U.S. standards of college readiness compare to the average performance of students in OECD countries internationally. Because the ACT College Readiness Benchmarks are empirically based on actual student performance in a nationally representative sample of two- and four-year postsecondary institutions, they can serve as an objective U.S. performance standard for college readiness. By linking PISA results to the Benchmark levels (via PLAN testing), it can be determined if the U.S. standard of college readiness is above, below, or the same as the average performance of like students in OECD countries.

In fall 2009, following the national assessment of PISA in the U.S., a sample of students was selected using a two-stage sampling design: schools were selected with probability proportional to size, and a random sample of students in the correct age range was chosen. Each student in the study was administered one PLAN battery and one PISA booklet, within a window of 4 months. ACT scored the PLAN tests according to standard procedures, and although these students were not included in the U.S. national PISA administration, the tests were administered in a manner consistent with the national program. ACT was responsible for coding and scoring PISA items and the Australian Council for Educational Research (ACER) was responsible for generating the PISA plausible values.

The analysis goal of the study was to link the PLAN College Readiness Benchmarks to the PISA tests. Given the score distributions of both PISA and PLAN tests for the study sample, each PLAN College Readiness Benchmark was linked to the corresponding PISA test using the traditional un-smoothed equipercentile method. Because there were five sets of plausible values for each subject, the linking was conducted five times, each using a different set of plausible values; the final reported linkage was the average of the five values. Linking variances were also estimated following the guidelines on conducting analysis using plausible values as described in PISA technical manuals. In addition, a series of cross-validation studies were conducted on additional linkages using publically available PISA data and PLAN data. Public PISA data was from PISA testing cycles 2003, 2006, and 2009. PLAN public data came from the PLAN norm distribution for Grade 10 in fall 2005. Linkage results from these additional sources were consistent with the current study.

Table 4 presents the linking results for mathematics. For mathematics, the PISA score that is equivalent to the PLAN College Readiness Benchmark is 530.1 (the average of the five linkages conducted on the five plausible values), which falls in a PISA mathematics literacy proficiency level 3. All three validation linkages also resulted in a PISA level 3. Linking error estimates between the study linkage and validation linkages are very close to each other in that all the 95% confidence intervals overlap with each other (the entire 95% confidence intervals for all four linkages fall in the upper range of PISA level 3).

A verbal description of PISA level 3, and the ACT College Readiness Standards related to the Benchmark score of a score of 19 for mathematics, is presented in Table 5.

Table 4. PLAN College Readiness Benchmark Equivalent on PISA for Mathematics

Linkages	Plausible Value					Average	PISA Level	%95 CI†
	1	2	3	4	5			
Study Sample	530.7	529.1	531.9	529.1	529.7	530.1	3	[523.5, 536.7]
Validations:								
PISA2003	533.0	532.5	535.3	533.3	533.3	533.5	3	[526.9, 540.1]
PISA2006	522.9	520.2	520.2	520.6	520.6	520.9	3	[513.1, 528.7]
PISA2009	534.9	536.7	535.3	535.3	534.2	535.3	3	[526.2, 544.4]

† %95 Confidence Interval = Estimated concordant score \pm 1.96*(linking error)

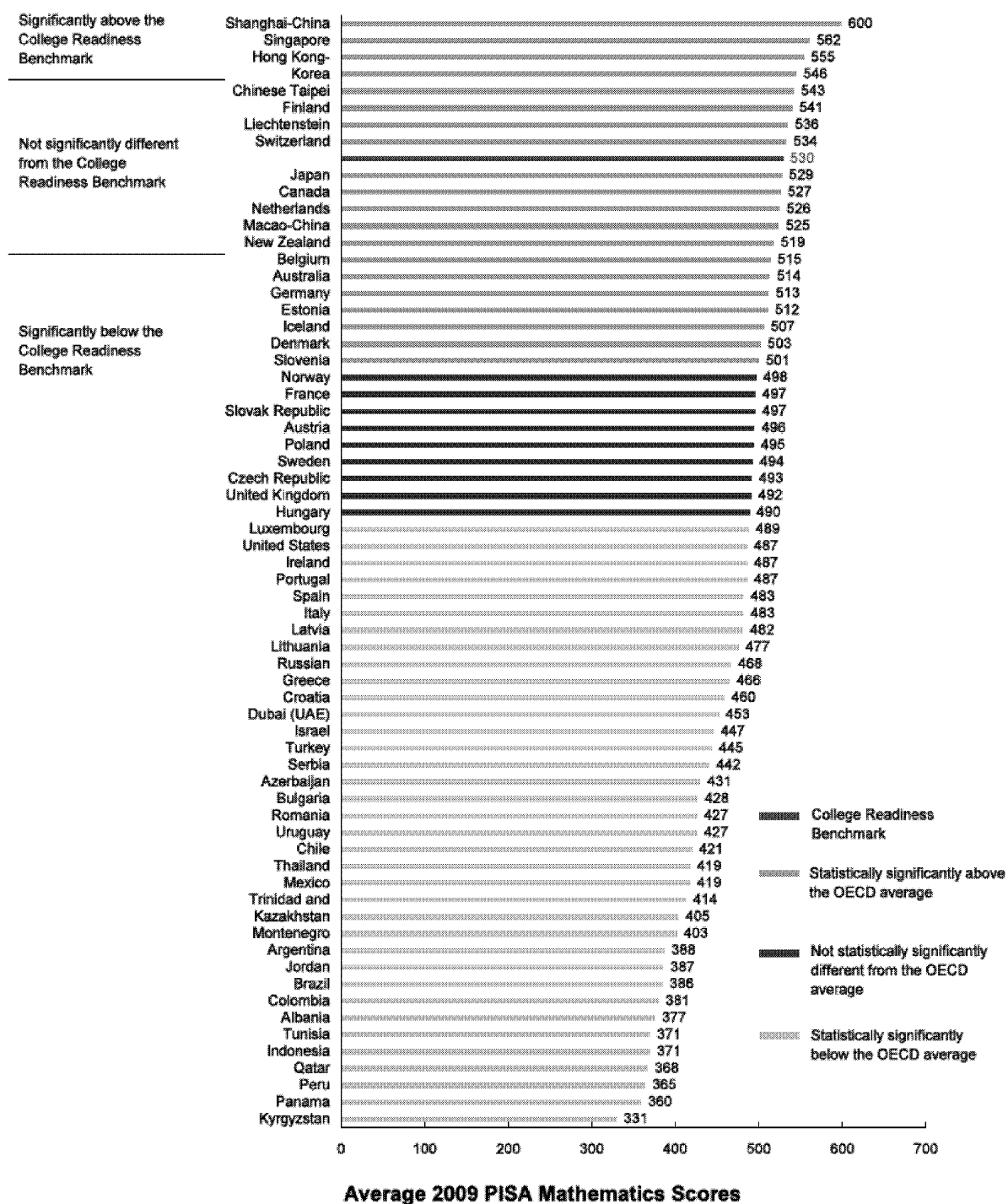
One question of interest is where on the PISA scale a country would rank if it was scoring at the level of the benchmark value; in mathematics, the PISA score of 530 that is comparable to the PLAN mathematics benchmark would rank a country 3rd among the 34 OECD countries, and 9th overall among the 65 entities that took the most recent PISA assessment, illustrating that the College Readiness Benchmark corresponds to a level that is internationally competitive, and students achieving this level would be well placed to succeed in the international marketplace, as illustrated in Figure 4.

Table 5. Descriptors of U.S. College Readiness Standards and Corresponding PISA Proficiency Level

	PISA ¹	U.S. College Readiness Standards
Mathematics	<p>PISA Level 3: At level 3, students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.</p>	<p>Score Range 16-19:</p> <ul style="list-style-type: none"> ▪ Solve routine one-step arithmetic problems (using whole numbers, fractions, and decimals) such as single-step percent ▪ Solve some routine two-step arithmetic problems ▪ Calculate the average of a list of numbers ▪ Calculate the average, given the number of data values and the sum of the data values ▪ Read tables and graphs ▪ Perform computations on data from tables and graphs ▪ Use the relationship between the probability of an event and the probability of its complement ▪ Recognize one-digit factors of a number ▪ Identify a digit's place value ▪ Substitute whole numbers for unknown quantities to evaluate expressions ▪ Solve one-step equations having integer or decimal answers ▪ Combine like terms (e.g., $2x + 5x$) ▪ Locate points on the number line and in the first quadrant ▪ Exhibit some knowledge of the angles associated with parallel lines ▪ Compute the perimeter of polygons when all side lengths are given ▪ Compute the area of rectangles when whole number dimensions are given

¹ Adopted from *Highlights from PISA 2009: Performance of U.S. 15-year-old students in Reading, Mathematics, and Science literacy in an international context*, by Fleischman, Hopstock, Pelczar, Shelley, and Xie (2010).

Figure 4: Tenth-grade College and Career Readiness Performance Benchmark Compared to the Performance of Countries/Economies on PISA 2009 Mathematics



Note. The data is adopted from *PISA 2009 results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*, by OECD (2010b).

Work Readiness

Not all students enter college when they finish their secondary education. Some go directly into the workforce or into a work training program. One of the challenges of the work arena is placing the right person in the right job, a component of which is ensuring a person has the requisite skills for a particular position.

ACT's WorkKeys is an assessment system designed to help employers select, hire, train, develop, and retain a high performance workforce. ACT's WorkKeys system is being used, to some degree, on every inhabited continent on the planet. One component of the system is analyzing what skills are needed in different job. ACT has an occupational profile database with more than 18,000 job titles, ranging from white-collar professional to blue-collar technical positions. Extensive research has been done on these job to identify both the essential skills, and the skill levels, necessary for employee selection and training.

WorkKeys assessments include Applied Mathematics, Locating Information, Reading for Information, Applied Technology, Business Writing, Listening for Understanding, Teamwork, Workplace Observation, and three soft skills assessments: Fit, Performance, and Talent.

The Applied Mathematics assessment is a 33 item test that can be delivered in paper&paper or computer mode, and in English or Spanish. Applied Mathematics is designed to measure the skills people use when they apply mathematical reasoning, critical thinking, and problem-solving techniques to work-related problems. The items require examinees to solve the types of problems and do the types of calculations that actually occur in the workplace. Applied Mathematics has five levels of difficulty, ranging from Level 3, the least complex, to Level 7, the most complex. The levels build on each other, incorporating skills assessed at the previous levels. For example, to solve Level 5 items, individuals need the skills from Levels 3, 4, and 5 because Level 5 items require several steps of logic and calculation (e.g., a problem may involve completing an order form by totaling the order and then computing tax) that necessitate using skills associated with Levels, 5 and lower levels. (See <http://act.org/workkeys/assess/math/levels.html> for a description of the specific skills associated with each level.)

The job profiling process ACT uses has four steps. The first step is to create an initial task list, covering the tasks most relevant to the job. The second step is a task analysis, where the initial list of tasks from step one is reviewed and revised as needed by subject matter experts, and each task is rated according to importance and the relative time spent on the job doing that particular task. This data is then used to produce a criticality rating for each task, which are then reviewed. The end result of step two is a final task list which indicates which tasks are most critical to performing to the job. The third step is skills analyses, where the skills are reviewed. Detailed descriptions of the skills covered in the WorkKeys assessments are presented to the subject matter experts, who determine the relevance of those skills to the job of interest, deciding which skill levels are necessary for entry level. The fourth step is documenting the results in a report which establishes the link between the tasks of the job and the WorkKeys assessments skills and skill levels.

An example of the skill levels required for a sample of job titles, taken from <http://act.org/workkeys/skillsearch.html>, is shown in Table 6. For example, the job title "accountants" was profiled as requiring a Level 6 in Applied Mathematics, a Level 5 in Locating Information, and a Level 5 in Reading for Information. The job title "advertising

sales agents” requires a Level 3 in Applied Mathematics, indicating a lower level of math achievement is needed than that required for “accountants”.

For those examinees who are not functioning at the WorkKeys level they would like to be at, the WorkKeys system also includes a training component. KeyTrain is an interactive tool for career readiness skills, based on a targeted curriculum written specifically to assist people in mastering the applied workplace skills as defined by WorkKeys.

Table 6. Example of Average WorkKeys Scores Needed by Job Title

Job Title	Applied Mathematics	Locating Information	Reading for Information
Accountants	6	5	5
Adjustment Clerks	4	4	4
Administrative Services Managers	4	4	4
Advertising Sales Agents	3	4	4
Agricultural Crop Farm Managers	5	5	4
Aircraft Body and Bonded Structure Repairers	5	5	5
Aircraft Structure Assemblers, Precision	4	4	5

Noncognitive Assessments

ACT and others have conducted research that suggests other factors, such as interests and motivation, have an impact on success, over and above cognitive achievement, in both academic and workforce settings. In addition, these noncognitive factors may help in understanding what types of coursework, careers, and or training programs may work well for a particular individual.

ACT’s Interest Inventory

Because individuals often need assistance in exploring careers options, and in determining how their skills and interests align with various occupations, ACT provides an interest inventory. In addition to use by individuals in identifying careers they may be interested in and skills they may need to acquire, entities such as trade organizations and cities need methods to match potential employees with the needs of current and future jobs.

The ACT Interest Inventory is a component of EXPLORE, PLAN, and the ACT, providing career planning information to over four million individuals every year. The questions in

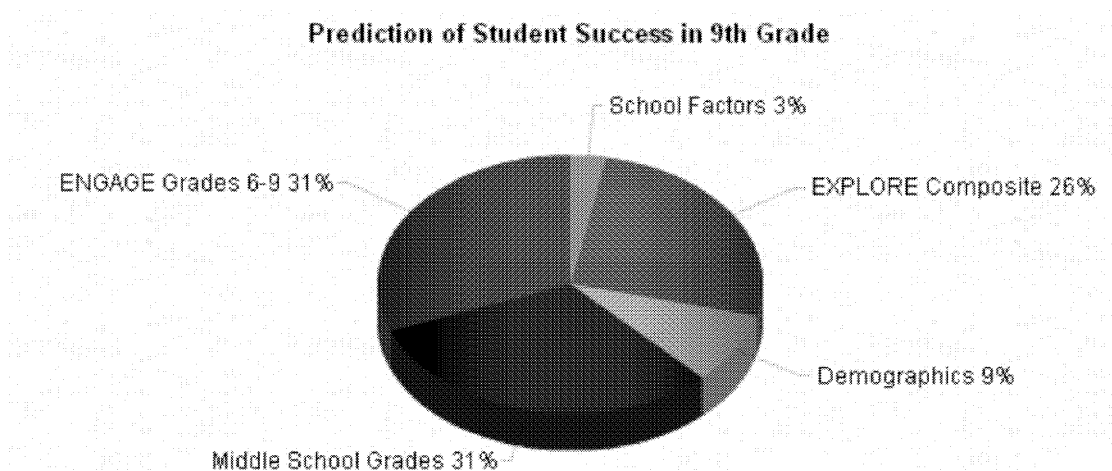
the inventory emphasize work relevant activities (like conducting a meeting or exploring a science museum), that should be familiar to examinees, rather than potentially unfamiliar occupational titles. The ACT Interest Inventory is based on extensive research to ensure reliability and validity, with research using this inventory indicating that the fit between interests and environment (college major or occupation) is related to a diverse set of positive outcomes including persistence in a college major, attainment of a college degree, job earnings, and job satisfaction. The individual Interest Inventory results can help focus career exploration, by pointing to groups of occupations that are aligned with an individual’s interests, rather than singling out one “right” occupation, with results reported graphically on the World-of-Work Map as well as in text.

ENGAGE

ACT has found in its research that 8th grade students' academic achievement has a larger impact on their readiness for college by the end of high school than anything that happens academically in today's high schools. Student readiness is also influenced by psychosocial development: a student's academic readiness for college and career can be improved if the student develops behaviors that are known to contribute to successful academic performance. The ENGAGE assessments provide measures of the academic behaviors that are associated with academic success, and include Academic Discipline, Optimism, Family Involvement, as well as other areas under the broad domains of Motivation, Social Engagement, and Self-Regulation.

ACT has tested thousands of students with ENGAGE, tracking them as they progress through middle school and into high school, finding that ENGAGE administered during middle school was a valid predictor of high school grades. In fact, even after taking into account previous grades and academic readiness (e.g., EXPLORE scores), ENGAGE provided additional information that helped to more accurately identify students who were at risk of poor grades and academic failure. Figure 5 shows that a measure of middle school academic achievement (EXPLORE) and middle school grades used in combination are clearly the best predictors of early high school GPA, however, academic behaviors are also substantial predictors.

Figure 5: Relative Strength of Predictors of Early High School GPA



Note: Based on a linear regression model predicting 9th-grade GPA ($R^2 = .55$).

Table 7 illustrates the importance of both academic readiness and academic behaviors for subsequent student achievement. Student A scored low on both ENGAGE’s Academic Success Index and EXPLORE, and subsequently failed 6 classes and has an extremely low GPA. Student B, who had the same low EXPLORE score (9) but scored high on the ENGAGE Academic Success Index, did not fail any classes and has a GPA of almost 3.0. Student C had a high score on EXPLORE, but scored low on ENGAGE and subsequently failed one class and has a GPA of only 1.56). Student D with the same high score on EXPLORE (21) and a high ENGAGE score failed no classes and has the highest GPA.

Table 7. Four Example Students’ Academic Behavior and Academic Readiness Scores and Later Academic Outcomes.

		<u>Explore</u>	
		Low (Composite = 9)	High (Composite = 21)
ENGAGE Grades 6-9 Academic Success Index (percentile rank)	Low	Student A Success Index = 3 High school GPA = 0.41 Failed high school classes = 6	Student C Success Index = 5 High school GPA = 1.56 Failed high school classes = 1
	High	Student B Success Index = 95 High school GPA = 2.99 Failed high school classes = 0	Student D Success Index = 99 High school GPA = 4.16 Failed high school classes = 0

ENGAGE™ Grades 10-12, and ENGAGE™ College are also available. In addition, ACT has developed a set of scales called ENGAGE Teacher Edition that provides assessments of specific behaviors related to academic success. These behaviors are tangible, observable, and connected to academic performance and other student success outcomes. When the scales and ENGAGE are used in combination, the resulting data allow teachers to identify at risk students early, diagnose individual students’ strengths and weaknesses, identify appropriate curriculum activities and behavioral interventions, assess the effectiveness of those activities and interventions, and track a student’s progress in developing effective academic behaviors.

Summary

The intent of this paper, and the accompanying talk, is to provide information about some of the wide range of products and services ACT, Inc. provides in both the education and workplace arenas. The ACT website offers much more information on the assessments and services ACT has developed, as well as the research that provides the foundation for them.

A First Look at the Common Core and College and Career Readiness
<http://www.act.org/research/policymakers/pdf/FirstLook.pdf>

Solutions for Success in an evolving global market: ACT Annual Report September 1, 2009

<http://act.org/aboutact/pdf/AnnualReport10.pdf>

A Better Measure of Skill Gaps

<http://www.act.org/research/policymakers/pdf/abettermeasure.pdf>

What are College and Career Readiness Targets?

http://www.nc4ea.org/files/what_are_college_and_career_readiness_targets-01-14-11.pdf

Enhancing College and Career Readiness and Success: The Role of Academic Behaviors

http://www.act.org/engage/pdf/ENGAGE_Issue_Brief.pdf

Affirming the Goal Is College and Career Readiness an Internationally Competitive Standard?

<http://www.act.org/research/policymakers/pdf/AffirmingtheGoal.pdf>

How Much Growth toward College Readiness Is Reasonable to Expect in High School?

<http://www.act.org/research/policymakers/pdf/ReasonableGrowth.pdf>

Impact of Cognitive, Psychosocial, and Career Factors on Educational and Workplace Success.

<http://www.act.org/research/policymakers/pdf/CognitiveNoncognitive.pdf>

The Forgotten Middle: Ensuring that all Students Are on Target for College and Career Readiness before High School

<http://www.act.org/research/policymakers/pdf/ForgottenMiddle.pdf>

Enhancing College and Career Readiness and Success: The Role of Academic Behaviors

http://www.act.org/engage/pdf/ENGAGE_Issue_Brief.pdf

Ready for college and ready for work: Same or different?

<http://www.act.org/research/policymakers/pdf/ReadinessBrief.pdf>

Rigor at risk: Reaffirming quality in the high school core

http://www.act.org/research/policymakers/pdf/rigor_summary.pdf

The ACT Technical Manual

http://www.act.org/aap/pdf/ACT_Technical_Manual.pdf

The PLAN Technical Manual

<http://www.act.org/plan/pdf/PlanTechnicalManual.pdf>,

ACT(American College Testing)の現状と教育テスト開発の新たな展開

Deborah Harris

(ACT, Inc., Measurement and Reporting Service 部長)

本稿の目的は、ACT が提供する製品やサービスに関する情報を提供することである。主に、教育テスト、そして、認知的または非認知的な側面の双方において、ACT が如何に学生や教育関係者に最新の情報とガイダンスを提供し続けているか、その方法に焦点を当てている。職業選択支援の領域におけるACT のノウハウについても議論する。ACT が提供する情報は、サービス、製品、そして研究の編纂である。更に、ここでの話題になる情報は、本稿の末尾にリストアップした関連のリンクから見つけることができる。

概要

ACT (非営利団体である) は、1959 年に、一つのアセスメントプログラムを核として設立された。今日、ACT は世界中にオフィスを構えており、高校、単科大学、総合大学、専門機関、ビジネス、及び政府系機関などのためにサービスを提供している。

また、ACT 社の国際的な事業は、数十年前に、ACT Assessment のアメリカ外での運営から始まった。ACT は現在では 135 カ国以上で運営され、アメリカ外の大学で ACT を入学者選抜プロセスの一部に使っているところもある。その数は過去 2 年間で 2 倍以上になった。

ACT のサービスは、国際的なマーケットで役立つようにもデザインされ、めざましい成長を遂げている。ACT Education Solutions, Limited (AES) は、現在、シドニー、ジャカルタ、上海、シンガポールにオフィスを構え、AES は Global Assessment Certificate (GAC) を含む、いくつかのトレーニングプログラムを提供している。それは、英語を母国語としない学生のために、英語圏での学部生としての勉強を支援する事を目的としたものであり、第 1 言語が英語でない生徒の大学進学準備のプログラムという点で、広く世界的に認知されている。中国を含む 12 カ国には、90 以上もの GAC 教育センターがある。アメリカ、カナダ、イギリス、アイルランド、オーストラリア、シンガポール、ニュージーランド、メキシコにある 100 を超える Pathway University は、GAC を取得した生徒を受け入れる。

本稿では、教育はもとより職業選択支援の領域における ACT の製品、サービスの概要をいくつか取り上げる。

注釈: アメリカでは、単科大学も総合大学も独自の入試基準を設けている。大学へ進学希望の高校生は、志願プロセスの一部として通常、ACT のようなアセスメントをうける。一部の生徒には SAT を受ける、または ACT と SAT の両方を受けることを選ぶものもいる。単科大学、総合大学は通常追加に特定の入学試験は行わない。さらに、ACT と SAT は年間に複数回行われ、受験生はどちらかを、もしくは両方を複数回受けることができる。

ACT

ACT アセスメントプログラムは、高校生の高等教育への計画を支援し、そして、中等教育機関が生徒のニーズを充足することを支援する包括的なシステムである。ACT のテストは、英語、数学、読解、科学の多肢選択方式の到達度テストであり、その他、選択でライティングがある。ACT では、また、生徒の高校でのコース選択、成績、教育や職業への関心、課外活動や特別な教育のニーズなど、生徒自身が申告した情報を集めている。

収集された ACT のデータは、多くのユーザーのために様々な目的で使用される。高校では、ACT データをアカデミックな助言やカウンセリング、評価の研究、そして、認定基準に利用する。大学は、ACT の結果を入学者選抜やコース分けに利用する。州では、ACT を州全体のアセスメントの一部として利用する。奨学金や学生ローンなど財政援助を提供する多くの機関は、ACT を学生の資格認定の指標として利用している。

ACT は、ACT プログラムとしての機能と ACT の持つ Educational Planning and Assessment System (教育計画評価システム：EPAS) の中等学校レベルの一部としても機能する。

EPAS

The ACT Educational Planning and Assessment System (EPAS) とは評価とキャリア計画のプログラムを統合したものである。EXPLORE (8, 9 年生用)、PLAN (10 年生用)、ACT (11, 12 年生用) に分けられ、それぞれの生徒が大学や高等教育のためのアカデミックなレディネスを高める手助けとなるように設計されている。そのシステムは教育とキャリアの計画、アセスメント、教育支援や評価の長期的で体系的なアプローチを提供している。

これら 3 つのプログラムは英語、数学、読解、科学の分野において、カリキュラムベースのアセスメントを含んでおり、高等教育での成功と関連している。これらの内容では、高校生が学校で学んだこと、そして、高校を卒業する時点での大学進学や職業のレディネスとして必要となるものを測定する。さらに、これら 3 つのプログラムの得点は、テストバッテリーを通して同じ得点尺度で報告される。たとえば、ACT の得点は 1 から 36 のスケールで報告されるが、PLAN の最大スコアは 32 で、EXPLORE は 25 である。EPAS システムにより、ACT は 8 学年から大学までを通して生徒の成長をモニターする縦断的なデータシステムを確立した。その結果、生徒が K-12 (初等中等教育) で達成する到達度と準備の状況を測るレディネスのレベルは、高等教育での彼らの実際の成功と比較され、評価することができる。

The U.S. College Readiness Benchmarks

ACT のカレッジレディネス・ベンチマークは、高校生が ACT サブジェクト・テストにおいて、その約 75% が C グレード、もしくはそれ以上を獲得し、約 50% が B もしくはそれ以上を獲得する可能性があることが要求される得点である。そして、大学の初年次に一般的に履修される、英作文、代数、生物、社会科学 (たとえば歴史、哲学、社会学、政治学、経済学など) の単位付与コース (credit-bearing course) において、成功するために必要な得点である。EXPLORE と PLAN のベンチマークは ACT のベンチマークにおいて水準を満たすための生徒の成長の指標として確立された。

カレッジレディネスベンチマークは、全国各地の 9 万人の学生の 1998 年の 2, 4 年制の高等教育機関

でのデータを使用して、大学初年次の単位付与（credit bearing）コースでの実際の学生のパフォーマンスに基づいて経験的に導き出された。このベンチマーク得点を表1に示す。

表1 ACT カレッジレディネスベンチマーク

TEST	EXPLORE	PLAN	ACT
English	13	15	18
Mathematics	17	19	22
Reading	15	17	21
Science	20	21	24

2010年の米国での高校卒業生の47%、つまり約160万人が高校時代にACTを取得した。そのうちの24%がACTカレッジレディネス・ベンチマークの4つを取得しているか、それ以上を取得しており、これは2006年の21%、2009年の23%を上回っている。カレッジのコースワークに成功する用意ができていない高校卒業生の割合は、66%と英語が一番高く、次に読解の52%、数学の43%、化学の29%と続き、そこには大学とキャリアのレディネスの改善の余地があることを示している。全ての生徒において、高校卒業時までには獲得されるべき大学とキャリアのためのレディネスの重要な問題は、高校在学中にアカデミックなアチーブメントがどれくらい伸びたかであり、そして、彼らが高校を卒業する時点で、多くの生徒が大学やキャリアのレディネスを獲得していることを保証するために、その成長を促すことができるかどうかである。

Growth

EXPLORE、PLANそしてACTを受けたおよそ15万人の生徒のサンプルが、こうした成長問題の観察のために使用された。それぞれの科目における3つの評価の平均点は図1に示した。それぞれの科目は同じスケールで報告されるので、8年生から12年生までの間の平均点の伸びは容易にみることができる。その平均点の伸びは科学では3.3であり読解では5.6となっている。

生徒の総サンプルは、大学とキャリアのための準備が獲得されているかどうかによって、3つに分けられた。1つめは、EXPLOREへのカレッジレディネスベンチマークに到達しているか、それを超えているもの。2つめは、各ベンチマークまで2スコアポイントかそれ以下で到達するもの。3つめはベンチマークまで到達していないもの、に分けられた。数学における、これら3つのグループの平均点は図2に示した。

図1 8～12年生の平均到達度

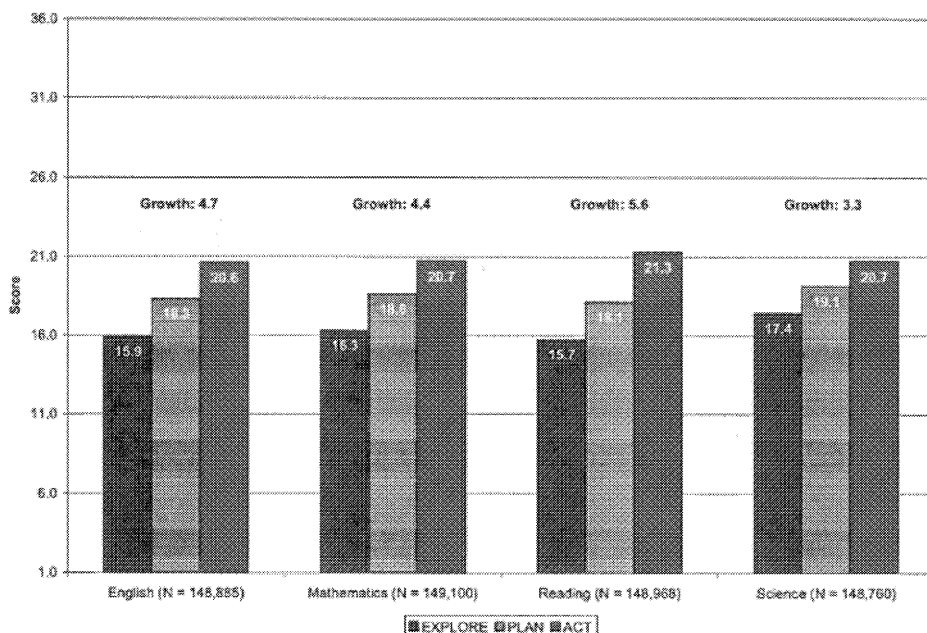
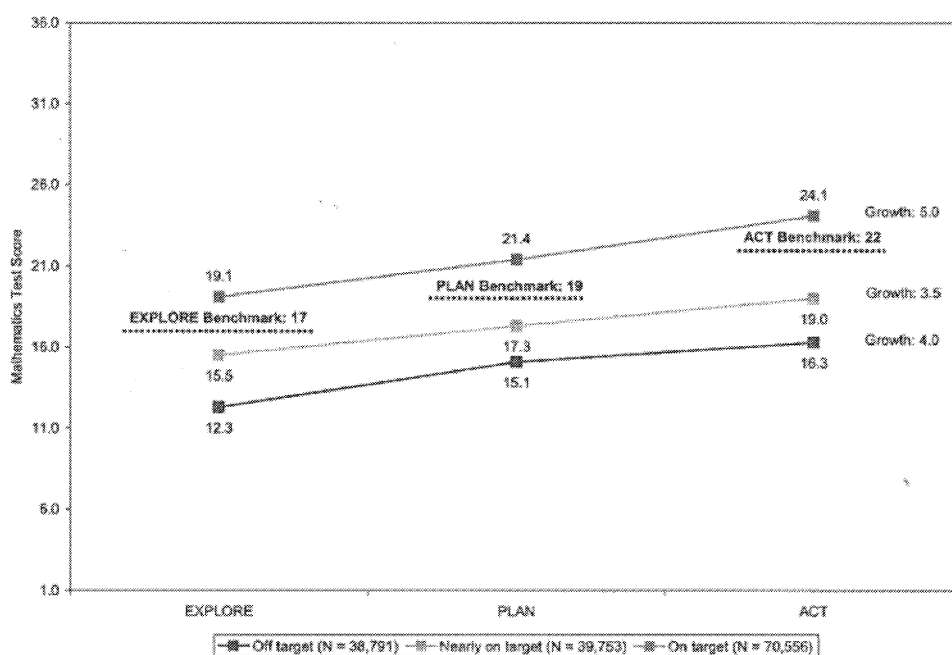


図2 カレッジレディネスの程度による8～12年生の平均到達度



大学とキャリアのための準備をしている8年生の平均的な成長が最も大きい。個々の生徒の成長目標は、尺度としてのカレッジレディネス・ベンチマークや成長曲線を使って設定される。8年生でオフターゲットの生徒に関しては、連続するテストの挑戦的ではあるが、合理的な目標は、生徒に与えられた科目の得点と対応するカレッジレディネス・ベンチマークの間の差の半分によって縮小すると思われる。成長の目標を設定するその他のアプローチとしては、高パフォーマンスを示している高校（最も大きな

成長を示している学校)の平均的な伸び率を測定し、次に、普通の、もしくは低パフォーマンスの高校の生徒のために、高パフォーマンスの高校の成長水準に沿って、目標を設定することである。

8年生で著しく大学とキャリアのレディネスが獲得できていない生徒は、高校時代に大学やキャリアの準備ができないということである。そのため、彼らが大学やキャリアのレディネスに必要な基礎的なスキルをつけることを支援するために、アカデミックな介入が必要である。

ACTはまた、3~7年生のアセスメントとリンクさせることによって、EXPLOREの前段階のレディネス目標を設ける試みをしている。これは、EXPLOREにおけるカレッジレディネス・ベンチマークに7学年のアセスメントと比較できるスコアを測定することによって行われる。そして、バックマッピングは6学年のスコアによって3学年を設定するために使用される。表2は、College and Career Readiness (CCR) Rampを示す。さらに、表3はNCEA成長目標と呼ばれる年度ごとの目標を示し、生徒がCCR rampに到達するためのパスを定めている。

表2 カレッジとキャリアレディネス指標の例

Backwards-Mapping the College and Career Readiness Targets

ACT/NCEA continues this process down to the lowest test grade—typically Grade 3. The trajectory defined by College and Career Readiness Targets from Grades 3-7 and ACT’s College Readiness Benchmarks from Grades 8-12 creates what is known as the College and Career Readiness Ramp.

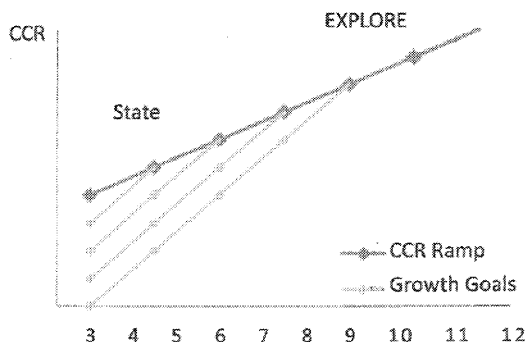
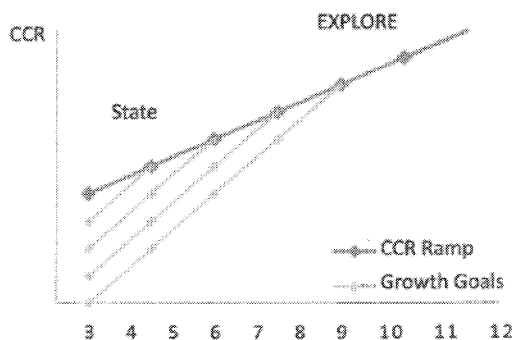


表3 年度ごとのゴールの例

Establishing Students’ Yearly Growth Goals

ACT/NCEA then identifies yearly growth goals that a student must achieve in order to get themselves onto the College and Career Readiness Ramp. These yearly goals are known as NCEA’s Growth Goals, and they define a path for a student to reach the College and Career Readiness Ramp in no more than four years.



Common Core

The common core state standards Initiative は、アメリカの教育への重要な改善を示しており、そして、ほとんどの州や地域において、全ての生徒のカレッジレディネスやキャリアレディネスには知識やスキルが必要不可欠であるという大半のコンセンサスに至る。ACTは、高等教育や職業トレーニングで成功するために必要な知識やスキルを識別する長期的な研究と、ACTのカレッジレディネス・スタンダードが、The common core state standardsの作成に使用される根拠となることについて、The

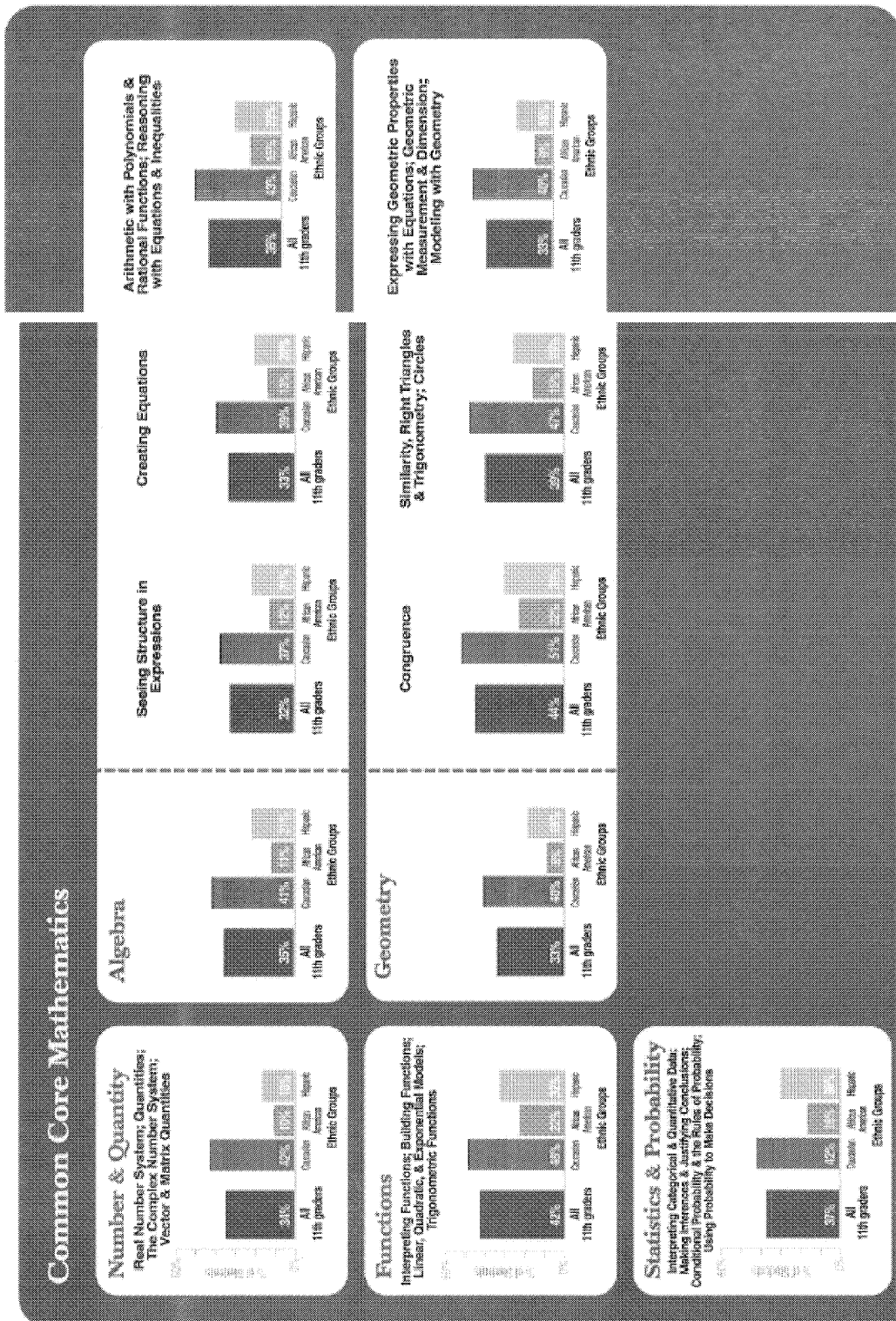
common core state standards のデザインに重要な役割を担っている。

高校生が、共通科目での成果を観察するために、ACT では、25 万人の代表的な ACT を受けた 11 学年の高校生のサンプルを利用し、共通科目クラスタでの項目のコード化を行った。パフォーマンス・スタンダードが、Common Core State Standards のために、未だ確立されていないので、ACT はその研究に基づいたカレッジレディネス・ベンチマークを、大学とキャリアにおける準備のパフォーマンス・レベルを推定するのに使用した。Common Core State Standards の各クラスタのために、ACT はデータ (Speaking & Listening と Research 以外) を利用して、カレッジへの進学準備ができていない状態のパフォーマンス・レベルを満たす、もしくは、上回るサンプルである生徒の割合が計算された。これらの分析は、州の実現努力に先立って、共通科目に関連する到達度を評価するための出発点として用いられる。図 3 は、全ての共通科目の領域、スタンダード、クラスタを越えて、高校生の 3 分の 1 から 2 分の 1 が、期待される到達度、すなわち、大学とキャリアのレディネス・レベルに達していることを示し、選ばれたサブ・グループの割合だけでなく、数学における ACT のカレッジレディネス・ベンチマークを満たしたサンプルの生徒の全体的な割合を示している。

共通科目の採択、およびその実施の期間は、全てのレベルで教育政策と活動を評価し、再構成するための重要な機会を提供する。

ACT は、ステイクホルダーが、共通科目をよく支援する政策とプログラムを始動するために、そして、生徒のカレッジ／キャリア・レディネスの現状を理解するために利用できる情報を、これらの分析結果から提供すると信じている。

図3 共通コア数学の最初の観察結果



International Benchmarking

ACT では、アメリカと海外の生徒のパフォーマンスとの比較を改善する目的で、読解と数学において OECD の参加国のパフォーマンスと ACT カレッジレディネス・ベンチマークとの関連を図るために PLAN と PISA を使用して研究を実施した。異なる集団や異なる管理条件のもとで運営されたが、PLAN と PISA には多くの類似点がある。両方とも、読解、数学、科学の内容の領域を提供している。そして、PLAN と PISA は、両方とも学生が学校で習ったことをもとに何ができるかということの評価しており、どちらも義務教育後の生活にとって重要である高次の批判的思考法のスキルを測定することに焦点を当てている。より詳細な比較については以下のホームページを参照：

<http://www.act.org/research/policymakers/pdf/AffirmingtheGoal.pdf>

アメリカは、グローバル化した経済の要求に適応するという挑戦に取り組んでいる。仕事はより専門化し、科学技術によってより追い詰められ、高度な教育とトレーニングが、特に数学と科学で求められている。アメリカの労働力はかつてないほど国際的に競争に直面している。また他の多くの国々の学生と比較されたとき、学業成績の国際比較で、アメリカの生徒が劣勢にいることを示している。現在のスタンダードが、国際競争のために必要なレベルに対して、生徒の到達度と準備の結果が十分かつ適正かどうかを判断するために、アメリカのカレッジレディネス・パフォーマンスは、国際的なパフォーマンスとの関係から検証される必要がある。単位付与 (credit-bearing) コース、および大学 1 年次のコースへのエントリーのための学力のアメリカのスタンダードが、グローバルに競争的なパフォーマンス・スタンダードであるかどうかを知るのには、一つはカレッジレディネスのように国際的な生徒のパフォーマンスを適切なパフォーマンス・スタンダードと比較する時である。アメリカにおけるカレッジレディネスのパフォーマンス・スタンダードと他国の生徒のパフォーマンスを比較することによって、国際的競争力のための生徒の準備はどれくらいの改革が必要なのかを知ることができ、単位付与 (credit-bearing) コースでの実際の生徒のパフォーマンスと国際的なパフォーマンスを繋げることができる。

ACT は、この研究を行う上で、独特な位置付けとなっている。それは、ACT カレッジレディネス・ベンチマークを使うことによって、カレッジレディネスのアメリカのスタンダードがどのように OECD の生徒の平均的なパフォーマンスに匹敵するかを確認することができる。ACT カレッジレディネス・ベンチマークは、実際の代表的な 2, 4 年制の高等教育機関でのサンプル学生のパフォーマンスに経験的に基づいているので、カレッジレディネスのためのアメリカの客観的なパフォーマンス・スタンダードとして用いることができる。PISA の結果を、PLAN テスト経由でベンチマーク・レベルとリンクさせることで、カレッジレディネスのアメリカのスタンダードと OECD の生徒の平均的なパフォーマンスは、同レベルか、上か下かを確認することができる。

2009 年秋には、アメリカで、PISA のナショナルアセスメントを受けて、生徒のサンプルが選ばれた。まず学校サイズにより抽出し、そして、同学年から無作為に抽出する二段抽出デザインを用いて、サンプルが選ばれた。対象の生徒にはそれぞれ 4 ヶ月の期間内に 1 冊の PLAN バッテリーおよび 1 冊の PISA ブックレットを提供した。ACT は、PLAN テストを標準的な手続きに則り実施した。そして、これらの生徒は the US national PISA administration に含まれていなかったが、テストはナショナルプログラムと同様の方法で行われた。ACT は、PISA のアイテムをコード化し、スコア化することに責任を負い、ACER (the Australian Council for Educational Research) は、PISA の推算値 (plausible values) の生成に責任を負った。

研究の分析目標は、PISA と PLAN、カレッジレディネス・ベンチマークをリンクさせることだった。研究サンプルに対して PISA の PLAN の得点分布が与えられ、それぞれの PLAN、カレッジレディネス・ベンチマークは、従来の非平滑化等パーセンタイル法を使用して対応し、PISA とリンクされた。

それぞれの科目に 5 セットの推算値があったので、それらの異なるセットを用いて 5 回のリンクが行われた。最終的に報告されるリンケージは 5 値の平均である。PISA テクニカル・マニュアルで定める推算値を用いた分析方法に関するガイドラインに従って linking variances も推定された。さらに、一般公開されている PISA データと PLAN データを使用した additional linkages に関して、交差妥当化研究が行われた。PISA の公開データは 2003 年、2006 年、2009 年のサイクルで、PLAN のデータは、2005 年秋の 10 年生の PLAN の標準分布 (norm distribution) からのものである。これらの付加的な資料からのリンケージの結果は、最新の研究と一致している。

表 4 は数学のリンク結果を示す。数学においては、PLAN カレッジレディネス・ベンチマークと同等の PISA スコアは 530.1 (5 つの推算値によって導かれた 5 つのリンケージの平均) で、それは PISA の数学リテラシーにおいてはレベル 3 となる。3 つの全ての validation linkages もレベル 3 である。study linkage と validation linkage 間のリンク誤差推定は、全ての 95% 信頼区間が互いにオーバーラップと言う点で非常に接近している (全てのリンケージに対する 95% 信頼区間は PISA のレベル 3 の上の範囲に入る)。

PISA レベル 3 と数学のベンチマークスコアでスコア 19 に関連した ACT カレッジレディネススタンダードについての言語的な説明を表 5 に示す。

表 4 数学における PLAN カレッジレディネス・ベンチマークと PISA の等化

Linkages	Plausible Value					Average	PISA Level	%95 CI*
	1	2	3	4	5			
Study Sample	530.7	529.1	531.9	529.1	529.7	530.1	3	[523.5, 536.7]
Validations:								
PISA2003	533.0	532.5	535.3	533.3	533.3	533.5	3	[526.9, 540.1]
PISA2006	522.9	520.2	520.2	520.6	520.6	520.9	3	[513.1, 528.7]
PISA2009	534.9	536.7	535.3	535.3	534.2	535.3	3	[526.2, 544.4]

* %95 Confidence Interval = Estimated concordant score \pm 1.96*(linking error)

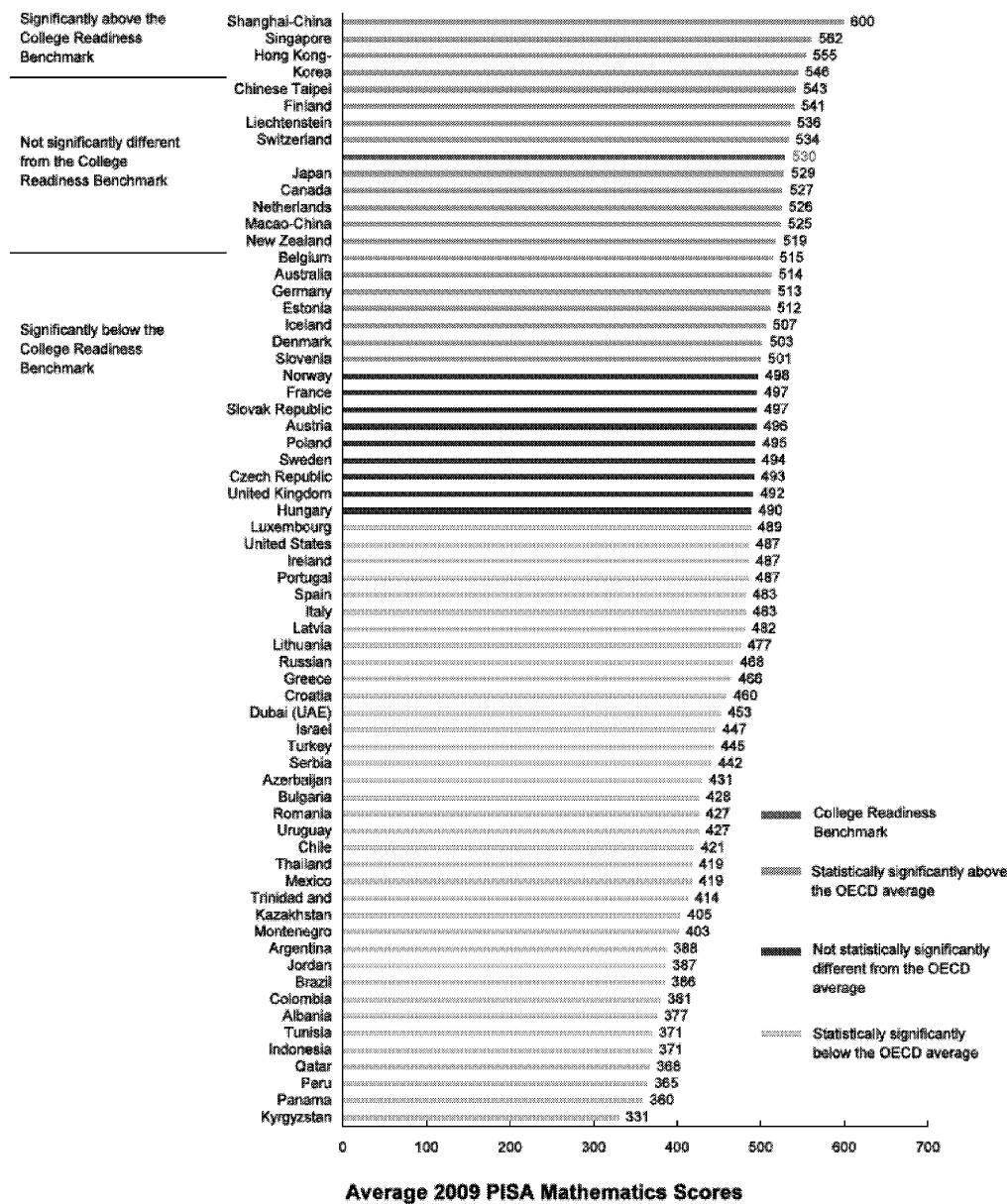
関心のある一つの問題は、ベンチマーク・バリューのレベルで得点しているならば、PISA のスケールで、アメリカがどこにランクインするのかということである。数学では、PLAN の数学ベンチマークに相当する PISA スコアの 530 は OECD 34 カ国の中で 3 位であり、最近の PISA アセスメントを行った 65 カ国の中では 9 位となる。カレッジレディネス・ベンチマークは国際的に競争力が高いレベルと一致することが示される。図 4 に示されるように、このレベルに到達している生徒は、国際的なマーケットで成功するだろう。

表5 アメリカのカレッジレディネススタンダードと対応する PISA Proficiency Level

	PISA ¹	U.S. College Readiness Standards
Mathematics	<p>PISA Level 3: At level 3, students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.</p>	<p>Score Range 16-19:</p> <ul style="list-style-type: none"> ▪ Solve routine one-step arithmetic problems (using whole numbers, fractions, and decimals) such as single-step percent ▪ Solve some routine two-step arithmetic problems ▪ Calculate the average of a list of numbers ▪ Calculate the average, given the number of data values and the sum of the data values ▪ Read tables and graphs ▪ Perform computations on data from tables and graphs ▪ Use the relationship between the probability of an event and the probability of its complement ▪ Recognize one-digit factors of a number ▪ Identify a digit's place value ▪ Substitute whole numbers for unknown quantities to evaluate expressions ▪ Solve one-step equations having integer or decimal answers ▪ Combine like terms (e.g., $2x + 5x$) ▪ Locate points on the number line and in the first quadrant ▪ Exhibit some knowledge of the angles associated with parallel lines ▪ Compute the perimeter of polygons when all side lengths are given ▪ Compute the area of rectangles when whole number dimensions are given

¹ Adopted from *Highlights from PISA 2009: Performance of U.S. 15-year-old students in Reading, Mathematics, and Science literacy in an international context*, by Fleischman, Hopstock, Pelczar, Shelley, and Xie (2010).

図4 10 学年における大学・キャリアレディネス・ベンチマークと PISA2009 Math の国別パフォーマンスの比較



Work Readiness

中等教育が終了したからといって全ての生徒が大学に行くとは限らない。幾人かは直接に就職する者、または職業訓練プログラムに進む者もいる。労働界の挑戦の一つは、適正な仕事に適任な人を配置することである。その構成要素は、人には、特定のポジションのために必要なスキルがあることを保証している。

ACT の WorkKeys は、雇用者が、高いパフォーマンスの労働力を選び、リクルートし、訓練し、そして維持するのを助けるようデザインされた評価システムである。ACT の WorkKeys システムはある程度においてどの場所でも利用されている。システムの一つの構成要素は、どのスキルが様々な仕事で

必要か、分析することである。ACT には、18,000 を超える職名を備えた職業のプロフィール・データベースがあり、それは、専門的なホワイトカラーから技術的なブルーカラーのポジションにまでわたる。この両方において欠くことのできないスキルとスキルレベルを識別するために、これらの仕事に関して広範な研究がなされ、そして、それは、従業員の選抜と訓練のために必要である。

WorkKeys アセスメントは、応用数学、情報の検索、情報の読み取り、応用技術、ビジネスライティング、理解のためのリスニング、チームワーク、職場での観察、そして適合、パフォーマンス、才能といった3つのソフトスキル・アセスメントを含む。

応用数学のアセスメントは、33 項目のテストで、英語かスペイン語で行う紙ベースもしくはコンピュータ・テストである。応用数学は、数学的推論、批判的思考、および問題解決の技術を職業関連の問題に適応する場合に、使用するスキルを測定することを目指している。それぞれの項目では、職場で実際に生じる問題を解き、そして、計算を行うことを受験者に要求する。応用数学の難易度は5段階のレベルで設定され、それはあまり複雑ではないレベル3から最も複雑なレベル7に渡る。レベルは直前のレベルで評価されたスキルを取り入れ、互いに関連している。レベル5の項目を解くためにはレベル3, 4, そして、レベル5のスキルを必要とするので、レベル5およびそれより低いレベルに関連したスキルの使用を必要とする数ステップの論理と計算が要求される。たとえば、ある問題では、注文をとり、注文書を作成し、そして、税金を計算することが必要かもしれない。(個々のレベルに関連する特定のスキルの説明は

<http://act.org/workkeys/assess/math/levels.html> 参照)

ACT が使用する Job Profiling process (職業のプロファイリング過程) は、4ステップである。ステップ1は、当該の職業に最も関連するタスクを網羅し、最初のタスクリストを作成することである。ステップ2はタスクの分析であり、ステップ1からの最初のタスクリストは、専門家によって必要に応じて調査され、修正される。そして、それぞれのタスクは、特定のタスクを伴う仕事で費やされる相対的な時間と重要性によって評価される。このデータは、それぞれのタスクの限界率 (criticality rating) を生じるのに使用され、その後、それは調査される。ステップ2の結果は、どのタスクが仕事を担うのに重要かについて示す、最終的なタスクのリストである。ステップ3は、スキルの分析であり、そこでスキルが調べられる。WorkKeys アセスメントにおいてカバーされるスキルの詳細な説明は、専門家によって提示される。彼らは、どのスキルレベルがエントリーレベルに必要なかを決定し、それらのスキルに関連のある仕事との関連を決める。ステップ4は、結果を記録することである。それは、スキルレベルと WorkKeys アセスメントと仕事のタスクの間の関連を確立するレポートである。

スキルレベルの一例は、<http://act.org/workkeys/skillsearch.html> から得られ、職名サンプルにとって必要であり、表6に示される。たとえば、職名が「会計士」の場合、応用数学ではレベル6、情報の検索と情報の読み取りではレベル5が要求される。一方、「広告会社の営業」という職名では、応用数学がレベル3であり、「会計士」に要求したことよりも低いレベルが必要なことを示す。

必要とされる WorkKeys レベルにおいて達していない生徒にとっては、WorkKeys システムはトレーニングの構成要素を含んでいる。特に、KeyTrain は、WorkKeys によって定義されるような応用の職場スキルをマスターすることの支援のために書かれた、目標カリキュラムに基づいたキャリアレディネス・スキルのための双方向ツールである。

表 6 職種によって必要な平均的 Work Keys スコアの例

Job Title	Applied Mathematics	Locating Information	Reading for Information
Accountants	6	5	5
Adjustment Clerks	4	4	4
Administrative Services Managers	4	4	4
Advertising Sales Agents	3	4	4
Agricultural Crop Farm Managers	5	5	4
Aircraft Body and Bonded Structure Repairers	5	5	5
Aircraft Structure Assemblers, Precision	4	4	5

Noncognitive assessments

ACT や他のものは、アカデミックと職業の両方において、認知的達成に加えて成功に影響を及ぼす興味や動機づけのような他の要因を示唆する研究を実施した。さらに、これらの非認知的な要因は、どのタイプのコースワークやキャリアかの適切性を理解する際に役立つかもしれない。あるいは、トレーニング・プログラムが特定の個人のために、うまくいくかもしれない。

ACT Interest Inventory

個人がキャリアのオプションを調べる際に、彼らのスキルや興味がさまざまな職業とどう結びついているか決める際には、しばしば支援を必要とするので、ACT は Interest Inventory を提供する。これは、個人使用として彼らが獲得する必要があるスキルや興味を持つようなキャリアを確認することに加えて、統治組織のような企業主体や自治体が、潜在的な従業員を現在や将来の仕事のニーズとマッチさせる方法としても必要とされる。

ACT の Interest Inventory は、EXPLORE、PLAN および ACT の構成要素であり、毎年 400 万人以上の個人にキャリアプランニングの情報を提供している。インベントリーの問題は、仕事に関連した活動（会議の運営、科学博物館の調査など）を強調する。それは受験者にとって潜在的になじみの薄い職業的なタイトルではなく、馴染み深いものでなければならない。ACT の Interest Inventory は、信頼性と妥当性を保証するために広範な研究に基づいている。個人の興味関心と環境（大学の専攻または職業）の間の適合することを示すこのインベントリーを使用する研究では、大学の専攻、大学の学位の取得、仕事の所得、そして仕事の満足度での永続性を含むポジティブな結果のさまざまなセットに関連がある。個人の Interest Inventory の結果は、テキストで伝えるように World-of-Work のマップ上で視覚的に報告された結果とともに 1 つの「正しい」職業を選び出すのではなく、個人の興味に合わせた職業のグループを示すことによって、キャリアの探索に集中させることを支援できる。

ENGAGE

ACTはその研究において、高校卒業までに今日の学校で起こることよりも、8年生のアカデミック・アチーブメントが彼らのカレッジレディネスに大きな影響を及ぼすことを発見した。生徒のレディネスはまた、社会心理的発達からも影響を受けている。成功したアカデミックなパフォーマンスの原因となると知られている態度を開発するならば、大学やキャリアのための生徒のアカデミックレディネスは改善することができる。ENGAGE アセスメントはアカデミックな成功と関連しており、アカデミックな態度の測定を提供している。そして、自己調整、社会参画、動機づけなど幅広い領域のもとで、学問分野、楽観主義、家族との関わりなどを含んでいる。

ACTは何千もの生徒にENGAGEをテストし、中学校から高校までの進歩をトラッキングした。そして、中学校で)行われたENGAGEが、高校での成績の有効な予測力となることがわかった。実際、前の成績やアカデミックレディネス(たとえばEXPLORE得点など)を考慮に入れても、ENGAGEはより正確に学業不振や低い成績のリスクを持つ生徒の特定を支援する追加情報を提供した。図5は、ある程度の併用で使用される中学校のアカデミック・アチーブメント(EXPLORE)と中学校の成績が、明らかに初期の高校GPAにおける優れた予測であることを示す。しかし、アカデミックな態度はさらに予測を十分なものにする。

図5 初期の高校GPAを予測する相対力

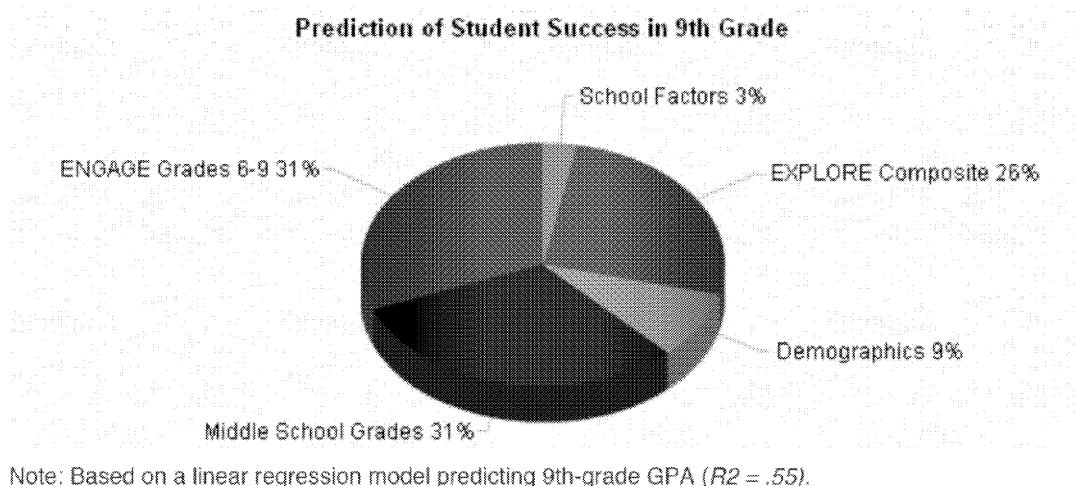


表7では、生徒のアチーブメントにおけるアカデミックなレディネスとアカデミックな態度の両方の重要性を例示する。生徒Aは、ENGAGEのアカデミックサクセス・インデックスおよびEXPLOREの両方でスコアが低く、6つのクラスを落とし、GPAが非常に低い。生徒Bは同じように、EXPLOREではスコア9と低いが、ENGAGEのアカデミックサクセス・インデックスにおいては高く、落としたクラスはなく、GPAが3.0である。生徒Cは、EXPLOREでは高スコアをとったが、ENGAGEではスコアが低く、一クラス落とし、GPAは1.56である。生徒Dは、EXPLOREでは21と高スコアで、ENGAGEでも高スコアなうえ、どのクラスも落とさずにGPAが一番高い。

表7 アカデミックな態度およびそのレディネススコアと後のアカデミックアウトカムにおける例

		Explore	
		Low (Composite = 9)	High (Composite = 21)
ENGAGE Grades 6-9 Academic Success Index (percentile rank)	Low	Student A Success Index = 3 High school GPA = 0.41 Failed high school classes = 6	Student C Success Index = 5 High school GPA = 1.56 Failed high school classes = 1
	High	Student B Success Index = 95 High school GPA = 2.99 Failed high school classes = 0	Student D Success Index = 99 High school GPA = 4.16 Failed high school classes = 0

ENGAGE グレード 10–12 と ENGAGE カレッジも同じく利用できる。加えて、ACT は ENGAGE Teacher Edition と呼ばれる一連のスケールを開発しており、アカデミックな成功に関連した特定の態度のアセスメントを提供している。これらの態度は、具体的で、観察が可能で、そして、アカデミック・パフォーマンスとほかの生徒の成功の結果に関係がある。そのスケールと ENGAGE が併用して使用される場合、結果として生じるデータは教師がリスクのある生徒を早期に識別し、個々の生徒の長所や短所を診断し、適切なカリキュラム活動と態度の介入を特定し、それらの活動と介入の効果を評価し、生徒の効果的なアカデミックな態度を開発するにあたって、進歩を追跡する。

Summary

本稿とそれに関連する話の目的は、ACT が教育と職業の領域で提供するさまざまな製品やサービスの一部について、情報を提供することである。ACT のウェブサイトでは、基礎研究と同様に、ACT が開発したアセスメントおよびサービスについてのより多くの情報を提供している。

New Approaches to Educational Testing and ACT

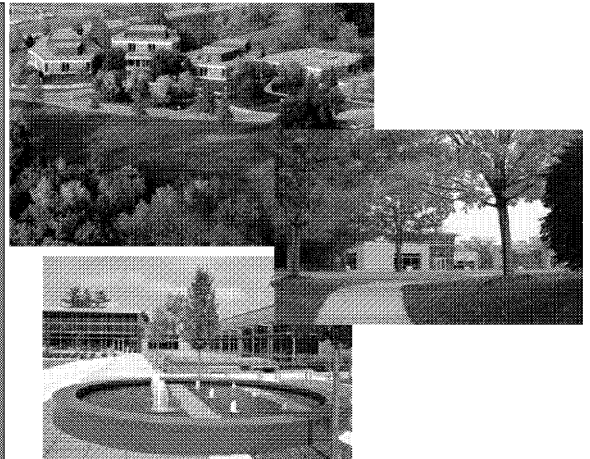
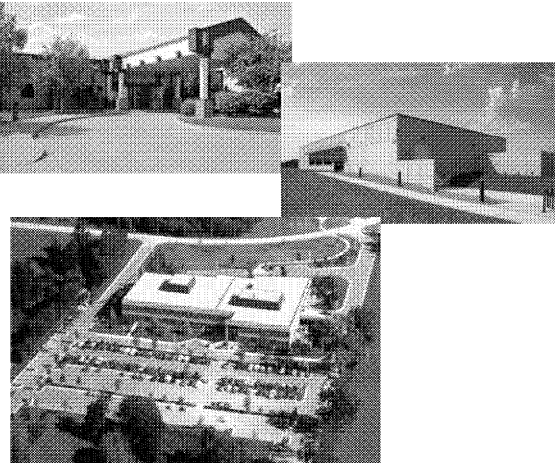
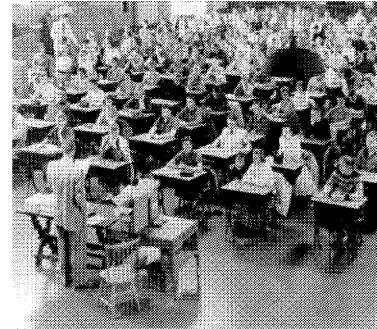
Deborah Harris
ACT, Inc.

Note: In the U.S., colleges and universities set their own admissions criteria. High school students who wish to attend college usually take an assessment, such as the ACT, as part of their application process. However, some students may choose to take the SAT, or both the ACT and the SAT. Individual colleges and universities usually do not have additional institution-specific entrance exams. In addition, the ACT and the SAT are given multiple times throughout a year, and an individual student may take one or both exams multiple times.

ACT's mission: Helping people achieve education and workplace success.



An early administration of the ACT test in the Old Armory Building on the University of Iowa campus



United States Offices

Regional Headquarters
300 ACT Drive
P.O. Box 198
Irvine, CA 92714-0198
Telephone: 949.851.4200

North Valley Office
Suite 100, P.O. Box 100
17150 N. Rockwood Road, Suite 100
Fort Valley, Georgia 31030-1000
Telephone: 478.244.4830

Northwest Office
One-Church Drive, 4th Fl.
1400, Suite 302
Seattle, WA 98101-3170
Telephone: 206.223.3333

Regional Director for Educational Achievement
6754 N. 34th Ave., Eastview, Suite 100
Scottsdale, AZ 85251-3804
Telephone: 480.350.1800

Florida Offices

Tallahassee 2000 S. Thomas Boulevard Suite 214 Tallahassee, Florida 32304-0214 Telephone: 904.487.9270	Orlando Office 3031 South Orange Ave. Suite 215 Orlando, Florida 32818-0207 Telephone: 321.257.2273	Fort Lauderdale Florida Office 3031 N. Maple Road Suite 300 Fort Lauderdale, Florida 33309-3000 Telephone: 312.259.3000
---	--	---

Chicago Office
300 Michigan Plaza
Suite 100
Chicago, Illinois 60601-4099
Telephone: 312.567.4500

Los Angeles Office
1601 S. Flower Street
Suite 400
Los Angeles, California 90019-0400
Telephone: 310.720.0200

Columbus Office
225 North Grand
Suite 410
Columbus, Ohio 43261-2019
Telephone: 614.470.1600

Indianapolis Office
144 Turpin Road
Suite 300
Indianapolis, Indiana 46203-1030
Telephone: 317.227.0200

Atlanta Office
1235 Lenox Road NE
Suite 100
Atlanta, Georgia 30309-1030
Telephone: 404.251.1900

Washington Office
1010 East Lafayette Street
Suite 300
Washington, DC 20003-4107
Telephone: 202.778-0700



International Offices



ACT Submissions Offices

Mexico Office
Plaza Manuel Gómez Morán 2
Edificio Alfredo Malpica Rivera 20
México, Spain 28020
Telephone: +52 55 51 517 6643

ACT Educational Institutions

Singapore Office
Room 11
401, 61 Heng Loong Road
#06-07 Bukit Timah Crescent
Singapore 650076
Telephone: 65-678 6940

Shanghai Office
Room 1204, Tian An Center
No. 258 Nan Jing West Road
Shanghai 200020
PR. C.H.R.A.
Telephone: 86 21 6259 1100

Jakarta Office
Marketing Executive Office
Mayapada Tower 17th Floor
Jl. Jenderal Sudirman No. 68
Jakarta 13000
Indonesia
Telephone: +62 21 5268 7343



The ACT Assessment

- The ACT Assessment program is a comprehensive system designed to help high school students develop postsecondary educational plans and to help postsecondary educational institutions meet the needs of their students.
- The ACT includes four multiple choice tests of educational achievement—English, Mathematics, Reading, and Science—and an optional Writing Test.
- The ACT also collects self-reported information about students' high school courses and grades, educational and career aspirations, extracurricular activities, and special educational needs.



ACT uses

- ACT data are used for many purposes, by many users. High schools use ACT data in academic advising and counseling, evaluation studies, and accreditation documentation. Colleges and universities use ACT results for admissions and course placement. States may use the ACT as part of their statewide assessment. Many of the agencies that provide scholarships, loans, and other types of financial assistance to students use the ACT as a measure of student qualifications.
- The ACT functions both as a stand-alone program and as part of the secondary school level of ACT's Educational Planning and Assessment System (EPAS.)



What is EPAS?

The ACT Educational Planning and Assessment System (EPAS) is an integrated series of assessment and career planning programs – EXPLORE (grades 8 and 9), PLAN (grade 10), and the ACT (grade 11 and 12) – that is designed to help students increase their academic readiness for college and post secondary training. The system provides a longitudinal, systematic approach to educational and career planning, assessment, instructional support, and evaluation.



ACT's College Readiness Benchmarks

ACT's College Readiness Benchmarks are the minimum ACT test scores required for students to have a high probability of success in entry-level, credit-bearing college courses, namely English composition, social sciences courses, college algebra, or college biology.

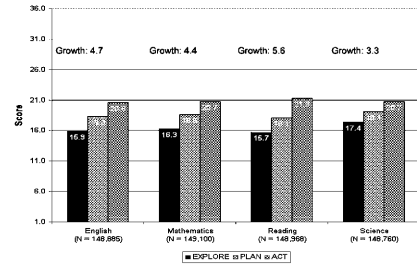


ACT's College Readiness Benchmarks

Test	EXPLORE	PLAN	ACT
English	13	15	18
Mathematics	17	19	22
Reading	15	17	21
Science	20	21	24

What is Reasonable Growth in High School?

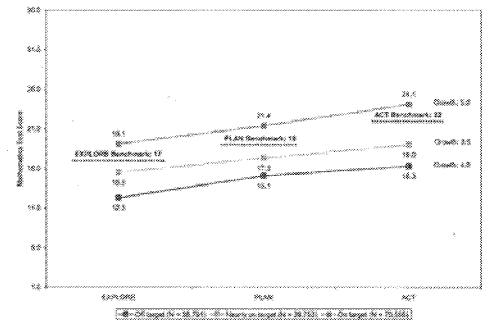
Figure 1: Average Growth in Achievement between Eighth and Twelfth Grades



Using the Benchmarks, the sample was divided into three groups:

- those who were on target in eighth grade (i.e., who met or exceeded the EXPLORE College Readiness Benchmarks),
- those who were nearly on target (i.e., who were within 2 or fewer score points of meeting each EXPLORE Benchmark),
- and those who were off target (i.e., who were more than 2 score points from meeting each EXPLORE Benchmark).

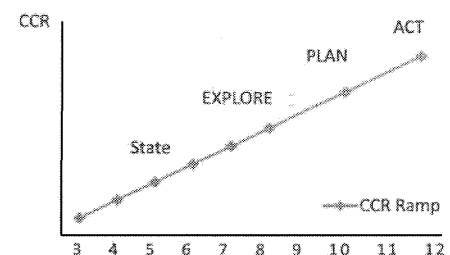
Figure 2: Average Growth in Achievement between Eighth and Twelfth Grades, by Degree of College Readiness



- Individual students' growth goals can be set using the Benchmarks and the growth trajectories as a yardstick.
- For students who are off target in eighth grade, a challenging yet reasonable goal on successive tests is to reduce by half the difference between the student's score in a given subject and the corresponding Benchmark.

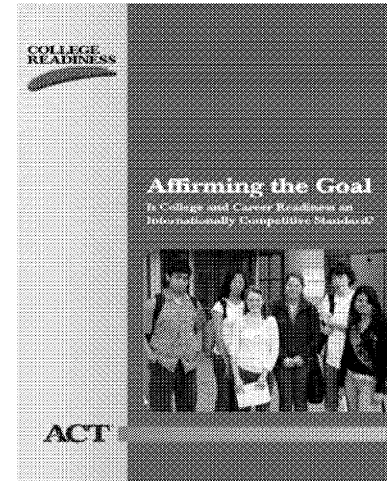
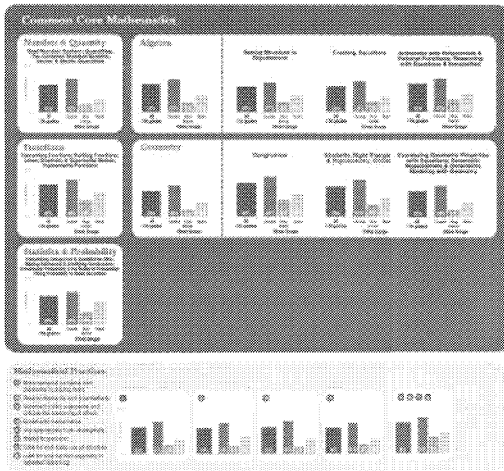
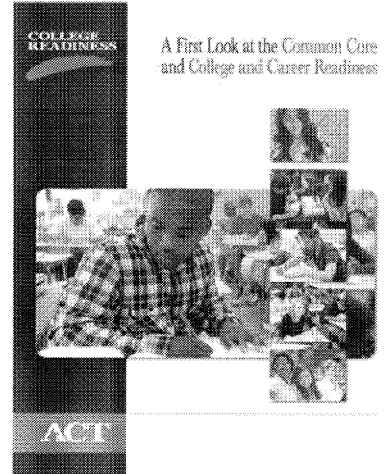
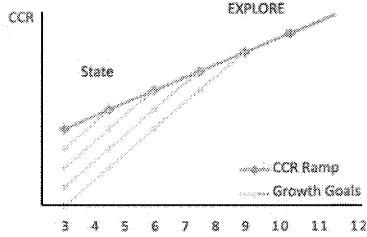
Backwards-Mapping the College and Career Readiness Targets

ACT/NCEA continues this process down to the lowest test grade—typically Grade 3. The trajectory defined by College and Career Readiness Targets from Grades 3-7 and ACT's College Readiness Benchmarks from Grades 8-12 creates what is known as the College and Career Readiness Ramp.



Establishing Students' Yearly Growth Goals

ACT/NCEA then identifies yearly growth goals that a student must achieve in order to get themselves onto the College and Career Readiness Ramp. These yearly goals are known as NCEA's Growth Goals, and they define a path for a student to reach the College and Career Readiness Ramp in no more than four years.



PLAN College Readiness Benchmark Equivalent on PISA for Mathematics

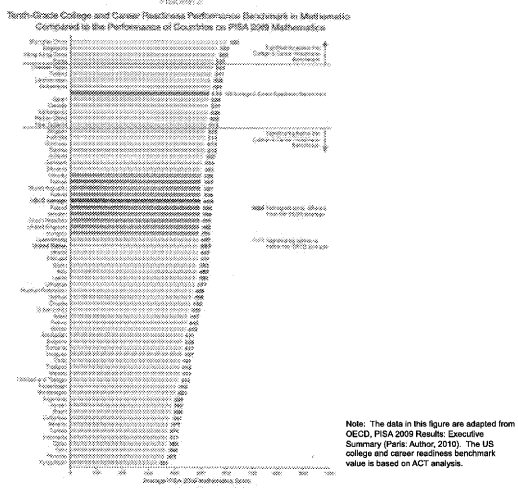
Linkages	Plausible Value					Average	PISA Level	%95 CI [*]
	1	2	3	4	5			
Study Sample	530.7	529.1	531.9	529.1	529.7	530.1	3	[523.5, 536.7]
Validations:								
PISA2003	533.0	532.5	535.3	533.3	533.3	533.5	3	[526.9, 540.1]
PISA2006	522.9	520.2	520.2	520.6	520.6	520.9	3	[513.1, 528.7]
PISA2009	534.9	536.7	535.3	535.3	534.2	535.3	3	[526.2, 544.4]

^{*} %95 Confidence Interval = Estimated concordant score \pm 1.96^{*} (linking error)

Descriptors of U.S. College Readiness Standards and Corresponding PISA Proficiency Level

	PISA ¹	U.S. College Readiness Standards
Mathematics	<p>PISA Level 3:</p> <p>At level 3, students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.</p>	<p>Score Range 18-19:</p> <ul style="list-style-type: none"> Solve routine one-step arithmetic problems (using whole numbers, fractions, and decimals) such as single-step percent Solve some routine two-step arithmetic problems Calculate the average of a list of numbers Calculate the average, given the number of data values and the sum of the data values Read tables and graphs Perform computations on data from tables and graphs Use the relationship between the probability of an event and the probability of its complement Recognize one-digit factors of a number Identify a digit's place value Substitute whole numbers for unknown quantities to evaluate expressions Solve one-step equations having integer or decimal answers Combine like terms (e.g., $2x + 5x$) Locate points on the number line and in the first quadrant Exhibit some knowledge of the angles associated with parallel lines Compute the perimeter of polygons when all side lengths are given Compute the area of rectangles when whole number dimensions are given

¹ Adopted from Highlights from PISA 2009: Performance of U.S. 15-year-old students in Reading, Mathematics, and Science Literacy in an international context, by Fleischman, Hopstock, Palczar, Shelley, and Xie (2010).



Note: The data in this figure are adapted from OECD, PISA 2009 Results: Executive Summary (Paris, Author, 2010). The US college and career readiness benchmark value is based on ACT analysis.

The screenshot shows the ACT Workforce Development website. The main heading is "REINVENT YOUR WORKFORCE". Below this, there is a section titled "Reinvent Your Workforce" with a sub-heading "In the 1980s, ACT sponsored college readiness in America. Now our Workforce Program continues to help students prepare for the workforce." There are several bullet points and a "Learn More" button.

WorkKeys Applied Mathematics Assessment

The screenshot shows the WorkKeys Applied Mathematics Assessment page. It includes a "Related Information" section with links to "Applied Mathematics", "WorkKeys", and "WorkKeys Applied Mathematics". There is also a "Sample Item" section with a math problem: "A car starts at 0 miles per hour and accelerates to 60 miles per hour in 10 seconds. What is the average speed of the car during this time?"

Example of Average WorkKeys Scores Needed by Job Title

Search Jobs by Skills

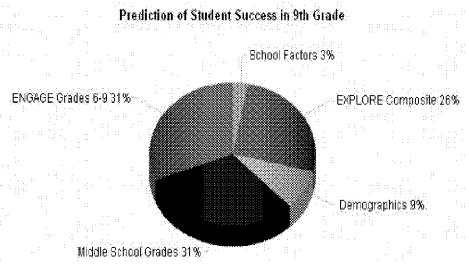
Average WorkKeys Scores Needed by Job Title

Browse Job Titles: A B C D E F G H I J K L M N O P Q R S T U V W

Job Title	Applied Mathematics	Locating Information	Working for Success
Accountants	5	5	5
Adjustment Clerks	4	4	4
Administrative Services Managers	4	3	4
Advertising Sales Agents	3	4	4
Agricultural Crop, Farm, and Forestry Managers	5	5	4
Account and Executive Clerks	5	5	5

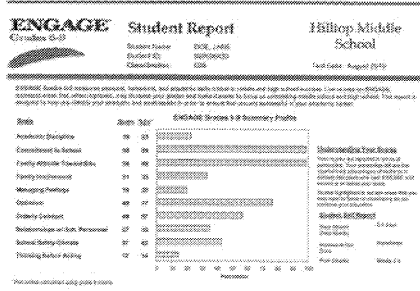
The screenshot shows the "World-of-Work Map" website. It features a circular map of the United States with various career clusters highlighted. A text box on the map lists "People & Data Person" characteristics: "People who like: helping, driving, caring for, selling things to other people, facts, computing numbers, and creating and organizing files." The website also includes navigation links like "About the Map", "Career Clusters and Areas", and "Understanding Research".

Relative Strength of Predictors of Early High School GPA



Note: Based on a linear regression model predicting 9th-grade GPA ($R^2 = .55$).

Sample ENGAGE Report



Sample ENGAGE Report

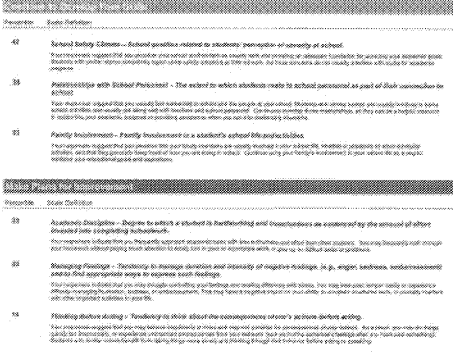


Table 1. Four Example Students' Academic Behavior and Academic Readiness Scores and Later Academic Outcomes.

		EXPLORE	
		Low (composite = 9)	High (Composite = 21)
Low	Student A	Success Index = 3 High school GPA = 0.41 Failed high school classes = 6	Success Index = 5 High school GPA = 1.56 Failed high school classes = 1
	Student B	Success Index = 95 High school GPA = 2.99 Failed high school classes = 0	Success Index = 99 High school GPA = 4.16 Failed high school classes = 0

Summary

The intent of this presentation was to provide information about some of the wide range of products and services ACT, Inc. provides in both the education and workplace arenas. The ACT website offers much more information on the assessments and services ACT has developed, as well as the research that provides the foundation they were built on.

Thank you

New Approaches to Assessment in the 21st Century

Eva L. Baker

University of California, Los Angeles

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Overview

This paper addresses how the predictable and unpredictable changes associated with globalization, media, demography, and new developments in learning together make the time exactly correct to consider how to change our approaches to testing and assessment. The paper will discuss a few aspects of greatest salience affecting our contexts, then move to an explicit discussion of 21st century skills and how they respond to emerging contexts. A set of steps is offered around the design of new measures, all with the purpose of finding an evidence-based and reasoned strategy to meet the future in the testing area. The excellent work conducted over the years by National Center for University Entrance Examinations in Japan, under the leadership of President Yoshimoto will serve as a frame for the quality of our expectations.

Why is the season ripe to rethink assessment? Globalization is no longer a general and abstract notion. It has been operationalized clearly and is at the heart of the rise and fall of national, regional, and international economies. Globalization involves—in a minimalist way—the current and emerging expectations for expertise in the labor force, including their compensation expectations, the length of time of their working life, levels of support in retirement, levels of preparation in high school and the university, and the roles and redistributions of universities and graduate schools. At one level globalization forces continuous benchmarking and may put developed nations, for a time at a disadvantage relative to price and productivity.

Globalization is also fueled by technology. The massive change in the last five years of the use of mobile and other media for work, social engagement, private amusement, instruction, and learning has had effects in many ways. First, there are effects on the learners, students, and adults themselves, in terms of sense of expertise, preferences for visual or graphic stimuli, reduction in reading or other former media pursuits, television, and film. There is also an ethos of personalization that the new media creates, an expectation that I can choose for myself what I am interested in, and more person to person rather than person to institution relationships. Given that formal testing, whether part of the conduct of instruction and the determination of effectiveness is on the rise, and selection tests have long held a strong position in the purposes of testing, both come from an institutional rather than personal perspective. Thus they conflict with the message of the present day media. While tests have flourished in an environment where they can predict performance, media and technology use the currency of rapid adaptation and change. Therefore, not only are new arrangements for assessment important, but we must

re think our present course and focus on both what and how assessment should take place.

Ideally, the goals for assessment should incorporate three components: (1) the strength and resilience of cultural values that should be maintained, despite the onslaught of globalization; (2) the disciplinary and practical knowledge incorporated in academic knowledge; and (3) personal learning goals of the student, or trainee if that is the case, in order to preserve motivation and relevance of learning. So if we agree that the learners' expectations for personalized learning increases, we will need to consider how those goals can be incorporated into measures.

Much of assessment consequences have been directed to the learner or student directly, for example, in admissions and placement in institutions. More recently, assessments have begun to play a role in the heretofore private realms of classroom instructors and professors, providing more standardized approaches to diagnosis, feedback, and certification of effectiveness. But the pendulum may be swinging back, toward a more individual set of performances that will allow highly adaptive fits with desired institutions.

This context leads us to a consideration of assessment purposes as they are arrayed now, in order to place in perspective the discussion of 21st century skills at the core of this discussion. In Table 1 is a list of common purposes for assessment categorized by whether the purpose is principal directed to individuals or institutions.

Table 1
Traditional Purposes and Uses of Assessment

STUDENTS	INSTITUTIONS
Admissions	Status
Placement	Comparisons
Communication	Improvement
Motivation	Personnel decisions
Diagnosis	Sanctions and rewards
Feedback	Public and policy estimates of quality
Improvement	
Certification	

While purposes are a critical element of assessment (and linked as we will see to validity evidence), a very common way of classifying assessments is in terms of their format, as commonly experienced in schools. These include problem sets given to master procedures, multiple choice items to sample understanding of content, projects, essays, research papers and other student-constructed, more extended responses. Sometimes this contrast is between objective and subjective tests (although constructed responses can be as rigorously objective as any measure). They may also be contrasted in terms of the surface features presented to students, for instance, paper based, manipulatives, or administered through computer or other technological means? The computational basis of assessment, however, involves more than the superficial form of administration, sometimes called electronic page

turning. New developments in technology permit assessments to adapt to the level of student's performance during a test. They may also offer either high fidelity representations of complex situations, in social, scientific, or mathematical environments, they may be embedded in a "game-like" setting focused on the acquisition of specific accomplishments or qualifications. Such computer assessments can be scored in an automated, instant way, and in fact can be composed of elements or modules and assembled rapidly. In the near future, such assessments will be created as needed in real time.

Yet purposes and surfaces features are attributes of assessments that glide over their real core.

At the heart of any assessment is its content: what skills, content, strategies are behavioral manifestations of the thought processes affected by learning. The major focus of all assessment of an educational nature is the understanding of the degree of learning that has occurred, as determined by a sample of student performance. Our major focus now must shift to the learning that is desired in this century.

The term "twenty-first century skills" is a widely adopted metaphor for a class of cognitive skills and social and affective competencies that are thought to be essential for future education and training. They are pertinent whether the goal is to develop the individual capacity, to ready the student for further or higher education, for instance, technical training or college, or to consider the types of skills thought to be of high value in the current and future work environments, including the military. In many discussions of cognitive skills the emphasis is often on the selection of individuals with generally measured aptitudes, such as intelligence or creativity. The problem is different from a training and education perspective. With the focus on how to change skills and proclivities, many argue that thinking skills should be developed within the context of explicated content domains, such as mathematics, science, and history. They may also augment the typical definitions of skill areas, such as literacy in the national language or additional languages. A second, relatively recent idea in workplace or academic settings is that it is sufficient to teach a cognitive, such as problem solving for instance, in a specific domain, such as algebra. Others, including this author, believe otherwise, and propose a phased approach. First, attention must be directed to helping the student learn to apply the skill in the principal areas of content. Second, to demonstrate significant competence in the domain, performance must traverse the full breadth and sufficient depth of the content domain of interest. Third, the focus on instruction and assessment should be directed to building and verifying that the learner has acquired a set of principles incorporated into a mental schema or pattern. This step is important if trainees are expected to retrieve efficiently the key aspects of the cognitive demand rather than to memorize surface features of a procedure (Sweller, 2003). Finally, attention is not only required for a broad range of content, but more explicitly to the notion that the attributes of the situation, constraints, elements of content, and quality of solution or action may in fact change simultaneously but to different degrees. The ability to respond to such unpredictable new states is the fundamental difference in the use of the term "cognitive readiness" as opposed to cognitive demands or 21st century skills. The expectation is that "readiness" implies the ability to undertake an unforeseen assignment. To achieve such readiness, the trainee or learner

must be exposed to a sufficiently diverse set of conditions, situations, and problem settings so that their ability to transfer their learning to a new setting will be developed (relative to applying the schema described above). In education, to date, most assessments do not explicitly call out 21st century or less modern formulations of cognitive demands, such as adaptation, risk taking, or situation awareness. They instead over cue on the content knowledge itself, applied in routine settings. Even recent research (Baker, 1997a,b; 2007a,b), the focus has been on designing learning and assessment tasks in the context of tried and true domains. This is true even when content knowledge has been improved conceptual refinement or greater specificity (see State Standards Common Core 2010). The continued explosion in knowledge suggests that routine contexts or notions of fixed domains are obsolete. Learners will need adaptive skills or else must relearn all the contexts and content to which these skills may apply, clearly an impossible goal. If adaptability to an unforeseen future is important, and analysts predict new careers and tasks within five years that are largely unknown today, learners must be comfortable with applying principles and schema they have learned as well and to determine how such schema may be modified in order to meet new requirements. Therefore, in addition to schema and transfer performance, the learner needs to develop unforeseen new skill sets which may be combinations or modifications of methods taught to solve problems, reason, or make decisions. To summarize, at this point, in military training and in the world of work, the emphasis is on the application of some of the 21st century cognitive skills, concentrating on the variable context of emerging and uncertain situations. In educational research on academic learning, the rendition of “uncertainty” of future context has been often limited and focused on “transfer” situations, that is, tasks where the learner needs to apply the skill to a heretofore unlearned situation, domain, or explicit set of constraints. The difference between “academic” settings and military or workplace training may be in a figure-ground difference, where emphasis on content, skill level or context is a matter of perspective, but different perceptions have importance for the design of both learning and assessment systems. That is, whether one primarily sees skills embedded in content or whether one’s attention is on adapting to changing contexts will modify one’s approach to design of assessments, and of the learning experiences that precede them. This difference in perspective may be one useful marker for deciding whether one is in the 21st Century Skills basket.

A Consideration of 21st Century Skills

Earlier in my own thinking of assessment design, it became clear that we should be focusing on cognitive skills when assessments were designed. The earlier conception is in contrast to the focus first on content and then on formulas of test formats, such as multiple choice items. The earlier list is represented in Figure 1 below.

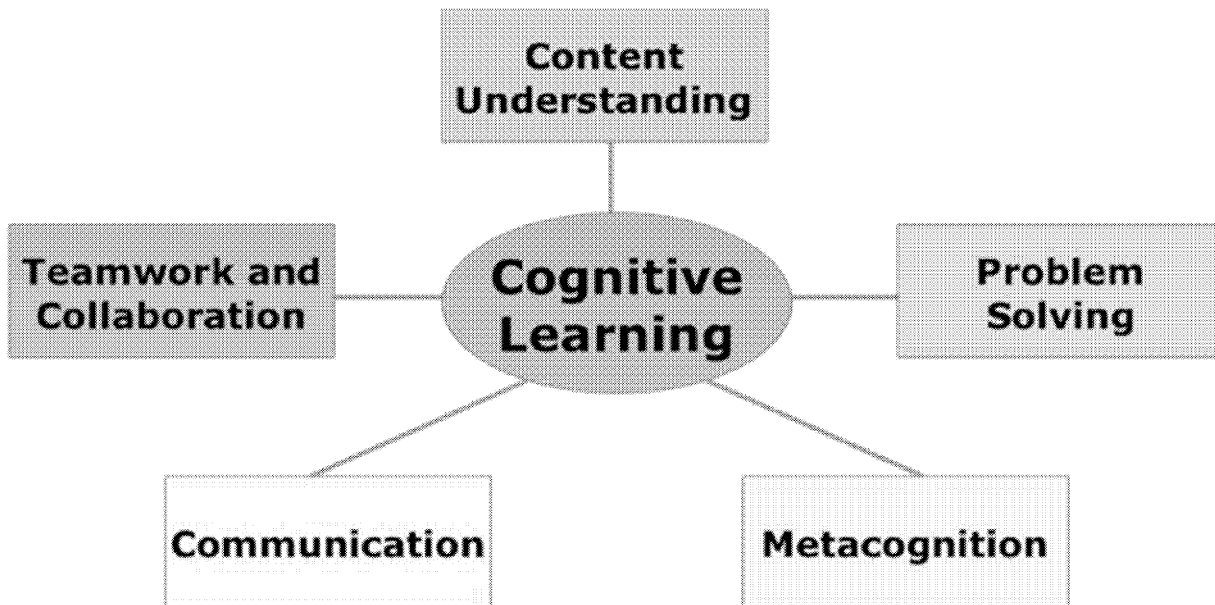


Figure 1: Simplified Cognitive Demands for Assessment Design

As the formulation of 21st century skills evolved to be more complex, as the context of use became less predictable, we must consider the sources of high performance. There are two main sources to weigh. The first is nascent talent, traits, in other words the individual differences that make some people more perceptive, more curious, more fluent, characteristics that students bring to school. The second source is the skills that are learned explicitly, whether in schools or less formal settings. Obvious, there are interactions among traits and learning. Curious people may make excellent problem solvers. For our discussion, we will put aside relatively stable traits, and focus on 21st century skills that are intended to be learned systematically in programs and institutions, and are to be measured to judge attainment for the purpose of certification or prediction.

What is the range skills included in 21st century skill sets? Any list of them is not purely original, for they overlap in language, using related synonyms to mean similar ideas. I have collapsed them into three major categories: (1) intellectual cognitive processes, (2) socially oriented processes, and (3) intrapersonal skills. The first and third focus on that can be enacted independently by the individual. These skills, may be either outward-looking, such as “how do I scan to identify key features of a problem?” or inward-directed, for instance, “how do I manage my own cognitive (or affective) processes to achieve my desired goals?” For the most part, all skills, especially those tasks in schools and training, are embedded in particular subject matter. A child may be asked to figure out how to cross a river, using given objects and applying fundamental principles of force and motion, including momentum and friction. A high school student might be asked to determine the optimum shape of a figure that meets given or inferred sets of constraints. A Naval officer might need to interpret signals and signs to determine whether to launch a defensive action. An internal medicine doctor might need to determine a diagnosis for muscle pain and suggest a course of action for the patient.

Intellectual Cognitive Processes

In the area of intellectual processes are those cognitive tasks that require sets of related sub-processes to achieve a particular goal. Consider problem solving (Baker & Mayer, 1999) which normally involves a different set of thinking skills under conditions where the problem is clearly defined as opposed to a situation where the problem is not well specified (Feltovich, Spiro, Coulson, & Feltovich, 1996). After the identification, and verification of the problem, the learner or examinee must represent the problem clearly, verbally, graphically, or symbolically and determine whether the task instructions demand a “correct” or convergent answer or if they permit a divergent response requiring a new combination of learned skills, or an innovative strategy. If convergent solutions are called for, then the learner must find the optimal path and match the designer’s general solution. Procedures to verify that the posited solution solves the problem may be related to another task. Some of these skills have a long and relatively unchanging set of definitions, for instance the difference between inductive reasoning (from examples) and deductive approaches (reasoning from a premise) Work in this area goes back to classical times, but a more recent publication by Johnson-Laird (2006) summarizes evidence in this area. A newer principle may be the cognitive skill of search, where students are encouraged to conduct searches, but to use a model of evidence to validate their findings. An earlier literature, mostly in the field of information science has given way to the ability to use search engines well, particular tips, for instance, offer at almost every university library, and varying sets of criteria to judge quality of output (see Dragan, Tepper, & Misso, 2011 for an example). The ability to apply criteria of quality will be increasingly needed as technology repositories may either be “crowd-sourced,” e.g., Wikipedia, or not subjected to any quality control. Another area of interest is the understanding of declarative, procedural, and systemic knowledge. This understanding is critical for as noted earlier, cognitive skills are embedded in subject matter and situations. Sources for the tripartite definitions of understanding (What, How, and Why correspondent to declarative, procedural and systemic knowledge) have a long history. Authors writing in this area including a series of articles by Alexander and Judy (1988) reviewing the literature in the area, and an excellent earlier piece by de Jong and Ferguson-Hessler (1996), on types and qualities of knowledge. These categories have been combined with Bloom’s *Taxonomy of Educational Objectives* (1956) in a revision by Anderson and other cognitive psychologists (2000).

Table 2, a list of potential 21st century skills is listed, focused on intellectual cognitive processes. When these are engaged effectively, “on-the fly” without much preparation, they merge into the realm of cognitive readiness.

Table 2

21st Century Skills – Intellectual Cognitive Processes

-
1. Adaptive problem solving
 2. Decision making
 3. Situation awareness
 4. Reasoning, inductive and deductive
 5. Searching for missing resources
 6. Understanding declarative, procedural and systemic knowledge
In particular sub domains
-

Two points are obvious. First each of the lists of intellectual cognitive tasks does not operate in isolation. A task nominally focused on one may depend upon the application of others. For example, problem solving may require situation awareness (Koenig, Lee, Iseli, & Wainess, 2009) and decision making (Lee, Bewley, Jones, Min, & Kang, 2009; Lee, Jones, & Min, 2009). Second, each of the skills may take many different surface forms. These skills play out differently in domains comprised of different principles, concepts, facts, and routine approaches. Going back to the problem solving task area above to illustrate, consider an ill-defined problem. A problem may be ill-defined in a number of ways. It may be vague without sufficient information to guide strategies for a solution. The respondent will then need to generate alternative, plausible interpretations of the problem, represent them in an appropriate fashion and then proceed to apply strategies or procedures for solutions. It is clear that there may be loops and returns to the beginning if the problem identified is not what was intended by the designer. A problem might also be ill-defined because part of it is hidden or occluded. That is, the learner may need to wade through extraneous material to find the real problem, stimuli that might be verbal, visual or both. The task writer may work very hard at deception, trying to lead the respondent to a wrong interpretation. To get through this type of ill-defined problem, the learner must know enough about the task situation and the content domain to avoid false lures. If in the problem set-up, the designer has included explicit or less obvious constraints, then the respondent must use skills of situation awareness to be able to focus down on the real problem. The learner will also need to use reasoning skills, even at the problem identification stage, if he or she is to discard less central material and determine which variables are critical determinants of the task. At key points of a complex, multi-stepped problem, the respondent will need to make decisions. These decisions will be made on the basis of prior experience, knowledge provided or obtained relevant to the task, and by the application of reasoning.

When analysts consider such skills as a “trait” or innate individual differences, they are positing that a good “problem-solver” logical thinker, or detector of change in environmental situations, will be able to apply these skills generally, without specific training in or across domains. In other words, that the cognitive skills are used independently of the particular content domain with about the same level of skill. It turns out, however, that individuals may have their abilities bounded by related domains, for instance, good problem solving in math will not bleed over to the same respondent’s

behavior in literature. What we are positing, however, that training that involves transfer of skills first within a broad domain, then in varying situations applicable to the domain and the intellectual skill, will need to be augmented by training that requires application across content domains. Whether such a domain-independent set of performance skills can be developed remains questionable. Each domain typically requires relatively deep declarative procedural and systemic knowledge before the respondent can use strategies to solve problems or make decisions. It is somewhat unlikely that ordinary rather than extraordinary minds will acquire deep knowledge over a wide range of domains. Thus, the plausibility of the domain-independent application of intellectual cognitive skills depends on the limits of time, interest, and capacity to learn domains that are far afield from one another.

Social and Interpersonal Skills

A second class of 21st century skills involves a set of socially oriented skills that include cognition, but require its application in interpersonal situations. These competencies may involve the collaborative nature of work on the one hand, or the ever present need to communicate to obtain approvals for plans, to discuss and understand work and to report it to a range of audiences. Moreover, socially oriented skills do not heavily depend upon sunny or out-going personalities. For instance, the area of collaboration requires the learner to be able to clarify goals of the team or collaboration, to modify behavior to acknowledge the value of ideas, even in situations where they may differ in opinion with those in the group.

In Table 3, we provide a partial list of such skills.

Table 3

Socially Oriented 21st Century Skills

-
1. Teamwork
 2. Collaboration
 3. Help
 4. Social situational awareness
 5. Communication—productive and receptive
-

To succeed at work using the intellectual tasks listed in Table 2, the learner may need to depend upon other people as a resource on the one hand, or provide only part of the effort needed for a solution that will involve many players. Writers about team work (Salas & Cannon-Bowers, 2001) have created a set of component skills that may involve providing leadership, feedback, motivation, redirection, clarification, incentives, and effort in order to succeed in the task. Not all tasks call for all of these components, nor should it be imagined that any single person is required to provide all of them for a given task. But from a learning perspective, to be a good team or group member, some of them will be required. For example, to be a leader requires from time to time different tasks for different requirements. Clarity of goals may be a constant requirement, but it may be wise to have a group clarify goals rather than have the goal specified always by a particular individual. Obviously, the

manner of application depends upon organizational hierarchy and individual status. To make the right choice, the team member must be aware of the various states of the group or team, including their willingness to participate, their role, and other information, such as their experience or desire to try new things. Research on the topic of the social situational awareness is often called empathy (see Keltner, 2004). Here it is differentiated because it is formulated as a task to be learned as opposed to an attribute that is inherent. It is clear that interactions in the social realm, whether face-to-face, continuously or occasionally, through media, or transmitted by text, depend upon some level of expertise in communication. The skills are required to formulate, compose and explain important tasks or to ask and answer key questions. Another aspect of communication is the sensitivity to the use of appropriate language, suitable to the audience, the organization, task, and to specific interpersonal contexts, including, where relevant, cultural issues. These then form an evolving set of related skills pertinent to the domain of social situational awareness, combining skills that are both intellectual and interpersonal. As interpersonal skills may have strong experiential, cultural, and personality bases, it is unclear that they can, in totality, be trained or taught. However, key components, for instance, communication has a long history of being learned and assessed. Some studies look at the relationship of receptive (listening) or reading and productive (speaking or writing) at very specific levels (see for instance Guess, 1969) or as a more specific set of interpersonal skills (Hargie, 2006). Vygotsky (1978) saw social interaction and communication at the heart of higher psychological processing. Elements of collaboration or teamwork, especially understanding and executing specific steps or roles related to identify tasks, also can be well taught. In working on teams, the interpersonal components will develop over time, over stress conditions, and in varied task situations. Tasks vary as those that can be completed independently by members of a group or where each team member makes a unique and interdependent contribution to the attainment of the goal (see Webb, 1985).

Intrapersonal Skills

Intrapersonal skills are those personal behaviors and internal thought processes that can be systematically acquired or enhanced by instruction or learning experiences and for which evidence of change can be directly or indirectly inferred. In Table 4, there is a partial list of such skill sets.

Table 4

Intrapersonal Skill Sets

1.	Planning
2.	Self-monitoring, on intellectual tasks, including feedback
3.	Emotional awareness and self-control
4.	Risk-taking
5.	Motivated effort and Attributions of success and failure

Whereas some of these skills may be clustered under categories of metacognition, for instance, planning, self-monitoring, and deliberate effort, others have a more emotional

component. For example, individuals can learn to control emotions, to practice a more balanced personal demeanor through mind-body regimens, or to estimate risk and change their propensity for risk-taking related to situations, e.g., costs, likely success, consequences of failure. (There is general evidence that each of these can be taught to some degree.) The social and emotional skills can be learned, and may involve attribution (Weiner, Graham, & Reyna, 1997), self-efficacy, and resilience. These are components of self-awareness. The process of emotional control, under stress, derives from personality psychology (see Roger & Neshoever, 1987 for example of measurement). There are many examples with long research histories, for example, desensitization research, that ranges from those documenting therapeutic approaches (Gordon & Berstein, 1973), effects of sexism (Linz, Donnerstein, & Penrod, 1988), or the effects of violent media and games (Cline, Roger & Courier, 1973).

Risk taking, moderated by a sense of payoff, is another intrapersonal skill that has more currency in the context of self-motivated actions, leadership, and entrepreneurship. This work has a lengthy history as well (see for example, Brockhaus, 1980) and has been investigated often in the context of management schools.

The interacting areas of effort and attribution have been well summarized by Graham (1991), where researchers have first shown a relationship between success and attributions to personal effort, and have developed training regimen to develop such concepts in those who think that they have been either selected to fail or have no control over their learning. This line of inquiry is singularly related to studies of stereotype threat and how to overcome such perceptions by African Americans. Claude Steele (Steele, Spencer, & Lynch, 1993) has continued to probe the role that self-perception plays to support resilience under socially threatening conditions.

There are other elements, such as creativity, critical thinking, reflectiveness, which may or may not be as amenable to instruction, depending upon how the construct is formulated, and the cultural context, age, and experience of the learner. In this 21st century skill discussion, however, these elements are imagined to be changeable through education, training, and experience.

Individual differences apply to each of these areas, and intensive training may enhance intrapersonal skills to a high level. On the other hand, some background differences, for instance, experience with failure, may make it harder for training to be effective.

If the elements exemplified in the intellectual, social and intrapersonal 21st century skills are meant to be used in educational and training situations, what is the difference between them and the elements of cognitive readiness described, by Fletcher (2004), O'Neil, Lang, and Perez (in press), and others? At some level, the cognitive readiness notion emphasizes the ability to be agile, adaptive and prepared for uncertainty. While many elements of 21st century skills also share that emphasis, it is fair to say the term "readiness" means the ability to act in unpredictable situations. For the most part, 21st century skills are vested in institutionally based learning, schools, university, or world of work. When the term "cognitive readiness" is used in the military context, it engenders an image of rapidly

evolving situations, unforeseen challenge and constraints, and the requirement of rapid rather than reflective or long-term analysis. In the cognitive readiness model, there are some writers (O'Neil, Lang, & Perez, in press) who classify cognitive readiness attributes on a continuum ranging from relatively easy to train to those that are difficult to train or best considered as individual difference traits.

Towards Building Assessments

To build assessments, we start with the array of 21st century skills, intellectual, social, or intrapersonal. The next step is to consider content to be included in the measurement.

This second step in implementing any of the three types of 21st century skills involves the use of an ontology, or map of the content domain(s) of interest. An ontology is a graphic representation of language with the following characteristics: a network of nodes, a set of links describing the relationship among nodes, and a data base which will modify the direction or arrangement of nodes and links based on performance information. If an ontology is a graphical representation of a verbal domain to be used for learning or assessment, what is its construed properties? An ontology features principles, concepts, key knowledge, and procedures. These are depicted as nodes in a network. The links among nodes have direction and meaning. They may convey, in a hierarchical representation, the components are subsets of others, and range from desired higher order complex content all the way down to fundamental principles and facts that beginners are expected to know. Mathematics is an excellent area to display hierarchical strands of topical domains. For different content, the ontology may also be structured in an appropriate, active way, where principles, concepts and facts are linked to display in a way to convey chronology, themes, or principles. Political history is an example of a domain that may have a structure based on chronology. Another different structural form of a content ontology may illustrate the relationships among nodes (containing principles, concepts, or examples) in terms of their mutual or directional influence on one another. The strongest case of such links might be those that exemplify causal relations. In most fields of study, one will find a mix of relationships, some reflecting part-whole relationships, others more thematic, chronological, or causal. In some areas, the structure may be loose where the broad domain is considered, for example, in literature, where relationships among forms, such as the novel or poetry may be parallel or horizontal rather than hierarchical. Yet, within a literary form, for instance, stronger vertical structure may be found. In cases of plot or character development, highly explicit relationships can be developed.

The structure of the ontologies must follow the essential character of the domain. These may vary by the extent to which interpretive processes are present as opposed to specific methodologies are intended to yield relatively clear outcomes. These domains will also differ in the light of the extent to which they represent abstract principles, addressed theoretically or empirically, for instance in Physics, as opposed to domains in which each example may only be loosely joined to the next, for instance, examples of lyric poetry. Figures 2, 3, and 4, depict ontologies that vary by content, granularity, and representation.

Map of Algebra I Big Ideas

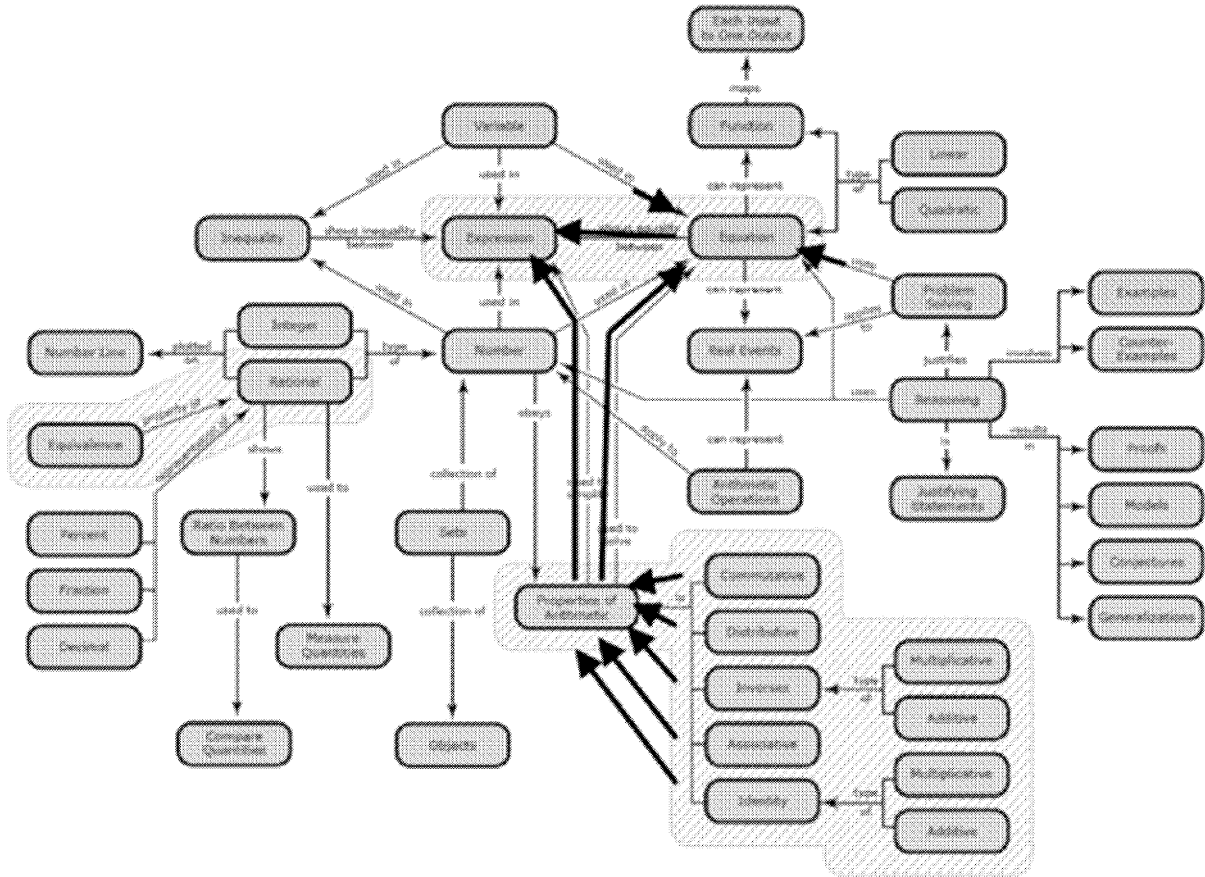


Figure 2: Content Ontology of Beginning Algebra

Map of Algebra I Big Ideas

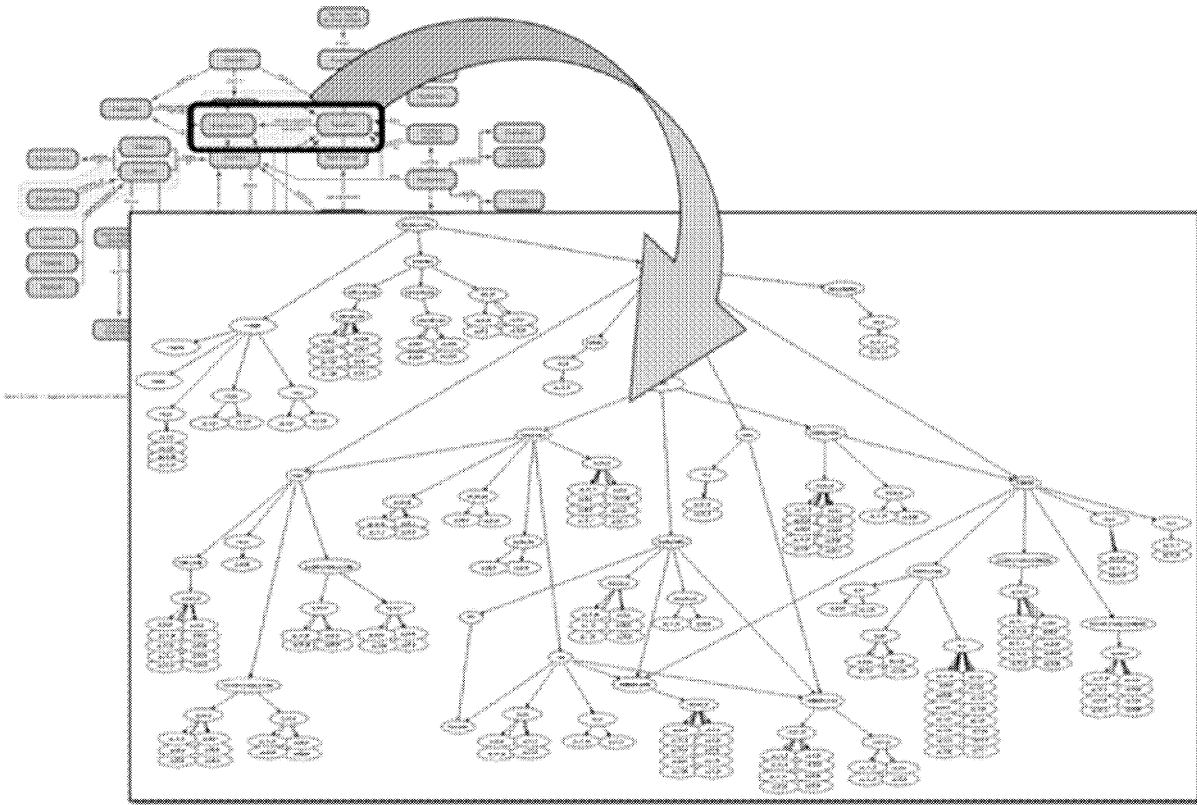


Figure 3: Algebra Ontology–Dynamic Bayesian Network

Map of Algebra I Big Ideas

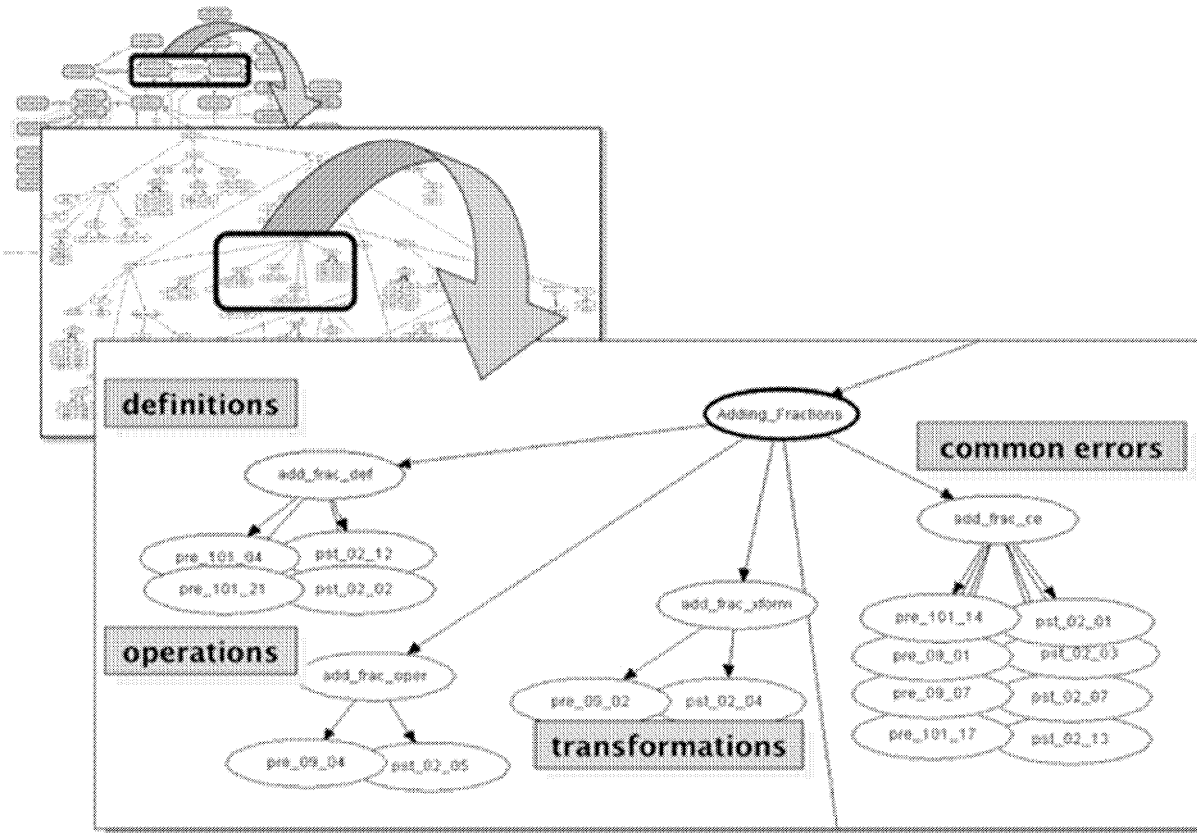


Figure 4: Algebra Ontology–Decomposing Data

Common Elements of Ontologies

Independent of structure, any ontology has particular features. First it is a graphical representation. Second, there is no restriction about how many nodes may be linked to any other. In point of fact, centrality of content can be determined simply by looking at the number of nodes with the highest frequency of links. One can also determine which content is remote, potentially non-essential because it is less well connected to more central ideas.

This operational depiction of importance can lead to direct inferences about the design of learning systems. While some may view an ontology to be a sequence of instruction, or the optimal arrangement of a computer adaptive test, no such inference should be directly made. For instance, it is fashionable to use the term “learning progression” as if one truly knew which sequence was optimal. However, only empirical study can generate stronger hypotheses about the order of learning or of assessment for learners and settings. The type of empirical study could be undertaken to contrast the processes followed by individuals known to be competent in the domain with those of novices, or with individuals only partially taught, or able to demonstrate middling levels of performance. Other empirical work can take hypothesized arrangements of tasks and delete systematically alternative elements in order to determine which content and processes are essential to the achievement of desired goals.

How Is an Ontology Made?

Ontologies are usually made iteratively and made by people. Imagine that an expert in a subject matter area were asked to represent the domain of interest in terms of its important topics and their relationship to other concepts, principles or facts. One way of thinking about an ontology is that we are asking the expert to externalize and to depict symbolically a mental model of the domain. In practical ontology development, experts are carefully selected, asked to create an ontology individually (using software, e.g. the Knowledge Mapper developed by Chung & Baker, 1997, at CRESST) where the expert can create the structure of nodes and links using a pull-down menu of options. When multiple experts each develop an ontology, they can be superimposed graphically to show areas of agreement and differences. A typical process then requires face-to-face discussions, explanations, and reconciliation of differences by the experts, iteratively, until consensus is reached. In addition, in the most recent research undertaken by Chung, Niemi, and Bewley (2003) and Iseli (2011) ontology development can begin or be augmented in process with the analysis of documents relevant to the domain. Automated extraction of key ideas in referent texts, articles, and other documents can be achieved using natural language processing placed in a network representation and included in the material that the experts are to reconcile. Content ontologies are in development for mathematics from kindergarten through secondary schools (Iseli, 2011; Iseli, Koenig, Lee, & Wainess, 2010), history (Phelan, Dai, Valderrama, & Herman, 2011), biological science (Phelan, Dai, Valderrama, & Herman, 2011), and language arts, (Phelan, Dai, Valderrama, & Herman, 2011).

Blending 21st Skills and Ontologies

In the ideal case, ontologies of 21st century skills should be merged with a content ontology to create an integrated architecture to guide learning and assessment. At CRESST, we have made some progress in creating ontologies for 21st century skills, in the areas of problem solving (S. Mayer, 2010), communication (Phelan, Dai, Valderrama, & Herman, 2011), situation awareness (Koenig, Lee, Iseli, & Wainess, 2009) and teamwork (O’Neil, Wang, Lee, Mulkey, & Baker, 2003). Each of these ontologies is based on theoretical and empirical analyses of the process domains. For instance, in skill of situation awareness, research by Endsley (1995) and her colleagues was essential, and in the teamwork area, frameworks, and empirical studies by Salas and Cannon-Bowers (2001), and O’Neil and colleagues (O’Neil, Chuang, & Chung, 2003; O’Neil, Chung, & Brown, 1997; O’Neil, Wang, Chung, & Herl, 2000). In S. Mayer’s problem-solving ontology, nationally recognized cognitive psychologists were used as experts and asked to create an ontology, one that is still under revision to refine its content and structure. It is our intention to continue to document the process by which intellectual and social skill ontologies are developed and combined with content domain ontology for the purposes of assessment, simulation, and game design (Chung, Delacruz, & Bewley, 2004; Chung, Niemi, & Bewley, 2003; Koenig, Lee, Iseli, and Wainess, 2009), marksmanship training (Chung, Delacruz, Dionne, & Bewley, 2003) and tactical decision-making (Bewley, Lee, Jones, & Cai, in press).

Unpredictable futures, however, may seem antithetical to an assessment approach that uses ontology as the way to represent “all possible” or known content. The unpredictability that cognitive readiness is specialized for inheres not in the details of the content, although one might posit that less frequent or more unusual content has a place in the definition of “unexpected.” The unpredictability may be more likely related to the unusual situation in which an individual is placed. There, the benefit of well-learned knowledge structures is in the rapid retrieval and testing relevancy against the situational requirements. There is not much evidence that this is the process that leads to achievement, survival, and avoidance of error, but there is a clear research path that could be taken to determine the components of knowledge needed to be agile in unpredicted environments.

Model Based Learning and Assessment of 21st Century Skills Embedded in Content and Cognitive Readiness in Unusual Situations

Design using 21st century skills, cognitive readiness and content ontologies must be realized in assessment and learning situations and systems. Let us focus on assessment. We have developed a model for the development of assessments and its history has evolved since 1992 (Baker, 2007c; Baker, Chung, & Delacruz, in press; Baker, Freeman, & Clayton, 1991). In the model of assessment we begin not with content specifications, which is the usual practice but with the selection of 21st century skills that will be embedded in the content domain of the ontology measure. Reasons for this explicit inclusion first focus on relevant 21st century skills can be explicated. First, it is done to assure that the intellectual depth of processing is included on the measure as intended by verbal statements on standards or doctrine. If not made explicit, many tests are found to over emphasize recognition or repetitive procedures simply because those are easy to generate. Second, the

operational definitions of 21st century enables the determination of which relevant content should be given higher weight in the intellectual skill domain. Third, particular intellectual skills suggest assessment formats to optimize caliber of measurement. For example, adaptive problem solving demands that the respondent create an original answer, whereas a problem identification measure might combine selected responses, i.e., which is the best statement or representation of the problem, with a verbal explanation about why the choice of problem statement was made. An explicit cognitive model of the 21st century skill allows the generation of one or more templates and or sets of modular objects that can be used in combination to build assessments. This modular feature looks forward to partially automated assessment design using computer-based authors systems. Next, the use of a common model task for assessment will increase the likelihood of coherent sampling within a domain. Construct-irrelevant tasks features (Messick, 1989) or item types that add noise to the understanding of student performance, will be identified and reduced. This focus on measuring the desired and relevant outcomes will affect positively reliability of findings and help detect real change as a function of intervention or experience. Moreover, the use of models will permit the development of subsequent extensions of the measures at a lower cost, because there will be three sources of guidance: the 21st century skills or cognitive demands, the ontology of relevant content, and the assessment task model. This economic utility should not be underestimated. The templates can be reused, and different situations, content, or response can be inserted, which will support longitudinal interpretations of growth. However, in the case of open-ended responses, scoring criteria or rubrics can also be reused and will greatly reduce cost. Furthermore, if rubrics are at a high level of connection to the ontology and 21st century skills, particularly if additional progress is made in computer-automated scoring (Chung & Baker, 2003), teachers and students can be recipients of more transparent requirements for learning. Moreover, the blended or integrated design of 21st century skills and content ontology has a more pervasive purpose than just developing a pool of tasks and criteria. The dual representation of skills and ontology can be the core of a database design that is intended to serve first as a repository for student performance on relevant assessments. To the extent the assessments are arrayed in a learning sequence, then the database structure may change as various patterns of correct and incorrect answers are accumulated. The skills and ontology serve as initial metatags for the database.

Model-Based Assessment Recap

When tasks are created using skills (either singly or in combination) and content, either separately or in a multidisciplinary way, there are still some issues that need to be addressed. In addition to the specification and realization of a coherent domain to be measured, assessment design can be facilitated by using models, templates and other components to deliver assessment tasks that are far more sophisticated than present item formats but achieved at a very low cost. Such an outcome can be achieved because these components, e.g., templates can be reused in subsequent tests, even in different subject matters, saving the need for reinvention for every new examination. To illustrate, one kind of template may involve a worked example (Sweller, 2003) which supports student retrieval of patterns or schema rather than small bits of knowledge. Another routine, but powerful approach is to require the respondent to explain the principle basis or “why” they selected

or constructed the answer they provided. This approach assures that examinees demonstrate that they have deeper understanding than may be inferred from the application of a straightforward procedure. Third, in the area of standards-based, or criterion-referenced assessment intended to measure a domain of expertise, procedures must be put in place first to assure that the marking of performance adheres to the skill and ontology demands in a reliable and accurate fashion. Second, a procedure needs to be developed to value the score obtained. In an admission setting, these procedures may involve ranking students or creating standard scores that translate to normal distributions. The alternative may be a set of comparisons between expert and novice performers.

Technology Challenges

As technology develops, it presents some challenges to current design processes as well as to accepted procedures for determining reliability and accuracy. In technology tasks, in simulations, for example, each assessment may be scenario-based, and a considerably longer task. Instead of having many items, the basis for most psychometric analyses, only a relatively long, interdependent task or two might be used in a technology setting. Research is needed to develop new ways to establish the quality of such tasks.

Validity Minimums

The approach to validity and other relevant instances of technical quality should be explicit. Validity is purpose driven. Validity is a chain of inferences linking the purpose of the assessment to data and subsequent inferences about the quality of decision the assessment yields (see Messick, 1989; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), with different purposes, types of data, and relevant inferences will vary. If the major use of data is formative assessment, for instance in order to help the designer to decide how a simulation or game should progress based on learners' prior performance, then the data of interest will be relatively granular, focused on misconceptions, hesitations, and errors that need to be addressed in a revision of the system or perhaps just relevant to the alternative paths provided for individual students. If the data were to be used to assess the acquisition of a set of skills for certification, or accountability, on the other hand, then the validity data would need to determine whether those skills were also represented in the repertoires of the skilled or performers or other individuals who had nominally achieved the desired goals. If the purpose were admissions testing, the assessment is used to make predictions, both the quality of the predictor and the predicted criterion measures must be subjected to empirical scrutiny. In addition, if public reporting for accountability were of interest, for instance, to categorize schools in terms of their own effectiveness, then information about learner, student and teacher, groups, mobility amount and types of instruction, and different levels of satisfactory achievement would be considered to infer validity. In all cases, however, data management systems would be needed to report outcomes to designers, teachers, or policymakers. They would need to be organized so that progress along skill dimensions could be tagged, as well as other potential variables such as content, situational, and linguistic complexity. This database function refers back to the larger use of ontologies, recalling that they serve not only a graphical representation role, but guide the design of

ongoing data management systems and reporting. In any case, a series of procedures can be used to establish quality. These procedures include (1) expert review of alignment to the target ontologies; (2) think-aloud protocols to determine whether expected learner processes are being applied in task examples; (3) identification of critical paths of actual performance; (4) psychometric studies of reliability, dimensionality, and fairness (total scores and diagnostic subscales); (5) usable by targets: teachers, students, administrators; and (6) validity evidence established before wide application of assessment.

Changes

Birthrates in Japan, South Korea, Taiwan, Hong Kong, Singapore, and most other parts of the developed world in Asia and Europe have dropped significantly below replacement levels. Aside from countries with strong religious inducements to have children or in poorer countries in Africa, numbers of students available to higher education may change. It is possible that families with fewer children may invest more in their intellectual capital and thus, the lower birthrate may not translate directly into fewer numbers of high-quality students for higher education. However, it is likely that except for the highest quality institutions, the quality of students applying for higher education may drop. Two predictions may be made about this view of the future. One is that stringent entrance examinations may only apply to a far smaller set of institutions (and a far smaller set of applicants) than in recent times. The second is that students will accept, in the name of cost containment or convenience, lower quality educations, for instance, consistent with some current mass market online institutions today and not seek admission to heretofore desirable institutions. As an aside, in the U.S., the birthrate is approximately two children per woman, and “replaces” the current population. Nonetheless, a similar phenomenon may be occurring as in countries with lower birthrates because of the historical difficulty in educating the majority of poor students to a high level of competence. Thus, unless there is a dramatic breakthrough in educational strategy, the numbers of highly qualified students applying to higher education may similarly drop.

In these cases, what are the options for systems that heretofore relied on entrance examinations to sort students. It seems as if, among a number of options, two appear to be probable and viable. One is to transform admissions tests offered on a national level to placement tests, that may be used to match students to institutions on the one hand, or to determine which entry level courses the examinee may qualify for. The second, related option is that instructors or professors in higher education who has benefited over the years from having well-prepared, motivated students, will now have to learn to conduct classes where they must adapt their teaching to reach less competent students. The risk in these scenarios is that the overall quality of the output of higher education could considerably drop. What may happen is that elite institutions of higher education may attract students from other countries, or in a broader context, countries with excellent reputations could well attract the top tier of students to their colleges and universities. A variety of adaptations would be required to deal with less homogeneous students, either intellectually or culturally.

Conclusions

So what may be the impact of uncertainty on admissions examinations? First we would assert that achievement is more than selection and purposes such as placement and certification may need to replace the current models. Second, that learning is a continuous process, not stopping with the success on a test. Twenty-first century skills will form an important content base for the future. These skills will have to be relearned in new contexts, new applications, new job requirements. It is very possible that that the assessment and examination process will be replaced and utilized in situations that are now formally beyond institutional settings.

Within the current context, examinations may match individuals to best available programs, identify and support undiscovered sources of talent that may meet varied institutional missions. The choices will be to determine and experiment with unconventional uses of assessment, to balance choice of students and flexibility versus control of educational institutions. The findings of examinations may also be analyzed and reported differently, using new techniques of data capture and modern analytics to pinpoint accomplishments and further needs.

Summary

What has been predictable may not be any more, because careers and content are exponentially changing in shorter and shorter times. Because of changing economic and work environments, the new focus must be on learning, in schools and throughout life. At this point, 21st century skills, content ontologies, and new methods of assessment design will be one strong pillar in support of essential learning of emerging, unpredictable requirements

This chapter has provided perspectives on 21st century skills and cognitive readiness, justifying our interest in them in the face of increasing uncertainty. The set was divided in three ways: (1) intellectual cognitive skills; (2) socially-oriented skills, and (3) intrapersonal skills. Comments related to selection or training of these skills was followed by an analysis of the term “cognitive readiness” and its intent in explicitly dealing with details of transfer and preparation to confront unexpected requirements. The role of content domains with respect to 21st century skills was treated, and a method for representing the details of content was developed. The graphical approach to representing content was defined and described in the concept of ontology or a symbolic representation of content, relationships and structure. Uses, experiences and research options were discussed with regard to ontologies as a method for identifying important skills for assessment sampling and for learning design. A brief acknowledgement of the need for ontologies for 21st century skills was described with topics of current work. The approach to creating assessments based on 21st century skills and content ontologies, “model-based assessment” (Baker, 1997b, 2007a) was described in terms of its utility in creating higher level tasks, increasing technical quality of measures and reducing cost. A brief discussion of validity included the key notions of minimizing construct-irrelevant variance and drawing correct inferences related to purpose. A discussion of an extension of the database development of ontologies was

presented. It is a more complex formulation of the relationship of individuals to explicit aspects of learning: skills, content, linguistics, assessment formats, experiences, and developmental trajectories. This strategy may have relevance as more automation of student experiences is systematically accomplished. All mechanics of design of learning and assessments ultimately hinge on the number and qualities of the students who present themselves for examination. In the face of changing birthrates, and new expectations in a globalized world, the future of examination systems is sure to change, but in change remain robust.

References

- Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Education Research, 58*(4), 375-404.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, L. W., & Krathwohl, D. R., (with Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C). (2000). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. Boston: Allyn and Bacon.
- Baker, E. L. (1997a, February). *Model-based assessments*. Paper presented at the 1997 AAAS Annual Meeting and Science Innovation Exposition (AMSIE'97), Seattle, WA.
- Baker, E. L. (1997b, Autumn). Model-based performance assessment. *Theory Into Practice, 36*(4), 247-254.
- Baker, E. L. (2007a). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment* (Special Issue), *12*(3&4), 179-194.
- Baker, E. L. (2007b, August/September). The end(s) of testing (2007 AERA Presidential Address). *Educational Researcher, 36*(6), 309-317. Retrieved October 2, 2007 from: http://www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3606/09edr07_309-317.pdf
- Baker, E. L. (2007c). Teacher use of formal assessment in the classroom. In W. D. Hawley with D. L. Rollie (Eds.), *The keys to effective schools: Educational reform as continuous improvement* (2nd ed., pp. 67-84). Thousand Oaks, CA: Corwin.
- Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (in press). The best and future uses of assessment in games. In M. Mayrath, D. Robinson, & J. Clarke-Midura (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 227-246). Charlotte, NC: Information Age Publishing Inc.
- Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.
- Baker, E. L., & Mayer, R. E. (1999, May/July). Computer-based assessment of problem solving. *Computers in Human Behavior, 15*(3/4), 269-282.
- Bewley, W. L., Lee, J. J., Jones, B., & Cai, H. (in press). Assessing cognitive readiness in a simulation-based training environment. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Teaching and measuring cognitive readiness*. New York: Springer.
- Bloom, B. S. (with Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of education goals. Handbook I: Cognitive domain*. New York: David McKay.
- Brockhaus, R. H., Sr. (1980). Risk taking propensity of entrepreneurs. *Academy of Management Journal, 23*(3), 509-520.
- Chung, G. K. W. K., & Baker, E. L. (1997). *Year 1 Technology Studies: Implications for technology in assessment* (CSE Tech. Rep. No. 459). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein, (Eds.), *Automated essay scoring: A*

- cross-disciplinary perspective* (pp. 23-40). Mahwah, NJ: Erlbaum.
- Chung, G. K. W. K., Delacruz, G. C., & Bewley, W. L. (2004). Performance assessment models and tools for complex tasks. *International Test and Evaluation Association (ITEA) Journal*, 25(1), 47–52.
- Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. *Proceedings of the IITSEC*, 25, 1811–1822.
- Chung, G. K. W. K., Niemi, D., & Bewley, W. L. (2003, April). *Assessment applications of ontologies*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Cline, V. B., Croft, R. G., & Courier, S. (1973). Desensitization of children to television violence. *Journal of Personality and Social Psychology*, 27, 360-365.
- de Jong, T., & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist*, 31(2), 105-113.
- Dragan, I., Tepper, K., & Misso, M. (2011). Teaching evidence based medicine literature searching skills to medical students during the clinical years – a protocol for a randomized controlled trial. *BMC Medical Education*, 11:49. Available at <http://www.biomedical.com/1472-6920-11-49.pdf>
- Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, 37, 65-84.
- Feltovich, P. J., Spiro, R. J., Coulson, R. L., & Feltovich, J. (1996). Collaboration within and among minds: Mastering complexity, individually and in groups. In T. Koschmann (Ed.), *CSCL: Theory and practice of an emerging paradigm* (pp. 25-44). Mahwah, NJ: Erlbaum.
- Fletcher, J. D. (2004). *Cognitive readiness: Preparing for the unexpected*. Alexandria, VA: Institute for Defense Analyses. Available at <http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA458683&Location=U2&doc=GetTRDoc.pdf>
- Gordon, P. L., & Bernstein, D. A. (1973). Anxiety and clinical problems: Systematic desensitization and related techniques. In J. T. Spence, R. C. Carson, & J. W. Thibaut (Eds.), *Behavioral approaches to therapy*. Morristown, NJ: General Learning Press.
- Graham, S. (1991). A review of attribution theory in achievement contexts. *Educational Psychology Review*, 3(1), 5-39.
- Guess, D. (1969). A functional analysis of receptive language and productive speech: Acquisition of the plural morpheme. *Journal of Applied Behavior Analysis*, 2(1), 55-65.
- Hargie, O. (2006). *The handbook of communication skills* (3rd Ed.). New York: Routledge.
- Iseli, M. (2011). *Ontology development: Overview and example* (Draft CRESST Whitepaper). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations*. (CRESST Report 775). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Johnson-Laird, P. (2006). *How we reason*. New York: Oxford University Press.
- Keltner, D. (2004). The compassionate instinct. *Greater Good*. Available at http://greatergood.berkeley.edu/article/item/the_compassionate_instinct/
- Koenig, A.D., Lee, J., Iseli, M., & Wainess, R. (2009). *A conceptual framework for assessing performance in games and simulations*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (IITSEC).
- Lee, J. J., Bewley, W. L., Jones, B, Min, H. & Kang, T. (2009). *Assessing performance in a Simulated*

- Combat Information Center*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).
- Lee, J. J., Jones, B., & Min, H. (2009). *Assessing performance in the Multi-Mission Team Trainer*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.
- Linz, D., Donnerstein, E., & Penrod, S. (1988). Effects of long-term exposure to violent and sexually degrading depictions of women. *Journal of Personality and Social Psychology*, *55*, 758-768.
- Mayer, S. (2010). *Problem solving ontology* (CRESST Whitepaper). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: McMillan.
- O’Neil, H. F., Chuang, S., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education*, *10*, 361-373.
- O’Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O’Neil Jr., (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411–452). Mahwah, NJ: Erlbaum.
- O’Neil, H. F., Lang, J., & Perez, R. S. (in press). What is cognitive readiness. In H. F. O’Neil, R. S. Perez, & E. L. Baker (Eds.). *Teaching and measuring cognitive readiness*. New York: Springer.
- O’Neil, H. F., Wang, S.-L., Chung, G. K. W. K., & Herl, H. E. (2000). Assessment of teamwork skills using computer-based teamwork simulations. In H. F. O’Neil & D. H. Andrews (Eds.), *Aircrew training and assessment* (pp. 245–276). Mahwah, NJ: Erlbaum.
- O’Neil, H. F., Jr., Wang, S., Lee, C., Mulkey, J., & Baker, E. L. (2003). Assessment of teamwork skills via a teamwork questionnaire. In H. F. O’Neil, Jr., & R. Perez (Eds.), *Technology applications in education: A learning view* (pp. 283-303). Mahwah, NJ: Erlbaum.
- Phelan, J., Dai, Y., Valderrama, M., & Herman, J. (2011). *Development of learning-based assessments in literacy: Towards the goal of college readiness* (Whitepaper Submitted to 2012 AERA). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Roger, D., & Neshoever, W. (1987). The construction and preliminary validation of a scale for measuring emotional control. *Personality and Individual Differences*, *8*(4), 527-534.
- Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, *52*, 471-499.
- State Standards Common Core 2010. <http://www.corestandards.org/the-standards>.
- Steele, C. M., Spencer, S. J., & Lynch, M. (1993). Self-image resilience and dissonance: The role of affirmational resources. *Journal of Personality and Social Psychology*, *64*(6), 885-896.
- Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 215–266). San Diego, CA: Academic Press.
- Vgotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Webb, N. M. (1985). Student interaction and learning in small groups: A research summary. In R. Slavin, S. Sharan, S. Kagan, R. H. Lazarowitz, C. Webb, & R. Schmuck (Eds.), *Learning to cooperate, cooperating to learn* (pp. 147-172). New York: Plenum.
- Weiner, B., Graham, S., & Reyna, C. (1997). An attributional examination of retributive versus utilitarian philosophies of punishment. *Social Justice Research*, *10*, 431-452.

21 世紀能力の育成とその成果を測定する新たな方法の展望

Eva L. Baker
(UCLA=CRESST 所長)

概要

本稿は、テストおよびアセスメントに対する私たちのアプローチをいかに変化させるかを考慮するために、予測可能な変化と予測不可能な変化が、いかにグローバリゼーション、メディア、デモグラフィ、そして、学習における新たな開発と関係しているのかを取り上げる。本稿では、まず、現状での特筆すべき点をいくつか述べ、それから 21 世紀スキルとそれらがいかに新たな文脈に対応するか、という点において議論を展開したい。こうした一連の流れは、テストの領域で将来を見据えて、エビデンスに基づきかつ推論された戦略を見つけるための新しい測定のデザインについて提案する。そして、吉本理事長のリーダーシップの下、日本の大学入試センターにおいてここ数年にわたって行われてきた素晴らしい仕事は、私たちの期待の質を更に高めることになる。

なぜ、いまアセスメントを問いなおすのか。グローバル化はもはや一般的で抽象的な概念ではない。それは、国家経済、地域経済、および国際的な経済のアップダウンの中心であり、明らかに操作できるようになっている。グローバル化は、報酬への期待、労働生活の時間、引退後の支援、高校と大学でのレディネスのレベル、そして大学と大学院の再分配を含んでおり、それは今現在そしてこれから必要となる労働力に必要な専門知識について明らかになる期待を、最小限に妥協できる方法で含んでいる。あるレベルでは、グローバル化は、連続的なベンチマーキングを強制し、価格と生産性を比較して不利な状況が、先進国に向けられるかもしれない。

また、グローバル化は、テクノロジーによって促される。過去 5 年の間に、仕事や社会参加、個人的な趣味、インストラクション、そして学習のために、モバイル（携帯電話）や他のメディアが使われるようになったのは劇的な変化であり、様々な点で影響を及ぼした。第一に、専門的知識を必要とし、視覚的またはグラフィックの刺激があり、さらに読書やテレビ・映画といった前メディア的な趣味の縮小が、学習者、生徒、そして大人自身に対して影響を及ぼしている。また、新しいメディアが作り出す個別化のエトスがあり、自分自身で興味関心を持つものを選択できるという期待や、人と関係機関との関係よりもっと対人間の関係に寄せる期待も生まれる。指導と効果が向上するかどうかにかかわらず、公式なテストと選抜テストが、テストの目的の中で強力な立場を維持し続けたとすると、両方とも個人の観点よりもむしろ組織の観点に根ざしている。したがって、それらは、現在のメディアのメッセージとは矛盾している。テストがパフォーマンスを予測できる環境の中では発展している間、メディアとテクノロジーは迅速な適応と変化の普及を利用している。そのため、変化のための新しい準備だけが重要なのではなく、今のコースを考え直すことも必要であり、アセスメントを行う際には、何をどのように行うべきなのか、に着目しなければならない。

理想的には、アセスメントためには、以下のような3つの構成要素を目的に取り入れなければならない。(1) グローバル化の猛追にも負けないような、維持されるべき文化的価値の強さ、しなやかさ。(2) アカデミックな知識に融合された規律的で実践的な知識。(3) 学習の適切さや動機づけを維持するために必要な、生徒やトレーニングを受ける個人の目標の3つである。そこで、もしわたしたちが、学習者の個別化した学習の増加の予測について同意するならば、どうやってその目標を測定することができるのかについて考える必要がある。

アセスメントの結果の多くは、直接学習者や生徒に伝えられる。たとえば、入試やプレイメントがそうである。ごく最近では、アセスメントは、教室における先生の個人的領域での役割を担いはじめ、学習効果の診断、フィードバック、そして保証へのより標準化されたアプローチを提供している。しかし、ふりこは後方へ揺れており、希望の学校と適合させるような一連の個々のパフォーマンスを重視する方向へ流れている。

このことは、わたしたちをアセスメントの目標の考察へといざない、この議論の中核は、21世紀スキルの議論を正しく位置づけることにある。

表1には、アセスメントの共通目標のリストがあり、その目的が本質的に個人に関係するか、学校に関係するか、によって分類されている。

表1 アセスメントの従来の目的と使用法

生徒	学校
入学	ステータス
配置	比較
コミュニケーション	改善
動機づけ	人員決定
診断	制裁および報酬
フィードバック	社会と政策評価の質
改良	
保証	

目的がアセスメントの重要な要素である一方、フォーマットの点からアセスメントの分類する一般的な方法は、学校で一般的に経験されていることである。これらには、手続きを習得する際に起きる問題のまとめ、内容の理解をサンプリングする選択肢の項目、プロジェクト、エッセイ、リサーチペーパーやそのほかにも学生がつくったもの、長期にわたる反応などが含まれる。時々、この対比は主観テストと客観テストの間にも起こる（構築された反応が任意の測定と同じくらい厳密に客観的になりうるとしても）。それらは学生に公開されている表面的な特徴という点でもコントラストをなすだろう。たとえば、ペーパーテストやコンピューターや他の技術的な方法を通して処理されるかもしれない。しかし

ながら、アセスメントにおける計算原理は、電子ページターニングと呼ばれる処理の表面的な形式を超えるものを含む。テクノロジーの新しい発展は、テストの間の生徒のパフォーマンスのレベルにまでアセスメントが適応されることを可能にする。そのことは社会、科学、数学の環境において、複雑な状況を忠実に表すことを提供するかもしれない。または、特定の成果や資格取得に焦点が絞られた「ゲームライク」な設定の中におかれるかもしれない。そのようなコンピューターアセスメントは自動的に、即席で得点化され、エレメントとモジュールからなり、素早く生成することができる。近い将来、そのようなアセスメントはリアルタイムで必要に応じて作られるだろう。

しかし、それでも、目的と特性は、それらの実際のコアに関して動くが、それもアセスメントの属性なのである。

どのようなアセスメントも中心となるのはその内容である；スキル、内容、ストラテジーは、学習によって影響を受けた思考プロセスの行動上の表現である。教育的な性質において全てのアセスメントの主要な焦点は、生徒のパフォーマンスのサンプルの決定に従い、生じた学習の程度について理解することである。わたしたちの主たる焦点は、今世紀に要求される学習へと変わらなければならない。

「21世紀スキル」という用語は、将来の教育とトレーニングにとって不可欠であると考えられる認知的スキルと社会的で感情的な能力として広く採用されたメタファーである。それらは、ゴールが個々の能力を高めるか、生徒をより高い教育（たとえば、技術的トレーニングや大学）のために準備できるかどうかに関係している。また軍隊を含む現在または将来の労働環境について高い価値があると思われるスキルについて考慮するかどうかということにも関係する。認知的スキルに関する多くの議論において強調されるのはいつも個人の選抜についてであり、インテリジェンスやクリエイティビティのような一般的に測定される適性である。この問題はトレーニングや教育の側面とは異なる。どのようにスキルや気質を変えるかという点に注目しながら、数学や科学、歴史のように詳説された内容領域のコンテキスト内においてスキルを発展させるべきだという主張がなされている。また、母語やその他の言語リテラシー能力のようなスキル領域の典型的な定義についても議論がなされている。二番目に、これは職場やアカデミックな環境では比較的最近のアイデアだが、問題解決能力のような認知力に関して、つまり特定の範囲でいえば代数に関しては十分に教えられているという。私も含め、他のものはこの点を信じている。さもないと、段階的なアプローチを信じ、提案する。最初に、生徒が内容の主要な領域のスキルを獲得することを学ぶために役立つようなことに注意が向けられなければならない。次に、その領域で重要な能力を示すために、パフォーマンスは興味のある内容領域の十分な広がりとは十分な深さを詳説しなければならない。三番目に、指導とアセスメントに対する関心は、学習者が精神的なスキーマやパターンに組み入れられた1セットの法則を得ることを構築し確認することに向けられるべきである。もし、トレーニングを受ける人がプロセスの表面的特徴を記憶するのではなく、認知的要求の鍵となる側面を効果的に取り出すと予測される場合は、このステップはとても重要である(Sweller, 2003)。最後になるが、注意は内容の幅広い範囲のため必要とされるだけでなく、むしろ、状況に付随した概念のために明確に求められる。状況、制約、内容の構成要素、そして解決や行動の質の特性が同時にそして異なる段階で変化することもある。そのような予測不可能な新しい状況に対応する能力に関して、「認

知的レディネス」という言葉を使用することは、認知的要求もしくは 21 世紀スキルに対比して使うこととは基本的な違いがある。「レディネス」という言葉の表す期待感には、思いがけない課題に立ち向かう能力が意味される。このようなレディネスを達成するためには、新しいセッティングのさいに彼らの学習を転移させるような能力が開発されるようにしなければならない。そのため、トレーニング受講者、または学習者は、条件、状況、そして問題セッティングの十分に多様なセットにさらされなければならない、それにより学んだことを新しい場面で生かす能力が磨かれるだろう（上記のスキーマを適応することと関連する）。

教育においては、現在まで、大部分のアセスメントは認知的な要求（たとえば適応、危険負担あるいは状況認識）を明示的には選られない。その代わりに知識内容そのものに関する手がかりに着目し、繰り返しの手順の中に対応させている。最近の研究についても (Baker, 1997a,b; 2007a,b)、その焦点は、絶対に確かな領域の文脈において、学習とアセスメントのタスクをデザインすることにあつた。これは事実であるが、知識内容が改善されると、概念が生成されるかあるいはさらに詳細化される (State Standards Common Core 2010 参照)。継続的な知識の爆発が示すことは、単調な文脈や一定の領域の概念はもはや古いということである。学習者は適応可能なスキルが必要となり、あるいは、これらのスキルが適応するかもしれない全ての文脈と内容を他に再び学ばなければならない。

それ以外のスキルに関しては、それらが適応する文脈や内容すべてを学び直すべきであるが、それは明らかに不可能な目標である。予見しがたい将来への適応性が重要であるなら、アナリストに期待されることは今日知られていない新しいキャリアとタスクを 5 年以内に予測することである。そうすれば、学習者は、新しい必要条件を満たすようスキーマがどのように修正されるか決定するために、彼らが同様に学んだ原理やスキーマを適用させることができるにちがいない。

したがって、スキーマとパフォーマンスの変換に加えて、学習者は予測できない新しいスキルセットを開発する必要があり、そのスキルとは問題解決のためや、理由づけ、決定のためにこれまで教えられた方法の組み合わせであり、あるいはその修正であるかもしれない。

まとめると、この段階において、軍事訓練や仕事の世界では、21 世紀の認知的スキルの何らかの適応が強調されており、そして、新出現で不確かな状況の可変的な文脈に集中することが重要だと言える。学術的な学びに関する研究では、将来の文脈における「不確実さ」の解釈は、しばしば限定され、「転移」状況という点に集中してきた。つまり、学習者がそのスキルをこれまで未習の状況に適応する必要のある仕事領域または制約に集中してきた。「アカデミック」な設定と、軍隊や職場でのトレーニングの違いは、図と地の相違にあるかもしれない。そこでは内容やスキルレベルもしくは文脈における強調は、見通しの問題であり、しかし認識の相違は学習とアセスメントシステムのデザインの両方にとって重要性を持つ。

つまり、人はスキルが内容の中へ組み込んだことを理解するか、それとも、人の注意が文脈の変化に適応することがあるのか。これらはアセスメントのデザインや先行する学習経験のデザインのアプローチを修正させることがあり、これまでの学習経験をも変化させるかもしれない。この見通しの相違は、人が 21 世紀スキルバスケットにいるかどうか決定するための、一つの有効的なマーカーかもしれない。

21 世紀スキルの考察

アセスメントデザインについての私が以前に持っていた考えでは、アセスメントがデザインされたとき、わたしたちは認知的スキルに注目しなければならないということだった。初期の概念は、内容の、そして、テストフォーマットの手法（たとえば多枝選択式の項目）についての関心とは対照的である。以前のリストは、以下の図1に表す。

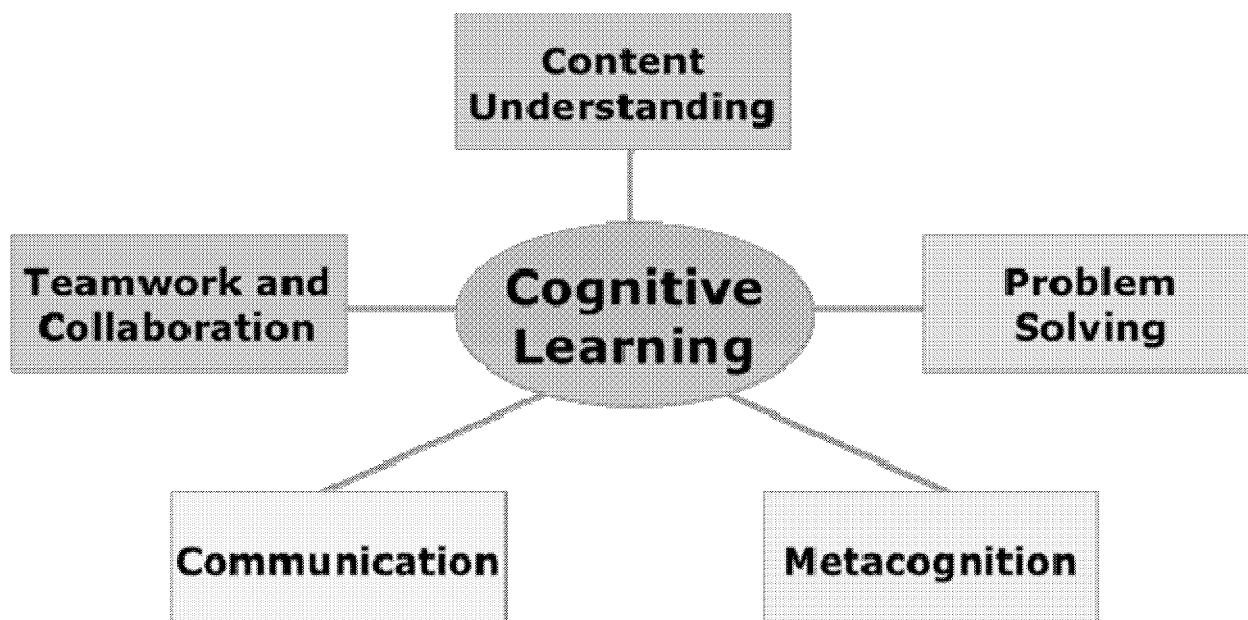


図1：アセスメントデザインの単純化された認知的需要

21世紀スキルの **formulation**（組織・構成）はますます複雑に発展するとともに、その使用についての文脈は予測しにくくなり、高度なパフォーマンスの源について考慮しなければならなくなった。比較検討すべきものが主として2つある。一つ目は、生まれながらの才能、特性であり、言い換えると、敏感であるとか、好奇心旺盛であるとか、流ちょうであるといったものであり、学校で生徒によくみられるものである。二つ目は、学習で得たスキルで、それは学校であるか、他の場面であるかどうかにかかわらず、はっきりと学習されるスキルである。これらの特性と学習間の相互作用は明らかである。好奇心の強い子は優れた問題解決をするかもしれない。わたしたちのテーマに関して言えば、比較的安定した特性はさておき、21世紀スキルは、プログラムや組織の中で系統的に学ぶことを意図するスキルに焦点をあて、そして、認証または予測のための何らかの達成を判断するために測定されるということに焦点を当てる。

21世紀スキルセットに含まれるスキルの範囲とは何か。それらのどんなリストも純粋なオリジナルではなく、似たような概念に関連した同義語を使用するので言語的に重なる。このスキルを主要な3つに分類した。それは、①知的な認知的プロセス②社会的に適応したプロセス③個人内スキルである。①と③は個人がそれを独立して実行することができることに焦点を当てている。これらのスキルはたとえば、「問題の重要な特徴を確認するためにどのように調べるか」といった外界からの見方、あるいは、内面に目を向けて、「私は望ましい目標を達成するために、どのようにして私自身の認識（もしくは感情）のプロセスを管理するのか」といった場合がある。ほとんどの場合、すべてのスキル、特に学校とトレ

ーニングのタスクは、特定の科目に深く埋め込まれている。たとえば、子どもに対して、川を渡る方法を考えるように課題を出す。そして、彼らは与えられた物を使って、勢いと摩擦についての力と運動の基本的な原理を適用させるだろう。高校生の場合は、与えられ推論された制約のセットにあうように、図の最適の形を決定するよう求められるかもしれない。また海軍士官は、防衛措置を開始するかどうかを決定するために、信号とサインを解読する必要があるだろう。内科医は筋肉痛に診断をくだし、患者に行動方針を提案する必要があるだろう。

知的な認知プロセス

知的プロセスのエリアでは、特定の目的を達成するために、関連した下位プロセスを必要とするような認知的タスクがある。問題解決について考慮すると、問題がよく特定できない状況と問題が明白な状況のもとでは、思考するスキルは異なるセットを有する。(Feltovich, Spiro, Coulson, & Feltovich, 1996)。

問題を特定し確認した後に、学習者ないしは受験者はその問題を明確にし、言葉で、図で、あるいは象徴的に表すに違いない。そして、タスク指示が「正しい」、もしくは収束する答えを要求するかどうかを判断し、あるいは、彼らが学習されたスキルか革新的なストラテジーの新しい組み合わせを必要としているような互いに異なる反応を許可するかどうか決定しなければならない。もし、収束するような解答が求められているなら、学習者は最適のパスを見出さねばならないし、テスト設計者の一般的な解と一致しなければならない。仮定された解決策が問題を解決することを確かめるための手続きは、ほかのタスクとの関連がある場合がある。これらのスキルのいくつかは、比較的変わらない定義を持っており、たとえば、この領域の（仮定からの）帰納的推論と（前提からの推理による）演繹的なアプローチの作業間の違いは、古典的な時代にさかのぼる。しかし、Johnson-Laird (2006)による最近の出版物には、この領域の論拠がまとめられている。より新しい原理は、検索の認知的スキルかもしれない。ここでは、生徒は検索を行うが、しかし、彼らの検索結果を確認するための証拠のモデルを使用することが奨励される。しかし、先行文献では、情報科学の分野の多くが、上手く検索エンジンを使う能力（特別なコツ）にとってかわっている。検索エンジンがほぼすべての大学の図書館に提供され、アウトプットの質を判断する基準は多様化しているのである（たとえば Dragan, Tepper, & Misso, 2011 を参照）。質の基準を適応させるための能力はますます必要となり、技術ポジトリが「クラウド・ソースである」場合もあり、質の基準を適応させるための能力はますます必要である。たとえば、Wikipedia である。いかなる質の管理にも影響されないという状態にまで行くにつれなおさら必要である。

もう一方の関心領域は、宣言的で、手続き的で、系統的な知識の理解である。この理解というのは、以前より重要であり、認知的スキルは科目と状況に深く埋め込まれている。この理解のための3定義（なぜ、何を、どのようにが、それぞれ宣言的、手続き的、系統的知識に対応する）には、長い歴史がある。Alexander and Judy (1988)はこの領域の論文をレビューしており、de Jong and Ferguson-Hessler (1996)による初期の優れた研究をも視野に入れて、知識と質のタイプについて言及している。これらのカテゴリーは、Anderson and other cognitive psychologists (2000)によって改訂された Bloom's Taxonomy of Educational Objectives (1956)と結合された。

表2は、潜在的な21世紀スキルのリストであり、知的な認知プロセスに重点が置かれている。これ

らが効果的にかかわるとき、それぞれのスキルはどの段階で効果的に作用するのか十分な準備もなく「実行中」になり、認知レディネスの領域に溶け込んでいく。

表 2 21 世紀スキル—知的な認知プロセス

-
1. 適応性のある問題解決
 2. 意思決定
 3. 状況意識
 4. 帰納的・演繹的である推論
 5. 見つかからない資源の探索
 6. 宣言的、手続き的、系統的知識の理解
特定の下位分野の中での
-

2つの点が明白である。第一は、知的な認知タスクのそれぞれのリストは、分離して操作されないということである。あるタスクが名目上焦点が絞られていても、ほかのアプリケーションに依存していることがある。たとえば、問題解決には状況認識(Koenig Lee, Iseli, & Wainess, 2009)と意思決定を必要とするかもしれない(Lee, Bewley, Jones, Min, & Kang, 2009)。第二に、それぞれのスキルは、様々な異なる表層の形をとるかもしれない。これらのスキルは異なる原理、概念、事実、繰り返されるアプローチからなる領域の中で異なって展開される。事例を示すために上記の問題解決タスクの領域に戻り、不明確な問題について考究する。問題はいくつかの方法で不明確である場合がある。これらの解決に向けたストラテジーを導くことは、十分な情報もないまま、あいまいになることもある。それから、解答者は、その問題についてもっともらしく思われる別の形に直し、適切な方法でそれらを表す必要がある。そして、解決のためのストラテジーや手続きを適用し始める必要がある。もしも確認される問題が作成者の意図したものでないならば、はじめに戻るか、延々繰り返すことは明白である。そして問題は、その一部が隠れたりふさがれたりしているので、問題自体も不明確かもしれない。つまり、学習者は、実際の問題、視覚的で口頭による刺激、あるいはその両方を見つけるために無関係な材料の中を努力して進む必要があるかもしれないということである。言葉か、ヴィジュアルか、またはその両者が刺激になるかもしれないが。タスク作成者、つまり出題者は、解答者を罠にかけようと熱心に働き、間違った解釈をさせようとする。この種の不確実な問題を突破するために、学習者は、間違ったおとりを回避するようなタスクの状況と内容領域について十分に知っていなければならない。問題のセットアップにおいて、作成者が明示的かそれほど明白ではない制約を含んでいる場合、そのとき解答者は、本当の問題に関して焦点を当てられるように、状況認識のスキルを使用しなければならない。解答者が、それほど中心的ではない材料を排除し、そしてどの変数がタスクの重要な決定要因であるかについて判断することができるならば、かれらはまた、問題識別段階でさえ、類推のスキルを使用する必要があるだろう。

複雑で多段階な問題のキーポイントについては、解答者が決定する必要がある。これらの決定は、過去の経験、つまり、タスクに関連して提供されるか得られる知識に基づいてなされ、そして、類推の応用に基づいてなされる。

分析する者が、そのようなスキルを「特徴」や生まれつきの個体差と考えるとき、優れた「問題解決者」は論理的思考者、または環境状況変化を探知できるものと仮定され、彼らは、その領域の中で、領域を超えた特定のトレーニングなしに、これらのスキルを一般的に適用することができるにとらえられている。言い換えれば、認知的スキルは、ほぼ同じレベルのスキルを備えた特定の内容領域とは独立に使用される。しかしながら個人では、関連領域によって規定された能力をもつだろう。たとえば、数学における優れた問題解決は、文学においても解答者が同じ態度をとるというわけではないのである。しかしながら、わたしたちが仮定している最初の幅広い領域内でのスキルの転移を要するトレーニングは、その後、その領域とその知識的なスキルを適用する様々なシチュエーションにおいて、内容の領域を超えた形での適用が求められるトレーニングによって拡大される必要があるだろう。このようなパフォーマンススキルについての領域から独立したセットを開発できるかどうかは、はっきりとしないままである。解答者が問題を解決するかあるいは決定するためにストラテジーを使うことができる以前に、それぞれの領域は一般的に比較的深い宣言的、手続き的、体系的な知識を必要としている。非現実というよりは現実に近いものの考え方が、幅広い領域を越える深い知識を要求することは傾向として幾分少ないだろう。このように、知的な認知スキルの領域から独立した運用のもっともらしさは、互いから遠い領域を学ぶための時間、関心、そして能力の範囲に依存する。

社会的、対人的スキル

21世紀スキルは、社会的に適応したスキルを含み、認知も含むが、より対人的な状況の中でのその運用を必要とする。これらの能力は一方では、仕事の共同的な性質を含んでいるかもしれない。あるいは、計画の承認を得るために交渉し、仕事について議論し、そして理解し、それを対象ユーザーに報告するために情報を発信する必要性を含んでいるかもしれない。

さらに、社会的に適応したスキルは、明るく、社交的な性格によって決まるというわけではない。たとえば、共同のエリアでは、グループの中で意見が異なるかもしれない状況において、アイデアの価値を認めるために態度を修正するように、チームの目標や共同を明確にすることを学習者に要求している。表3には、そのようなスキルの部分的なリストを示す。

表3 21世紀スキルの社会的適応

-
1. チームワーク
 2. 共同作業
 3. 手助け
 4. 社会状況の認識
 5. コミュニケーション—生産的、受容的
-

表2のリストされる知的なタスクを使用して仕事で成功するには、学習者は、資源として他者を頼る必要があるかもしれない。または、人材に関する解決に向けて必要とされる一部だけを提供することも

しれない。チームワークについて Salas&Cannon-Bowers (2001) は、タスクに成功するために、リーダーシップ、フィードバック、動機づけ、リダイレクション、説明、誘因および努力を提供することが必要かもしれない一連の構成要素スキルを作成した。すべてのタスクがこれらの構成要素のすべてを求めているわけではなく、また、どんな人でも、与えられたタスクのためにそれらすべてをこなすことを要求されていると思うべきではない。しかしながら、学習の展望から、よいチームメンバーまたはグループメンバーであるためには、それらのタスクのうちのいくつかは必要だろう。たとえば、リーダーに必要なタスクは、異なる条件下において時々違うタスクを必要とする。目標の明確さは永続的に求められる場合がある。しかし、目標は、一人一人がばらばらに持つのではなく、グループとして目標を明確にさせることが賢明な場合がある。もちろん、適用の方法については、組織的階層と個々のステータスに依存する。正しい選択をするために、チームのメンバーは、そのチームやグループの多様な地位を認識しなければならない。参加したいという意欲、それらの役割、および他の情報、たとえば経験や新しいものに取り組もうとする情熱といったことも含めて認識する必要がある。社会的状況認識のトピックについての研究は、しばしば感情移入と呼ばれている (Keltner,2004 参照)。ここでは、生まれつきの特性と対照的なものとして、学ぶことがタスクとして組織立てられるため、それは区別される。社会的領域における相互作用は、直接的にせよ、連続的または時折にせよ、メディアを通してあるいはテキストを媒介にしているにせよ、コミュニケーションについてのいくつかのレベルの専門知識に依存するかどうかにかかわらず、その相互作用は明らかである。

そのスキルは、重要なタスクを組織化し、構成し、説明するか、あるいは重要な問題に質問し答えるように要求される。コミュニケーションの別の面は、適した言語を使う繊細さ、という点である。適したオーディエンス、組織、タスク、特定の個人間の文脈で、むしろ文化的背景も踏まえて使うという点である。それから、これらは、社会的状況認識の領域について適切な関連したスキルにおける進化しているセットを形成し、知的スキルと個人間に関するスキルを組み合わせる。個人間のスキルは、きわめて経験的で、文化的および個性的な基礎を持つかもしれないため、それらを身につけるためのトレーニングまたは学習が可能かどうかについては、はっきりしていない。しかしながら、主要な構成要素、たとえばコミュニケーションは学習によってえられ評価されたという長い歴史がある。いくつかの研究では、非常に特定のレベル (たとえば Guess, 1969 参照) で、または個人間のスキルのより特定のセットとして、受容的 (聞く・読む) および生産的 (話す・書く) な関係を確認している (Hargie, 2006)。Vygotsky(1978)は、社会相互作用とコミュニケーションを高度な心理学的プロセスの中心となるものにとらえた。共同の要素、またはチームワークの要素、特にタスクを確認するために関連した特定のステップまたは役割を理解し実行するための要素もまた、十分に教えることができる。チームの中で働く際には、個人間の構成要素は、規定の時間を越え、ストレス状況下に置かれるにつれて、様々なタスクの状況をこなすうちに発達するだろう。タスクは、グループのメンバーが個々に独立して完成させるもの、または、各チームメンバーがその目標の達成に対して相互依存的に寄与するものまで多様に変化する (Webb,1985)。

個人内スキル

個人内スキルは、指導または学習経験によって組織的に得ることのできる、もしくは強化されること

のできる個人の態度や内部思考プロセスであり、その変化の根拠は、直接あるいは間接的に推論することができる。

表4には、そのようなスキルセットの部分的なリストがある。

表4 個人内スキルのセット

-
1. プランニング
 2. フィードバックを含む知的タスクの自己管理
 3. 情動認識および自己統制
 4. リスク負担
 5. 成功と失敗の動機づけられた努力および帰属
-

これらのスキルのうちのいくつかはメタ認知のカテゴリー、たとえばプランニング、セルフモニタリング、計画的な努力の下でのクラスター分けられるのに対し、ほかのスキルはより情動的な構成要素を持っている。たとえば、個人は、情動をコントロールすることや、よりバランスのとれた振る舞いを心身の管理を通して実践すること、あるいはリスクを推定し、状況（たとえば、コスト、成功の予測、失敗の結果）に応じて危機負担の傾向を変えることを学ぶことができる。（これらにはある程度は教えることができるという一般的な証拠がある）。社会的・情動的スキルは学ぶことができ、そして、その中に帰属、自己効力感、めげにくさ（resilience）も含む（Weiner, Graham, & Reyna, 1997）。これらは自己認識の構成要素である。ストレス下での情動のコントロールについては人格心理学に由来している（Roger & Nesselhoever, 1987）。長年の研究から数多くの事例があり、たとえば、脱感作研究（desensitization research）があるが、それは治療方法の文書化（Gordon & Berstein, 1973）から性差別の影響（Linz, Donnerstein, & Penrod, 1988）やメディアやゲームにおける暴力シーンの影響（Cline, Roger & Courier, 1973）にまで及ぶ。

結果についての判断感覚によって緩和されたリスクテイクは、自己動機づけされた行動やリーダーシップ、起業家精神の文脈において、より通用するもう一つの個人内スキルである。この研究は、同様に長い歴史を持っており（たとえば Brockhaus, 1980 参照）、そして、マネジメントの学部・大学院の文脈でよく研究されてきた。

努力と帰属について相互作用している領域は、Graham (1991)によってよくまとめられている。そこで、研究者たちは、個人の努力における成功と帰属の関係性を最初に示し、そして、失敗するよう選択された、あるいは学習のコントロールがきかないと思う人々に、努力と帰属の概念を発達させるためのトレーニング療法を開発した。研究におけるこのラインは、ステレオタイプ脅威の研究と、そして、アフリカ系アメリカ人によるそういった認識を克服する方法と非常に関連している。Claude Steele (Steele, Spencer, & Lynch, 1993)は、社会的に脅威となる状況の下において、めげにくさ（resilience）を支えるために自己認知が果たす役割について調査し続けた。

それは創造力、批判的思考力、思慮深さといったほかの要素もある。そういった能力は、インストラクションに従うか、そうでないかもしれないが、能力構造がどのように組織化されているのか、文化、

年齢、学習経験という文脈に依存しているのである。しかしながら、21世紀スキルの議論においては、このような要素は教育、トレーニングおよび経験によって変えられるとされている。

個性差は、これらの領域の各々へ適応され、そして、徹底的なトレーニングは、個人内スキルをハイレベルまで高める可能性がある。その一方で、何らかの背景の違い、たとえば失敗にともなう経験などは、そのトレーニングの効果が出るのをより難しくするかもしれない。

知的かつ社会的でそして個人内の21世紀スキルで例証されるような要素が、教育とトレーニングの状況の中で使用されることを意味するのならば、Fletcher(2004)、O'Neil, Lang and Perezらやその他によって記述された認知レディネスとの違いは何であろうか。あるレベルにおいて、認知レディネスの概念は、機敏で、適応性があり、不確かなことに対する準備といった能力を強調するものである。21世紀スキルの多くの要素は、そういう強調を共有しているが、「レディネス」という用語が予測できない状況で行動する能力を意味するということは、明らかである。ほとんどの場合、21世紀スキルは、組織に基づいた学習、学校、大学、または仕事の世界に与えられる。「認知レディネス」という用語が軍事的な文脈において使用される場合、それは急速な状況の展開や、予測できない挑戦と制約といった、熟考や長期的分析よりもむしろ迅速なことが必要条件であるというイメージを彷彿とさせる。認知レディネスモデルでは、比較的訓練しやすいものから訓練の難しいもの、そして個体差の特徴とみなされるものまで及ぶ、連続体の認知レディネスの特性を分類している幾人かの研究者がいる(O'Neil, Lang, & Perez 印刷中)。

アセスメントの構築にむけて

アセスメントを構築するために、わたしたちは、知的、社会的、個人内についての21世紀スキルの配列から始めたい。そして次のステップは、内容が測定に含まれると考えることである。

21世紀スキルのこれらの3つのタイプのどれかを実行する中で、この2番目のステップは、オントロジーの使用、もしくは興味の内容領域のマップを含む。オントロジーとは、言語を図表で表現したもので次のような特徴をもっている。ノードのネットワーク、ノード間の関係を描いた一組のリンク、そしてパフォーマンス情報に基づいたノードとリンクの配置と方向を修正するデータベースである。オントロジーが、学習またはアセスメントのために使われる言語領域の図表表現であるならば、その解釈された特性とは何だろうか。オントロジーは、法則、概念、重要な知識、そして手続きを特徴とする。これらはネットワークのノードとして描かれる。ノード間の関連として、方向と意味を持っている。それらは、階層的な表現で構成要素が他の部分集合であることを伝えるかもしれない。それは、期待される高次の複雑な内容の順序から、初心者でもわかると予想される基本的な法則と事実にいたるまで幅広く並べられる。数学は、この分野についての階層的な要素を示す優れた領域である。異なる内容については、そこでの法則、概念そして事実が、年代、テーマそして法則を伝える方法を示すためにリンクされる場合、オントロジーはまた適切で活発な方法によって構成させるかもしれない。たとえば、政治史は、年代に基づいた構造を持つだろう領域の例である。また、内容オントロジーの別の異なる構造型としては、ノード間の(法則や概念、または例を含む)関係について、相互関係や互いに方向のある影響が説明できるだろう。そのようなリンクの中で最も強いケースは、因果関係を例証するものだろう。大部分の研究分野では、人は、部分と全体の関係性を反映させ、他の主題や年代、または因果関係のミックス

された関係を見つけるだろう。いくつかのエリアにおいて、幅広い領域が考慮される場合、構造は不安定である場合があるだろう。たとえば、文学において、小説や詩のようなフォームの間の関係は、階層的であるよりもむしろ平行的か水平的である。それでも、文学的なフォームの範囲内では、より強い垂直構造は見つかるかもしれない。筋書やキャラクターの開発の場合には、非常に明示的な関係を開発することができる。

オントロジーの構造は、領域の本質的な特性に沿わなければならない。これらは、解釈の過程があらわされる範囲によって変化するかもしれない、それは比較的是っきりとした結果を生み出すことを目的とした特定の方法論とは対照的である。これらの領域はまた、物理のような、理論的あるいは経験的である抽象的な法則が表される範囲という点からみても異なり、叙情的な詩のように緩くつながれているような領域とは対照的である。図 2, 3, および 4 は、内容、細分性、そして表示によって変化するオントロジーを表す。

Map of Algebra I Big Ideas

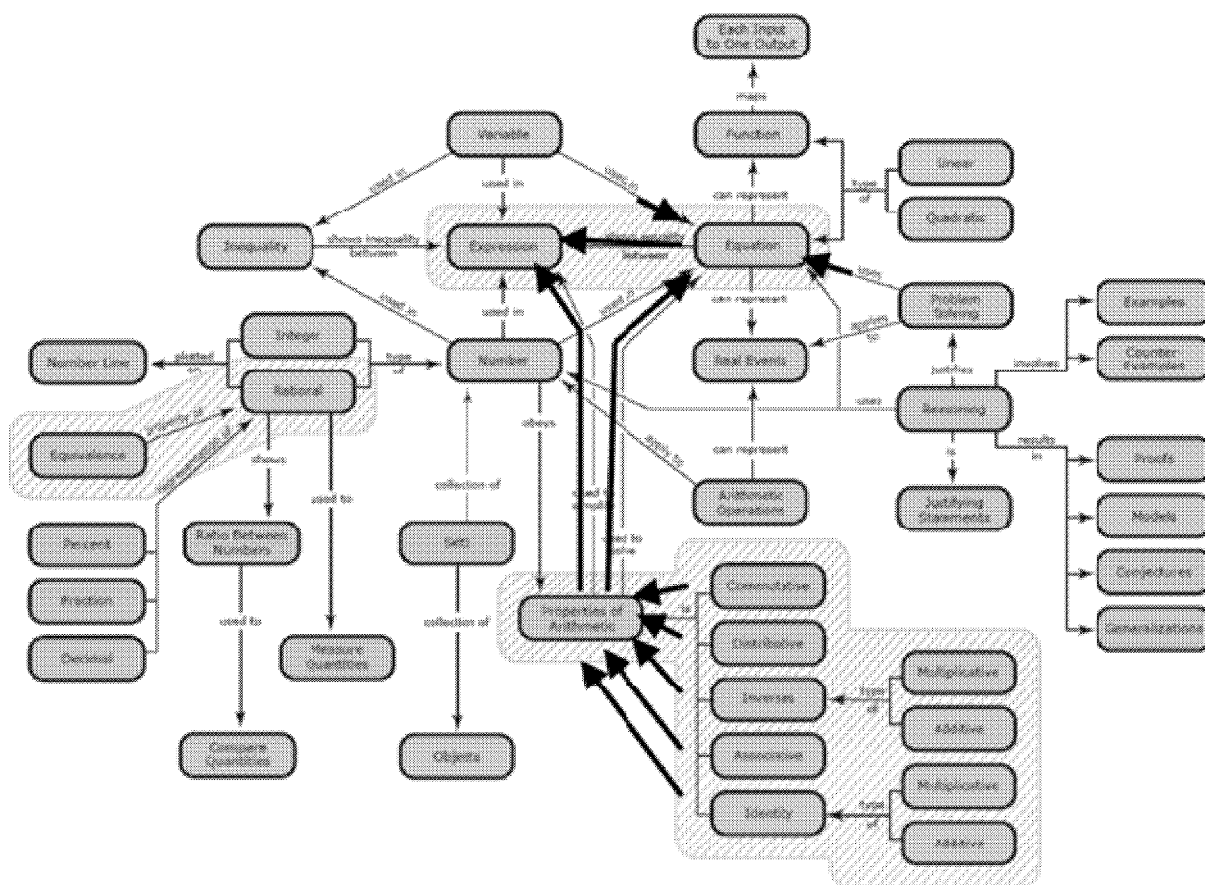


Figure 2: Content Ontology of Beginning Algebra

Map of Algebra I Big Ideas

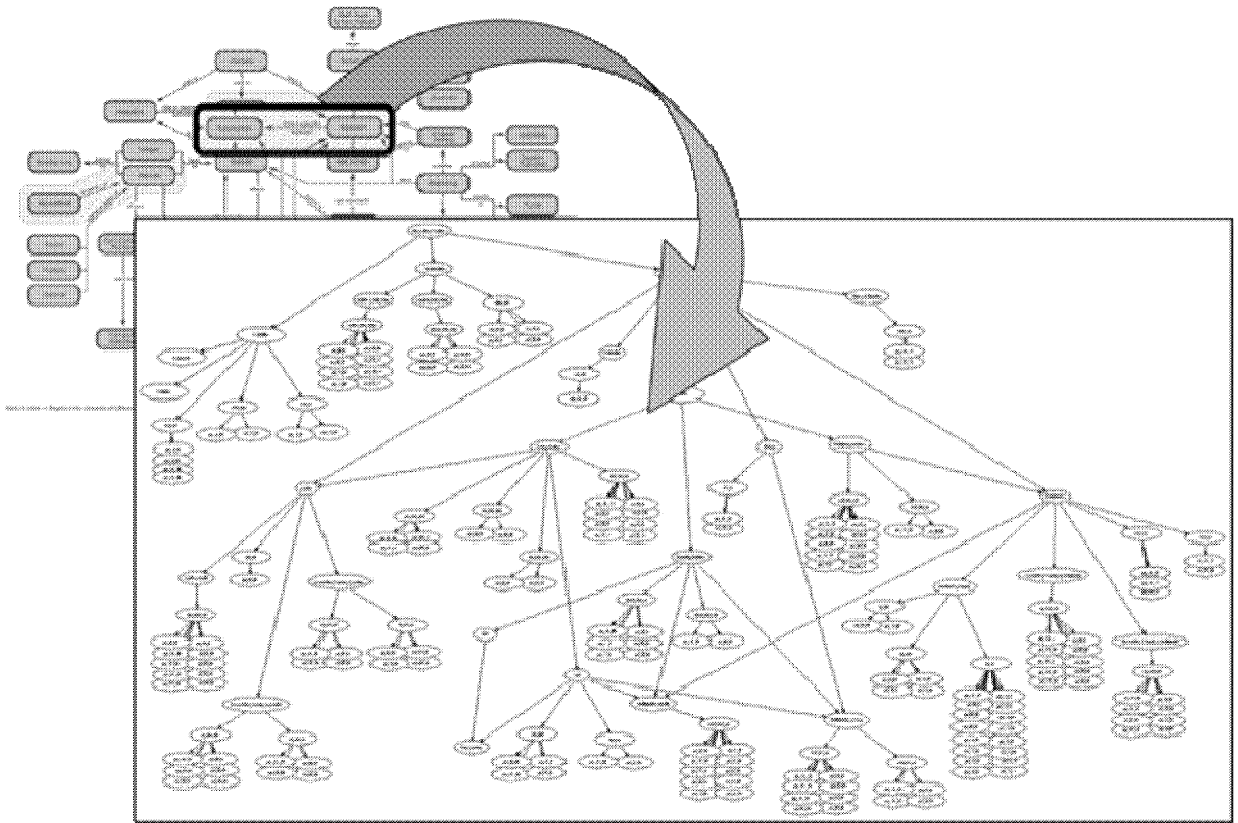


Figure 3: Algebra Ontology–Dynamic Bayesian Network

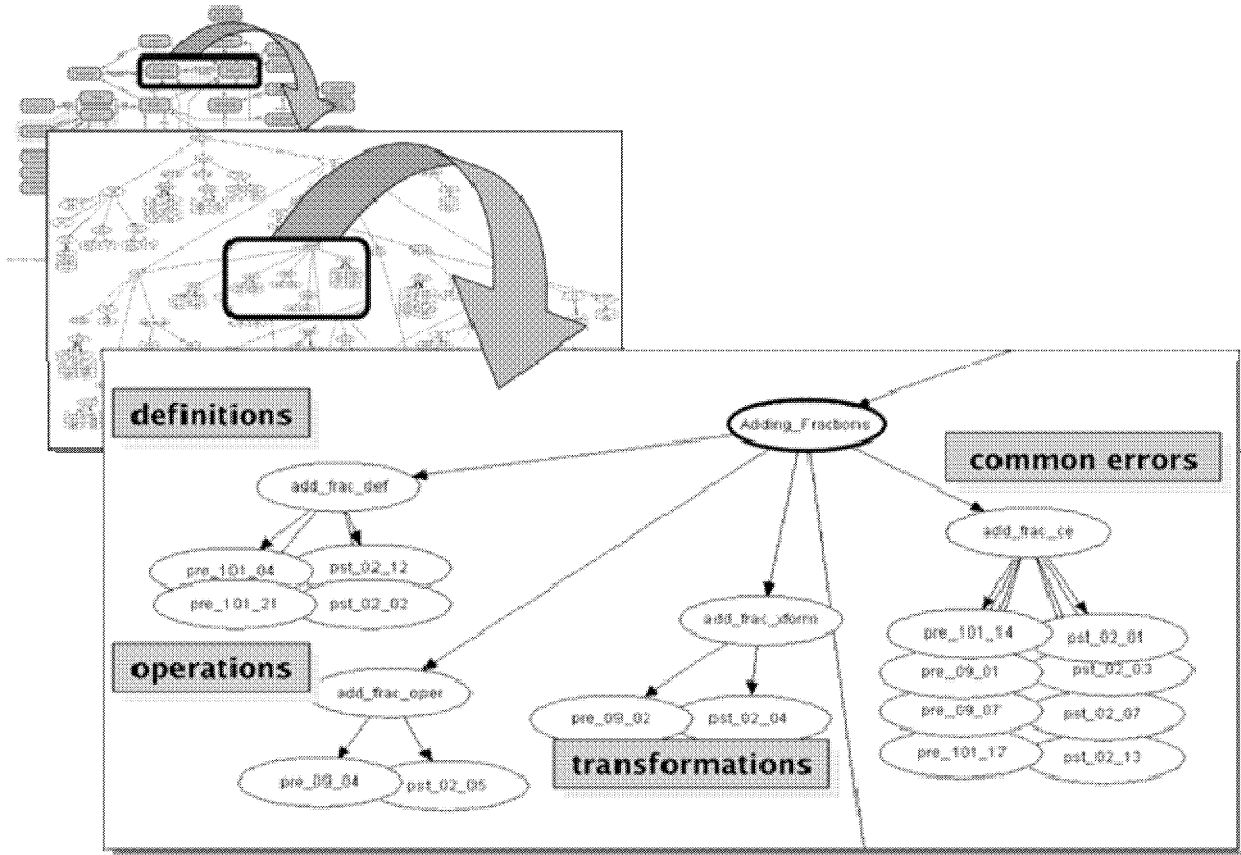


Figure 4: Algebra Ontology-Decomposing Data

オントロジーの共通要素

構造とは無関係に、どのようなオントロジーにも特定の特徴がある。第一に、図表を用いた表現である。第二に、どんなに多くのノードが他のものに相互にリンクしているかについて何の制限もない。実際の内容の中心的位置は、単純に高い頻度のリンクに伴うノードの数に注目することで決定することができる。またどの内容が、遠くて潜在的に必要なではないかどうかを、中心となるアイディアとの結びつきの弱さから決定することもできる。

この重要な操作上の描写は、学習システムのデザインについての直接的な推論に導いてくれる。いくつかは、オントロジーをシーケンスの指示、またはコンピューター適応型テストの最適な配列として見るかもしれないが、そのような推論は直接つくられるべきではない。たとえば、「学習の進歩」という用語を、あたかも、どのシーケンスが最適だったかについて本当にわかっているように使用することは格好良くみえる。しかし、実証研究だけが、学習者と環境のために、学習とアセスメントの順序に関するより強い仮説を生み出すのである。このタイプの実証研究は、ビギナーの中で能力を発揮する人や、ただ一部分的にしか教わっていない人、そしてパフォーマンスの中レベルを実行することができると知られた個人が、後続するプロセスを対比するためにこのタイプを保証することができるかもしれない。

他の実証的な研究は、タスクの仮定された配列に取り組むことができ、どの内容とプロセスが、期待される目標の達成にとって必要かということを決断するために系統的に代替の要素を削除することができる。

オントロジーはどのようにして生まれたか

オントロジーは、通常、人間の手で繰り返しつくられている。ある主題のエリアの専門家が、重要なトピック、および他の概念、法則あるいは事実とのそれらの関係に関して、関心のある領域をあらわすように示しなさいと言われたらと想像して欲しい。オントロジーについて考える方法のひとつとして、専門家に対して、その領域のメンタルモデルを外在化して、象徴的に表すように依頼しているということがある。実際的なオントロジーの開発では、専門家たちは慎重に選ばれており、ソフトウェア、たとえば Knowledge Mapper を使って(Chung & Baker, 1997 CRESST)、オントロジーを個別に構築するよう求められた。そうすることにより、オプションのプルダウンメニューを使用するノードとリンクの構造を作ることができる。多くの専門家が、オントロジーをそれぞれ開発する場合、かれらは、合意する部分とそうでない部分を示すのに図表を用いて重ね合わせるだろう。それから、典型的な過程としては、専門家によって直接議論され、説明され、和解し合うことが必要となり、それはコンセンサスが整うまで繰り返される。さらに、Chung, Niemi と Bewley (2003)や Iseli (2011)による最近の研究によると、オントロジーの開発が始められ、その領域に関連するドキュメントの分析を伴う過程の中で拡張することができる。参照文献、記事、そしてその他の文書の鍵となるアイデアを自動的に引き出すことは、専門家が一致させる予定である材料に含まれており、ネットワーク表現における自然言語処理を使用して達成することができる。内容オントロジーは、幼稚園から中等教育までの算数・数学(Iseli, 2011; Iseli, Koenig, Lee, & Wainess, 2010)、歴史(Phelan, Dai, Valderrama, & Herman, 2011)、生物科学(Phelan, Dai, Valderrama, & Herman, 2011)、言語科目(Phelan, Dai, Valderrama, & Herman, 2011)のために開発中である。

21 世紀スキルとオントロジーの融合

理想的なケースでは、21 世紀スキルのオントロジーは、学習とアセスメントをガイドする一体化構造をつくるために、内容オントロジーと融合されなければならない。CRESST では、問題解決(S.Mayer, 2010)、コミュニケーション(Phelan, Dai, Valderrama, & Herman, 2011)、状況認識(Koenig, Lee, Iseli, & Wainess, 2009)、そしてチームワーク(O'Neil, Wang, Lee, Mulkey, & Baker, 2003)の領域で、21 世紀スキルのためのオントロジーを構築することについて、いくらかの進捗があった。これらのオントロジーはそれぞれ、プロセス領域の理論的、経験的な分析に基づいている。たとえば、状況認識スキルにおいて、Endsley(1995)らの研究は、不可欠であり、そして、Salas and Cannon-Bowers (2001)と、O'Neil ら(O'Neil, Chuang, & Chung, 2003; O'Neil, Chung, & Brown, 1997; O'Neil, Wang, Chung, & Herl, 2000)によるチームワークの領域の研究、フレームワークの研究、経験的研究も重要であった。S.Mayer の問題解決オントロジーにおいては、全国的に認識されている認知心理学者が専門家として採用されており、その内容と構造を洗練するために修正を行うものについてオントロジーの構築が求めら

れている。それは、わたしたちの意図するところであるが、知的で社会的なスキルのオントロジーが開発され、その目的はアセスメントであり、シミュレーション、そしてゲームデザイン(Chung, Delacruz, & Bewley, 2004; Chung, Niemi, & Bewley, 2003; Koenig, Lee, Iseli, and Wainess, 2009)、射撃のトレーニング(Chung, Delacruz, Dionne, & Bewley, 2003)や戦術的な意志決定(Bewley, Lee, Jones, & Cai, 印刷中)のために、内容領域オントロジーと結合されるプロセスを文書化し続けている。

しかしながら、予測できない未来においては、「すべての可能性がある」、もしくはすでに知られている内容を表す方法として、オントロジーを使用するアセスメントのアプローチは、相容れないと思われるかもしれない。認知レディネスが特定化される予測の不可能さは、もともと内容の詳細にはそなわっていない。しかしながら、非日常的でありあまり使われない内容が「予期しないこと」という定義をもつと仮定されるかもしれない。予測の不可能さは、個人の置かれる非日常的なまれな状況にたいして、より関連がある場合がある。そこでは、しっかりと学習された知識構造の利点は、状況に対する迅速な検索にあり、そして状況による必要条件に対する関連性を試している。

これが達成、存続およびエラーの回避につながるプロセスであるという証拠はあまりないが、予測できない環境において機敏さを求められるような知識の構成要素を決定するために得ることのできた明確な研究への道筋がある。

非日常的状況における内容および認知レディネスに組み込まれる 21 世紀スキルのモデルに基づいた学習とアセスメント

21 世紀スキル、認知レディネスおよび内容オントロジーを使用するデザインは、アセスメントや学習環境、そしてシステムの中で行われるだろう。ここでアセスメントに注目してみたい。わたしたちは、アセスメントの発達のためのモデルを開発し、そしてその歴史は 1992 年以降進歩している(Baker, 2007c, Baker, Chung, & Delacruz 印刷中, Baker, Freeman, & Clayton, 1991)。アセスメントのモデルは、内容の仕様書からつくるのではない。いつもはそれが通例だが、まずオントロジーメジャーの内容領域に組み込まれる 21 世紀スキルの選択から始めていきたい。関連する 21 世紀スキルについて明示的に包含される第 1 の関心の理由は、詳細に説明することができる。第一に、スタンダードまたは方針について言葉の記述によって意図されるような方法において、処理の知的な深さが含まれると保証される。もし明確に作らないなら、多くのテストではそれを容易に作り出せるので、認知的または繰り返しの手続きを強調していることがわかる。第二に、21 世紀の操作的な定義は、どの関連する内容が知的スキルの領域の高い重要度を与えられなければならないかという決定を可能にする。第三に、特定の知的スキルは、測定の質を最適化するためにアセスメントのフォーマットを示す。たとえば、適応可能な問題解決では、解答者にオリジナルの答えを出すように要求する。そして問題認識の方法として、なぜ問題ステートメントの選択が行われたのかについての言葉の説明が行われ、問題についての最良のステートメントまたは表現である選ばれた答えを組み合わせるかどうか。明らかな 21 世紀スキルの認知モデルは、一つ以上のひな形(テンプレート)を生成し、あるいは、アセスメントを構築のために組み合わせるときに使われるモジュラーオブジェクトのセットの生成を促している。このモジュールの特徴は、コンピュータベースのオーサーズシステムを使用して、部分的に自動化したアセスメントのデザインを期待する。次に、アセスメントのための一般的なモデルタスクの使用は、領域の範囲内での一貫したサ

ンプリングの可能性を増加させるだろう。生徒のパフォーマンスの理解に雑音を加えるような構成概念に無関係なタスクの特徴(Messick, 1989)や項目タイプは、特定され、減らされるだろう。この関心は、期待された適切な結果を測定することであり、明らかに研究結果の信頼性に影響を及ぼし、介入または経験の機能として実際の変化を見つけることにつながる。さらには、ガイダンスについての3つのソースがあるので、モデルの使用は、測定後、低コストで拡張する開発を許すだろう。つまり、21世紀スキルまたは認知的要求、関連した内容オントロジー、そしてアセスメントタスクモデル、の3つである。この経済的な有用性は過小評価されるべきではない。つまり、テンプレートは再利用することができ、そして、異なる状況、内容、反応を挿入することができる。それは、成長についての長期的な解釈を支援するだろう。自由な解答の場合には、スコアリングの判定基準やルーブリックは、再利用することもできコストを大幅に下げる。さらに、ルーブリックが、オントロジーと21世紀スキルへ高水準で結びつくものであるならば、特にコンピューターによる自動化したスコアリング(Chung & Baker)の進歩が進んだならば、教師や生徒は、学習に対してより透明性を持った受容者となるだろう。さらに言えば、21世紀スキルと内容オントロジーとが調和し統合されたデザインは、単にタスクと判定基準のプールを開発するよりももっと普及力のある目的を持っている。スキルとオントロジーの二重表現は、データベースデザインの中核であり、それは適切なアセスメントで生徒のパフォーマンスのために保管場所として役立つように意図されるものである。学習の流れにアセスメントが配列されるという点では、データベース構築は、解答の正解、不正解の多様なパターンが蓄積されてくるにつれて、変化するだろう。スキルとオントロジーは、データベースのための最初のメタタグとして用いられる。

モデルベースのアセスメントについてのまとめ

スキルおよび内容を用いてタスクが作られるとき、別々にあるいは分野を複合した方法で扱われる必要があるいくつかの問題がある。測定されるべき一貫した領域の詳述と認識に加えて、アセスメントデザインが促されるには、モデル、テンプレートと他の構成要素を使用することが必要である。また、低価格であるが現アイテムフォーマットに比べてまだ洗練されているとは言えない段階であるアセスメントタスクを提供するために、他の構成要素を使用することが必要かもしれない。そのような結果はこれらの構成要素によってもたらされる。たとえば、テンプレートであるが、これは後の試験で再利用することができる。たとえ違う科目でも新しいテストに際して再び開発する必要性を省いている。実例を示すと、あるテンプレートは、作業例(Sweller, 2003)を含んでいるかもしれないが、知識の小さなビットというよりもむしろパターンやスキーマの検索に役立っている。別のルーチンでは、力強いアプローチにより、解答者にその原理に基づいた説明をさせ、「なぜ」それを選択したのか、解答したのかについての説明をさせている。このアプローチは、直接の手続きの応用から推論されるかもしれない深い理解があることを証明することを保証する。三番目として、スタンダードベースの領域で、または、尺度に基づいた専門領域を測るよう意図されたアセスメントでは、手順はまずパフォーマンスがスキルとオントロジーの要求にきちんと沿うような信頼性があり、正確な方法であるかどうか標記することに置かれる。次に、獲得したスコアを評価するために手続きを開発する必要がある。入学についての設定値としては、これらの手続きは生徒をランク付けるか、正規分布に変換できる標準得点を作成することも伴うだろう。また別の方法はエキスパートとビギナーの間の比較である。

テクノロジーの挑戦

テクノロジーの発達にともなって、信頼性と正確さを決定するために認められた手続きだけでなく、現在のデザインプロセスに対してもいくつかの挑戦が浮き彫りになっている。テクノロジーのタスクにおいては、シミュレーションの中で、たとえば、各アセスメントはシナリオに基づいているかもしれないし、また相当に長いタスクであるかもしれない。多くの項目を持つかわりに、テクノロジーの設定では、ほとんどの心理測定的な分析をするための基本原理、ただ長いだけの、相互依存的タスクないし両方が用いられているようだ。そのようなタスクの質を確立するための新しい方法を開発する更なる研究が必要である。

妥当性の最小範囲

妥当性と技術的な質の他の関連した事例に対するアプローチは明らかでなければならない。妥当性は導き出された目的である。妥当性はアセスメントの目的とデータとを結びつける一連の推論であり、アセスメントがもたらした判断の質に関しての付加的な推論でもある (Missick, 1989)。異なる目的、データのタイプによって関連した推論タイプは変化する。データの主だった使用が正式なアセスメントならば、例として設計者は、シミュレーションやゲームがどのように進行すべきかを、学習者が前もって行ったパフォーマンスに基づいて決めていけるようにするため、興味のあるデータは細分化され、誤解や逡巡、エラーに注目することで、さらにシステムの改善の必要性につながるほうがよいし、おそらく単純に別のやり方を個々の学生に提供するためでもある。もしデータが認証やアカウントビリティに必要なスキルセットを獲得したという証明のために使われるなら、信頼できるデータとは、それらのスキルが学生の身についているかを表すかどうか、また個人が名目上目的としたゴールに到達したかどうかを判断できる必要があるだろう。入学試験が目的だった場合、アセスメントは予測を立てることに使われ、予測機能としての質と、予測された基準値の質との両方が、経験的な詳細な調査を受けなければならない。

加えて、アカウントビリティのための公の報告なども興味深い。たとえば、学校をその効果や影響力という意味でカテゴライズすると、学習者についての情報、生徒と教師、グループ、移動性の量とインストラクションのタイプ、さらに異なるレベルでの満足度の獲得までも、妥当性を推論するために考察されるかもしれない。しかしながら、すべてのケースにおいてデータ管理システムは、結果を設計者、教師や政策担当者たちに伝える必要がある。内容的、状況的、言語的に複雑さのような他の潜在的な変数と同様に、スキル領域に沿った進歩にタグを付けることができるように、それらは組織化される必要がある。このデータベースの機能は、オントロロジーが広範にわたって使用されることを再度示唆し、オントロロジーは単に図式化した役割を担うだけでなく、進行中のデータ管理システムと報告のデザインを導くことを想起する。いずれの場合でも、一連のプロセスは質の向上に使われる。このプロセスとは(1) 専門家によるターゲットオントロロジーの調整チェック (2) 思考表出法プロトコルによって、望ましい学習プロセスがタスク事例の中に適応されているかを判定 (3) 実際のパフォーマンスのクリティカルパスを特定 (4) 信頼性、ディメンショナリティ、公平性の心理測定研究 (5) ターゲットとしている

人たちに使いやすいこと；教師、生徒、採用者など（6）アセスメントの広範囲な適用前に妥当性の論拠を確立すること

変化

日本、韓国、台湾、香港、シンガポール、およびアジアとヨーロッパなど多くの先進国の大部分の地域の出生率の低下は著しく、人口維持に必要な出生率を下回っている。宗教的な理由で子どもをつくる国、あるいはアフリカの貧困地域を別にして、高等教育を利用できる学生の数は変化している。子どもが少ない家族は学費により多く投資することが可能となり、したがって、出生率の低下は、高等教育を受ける優秀な生徒の数が減るという意味にはならないらしい。しかしながら、最高の教育機関を除いて、高等教育に進学を希望する生徒の質は落ちる可能性はあるだろう。この将来観については2つの予測が立てられる。一つ目は、厳格な入学試験は今よりもはるかに縮小された機関で行われる（もっと志願者も縮小される）。二つ目は、生徒たちが費用的な問題や利便さといった点で、レベルの低い教育を受けようとするだろう。たとえば、今日のマス市場と一致しているオンライン機関があり、これまで望んでいた機関への入学を求めなくなった。

ちなみに、アメリカの出生率は、女性一人につき子どもが二人であり、現在の人口を「リプレイス」している。それにもかかわらず、歴史的背景に貧しい生徒の大半を高等教育レベルに持っていくことの教育のむずかしさがあるために、出生率の低い国々と同じような現象が起きている。したがって、教育的な戦略において何らかの劇的な進展がない限り、高等教育に適応できる高い能力を持つ生徒の数は同様に減少するだろう。

そのような状況の中で、これまでは生徒たちを入試によって選り分けていた。そのようなシステムの他にどのような方法があるというのか。2つほど実行可能に思える方法がある。

一つ目は、国家レベルで提供される入学試験からプレースメントテストに変換する、ということである。このテストは生徒を一方では学校へマッチングさせるのに使われる、また、どのエントリーレベルコースが受験者に最適かを見分ける際にも使われることになる。二つ目に、高等教育に携わるインストラクターや教授陣は、これまできちんと準備しやる気のある学生を教えることで恩恵を得てきたが、これからは、クラス運営を学び、それほど優秀でない学生相手にも伝わるような教授法を採用しなければならない。これらの方法に危険性があるとしたら、高等教育における全体的な教育の質が、劇的に低下する可能性があるということである。何が起きるかということ、エリート高等教育機関が他国からの学生を惹きつけるようになり、さらに文脈を拡大すれば、とりわけ評判の良い国々は、トップ層の学生を自国の大学に呼び寄せることができるだろう。知的にも文化的にもそれほど均一でない学生を相手するには、より多様な順応性が求められるだろう。

結論

それでは、入学試験へおける不確実性の影響は何である場合があるだろうか。最初に、わたしたちは選抜よりもアチーブメントであると主張したい。そしてプレースメントや資格のような目的が現在のモ

デルにとって代わる必要があるだろう。二番目に、学習とは継続的な過程であり、テストで成功したからと言って終了するものではないということである。21世紀スキルは未来のためにとっても大切な内容のベースを形作るだろう。このようなスキルは新しい文脈や、新しいアプリケーション、新しい職務資格が提示される中で学び直されなければならない。アセスメントや試験のプロセスは状況ごとにおきかえられ、実用的に使われることが可能になり、その状況は今では個々の組織をまたいで起こりうる。

新しい文脈において、試験は個人を最適なプログラムにマッチングさせ、埋もれた才能の芽を見つけ援助し、その人材がゆくゆくは多様化した組織のミッションに沿うように試験を行うようになっている。その選択は、型にはまらないアセスメントの活用を決定し試験的に試すためにあり、学生の選択と柔軟性、教育機関のコントロールの両方をバランスよく保つためのものでもある。試験の結果から得た発見は分析され報告されるだろう。その分析は、さらなるニーズを正確に示すためにデータ収集および現代の解析論の新しい技術を用いて、異なって報告されるかもしれない。

まとめ

これまで予測可能であったものはなくなりつつある。それは、キャリアや内容がより短いスパンで急激に変化し続けているからである。経済や仕事の環境の変化のために、これからは学校そして人生を通しての学習に注目しなければならない。この点において、21世紀スキル、内容オントロジー、そしてアセスメントデザインの新しい方法は、今現在起きている予測不可能な要求に応じた重要な学習を支える一つの大きな柱となるだろう。

本章では21世紀スキルと認知レディネスについての展望を提供し、そして不確かさが増えていく状況に直面しながらも、わたしたちの見通しに対する関心について説明する。場面は3つの方法に分けられる。(1) 知的認知スキル (2) 社会的スキル (3) 個人内スキル である。選抜またはこれらのスキルのトレーニングに関しての論評は、「認知的レディネス」という専門用語の分析、また認知的レディネスは予測できない要求に立ち向かうために明らかに転移とレディネスの詳細について対処するという意図であるということである。21世紀スキルに関しては、内容領域の役割が扱われており、そして内容の詳細を表すための方法は開発された。内容を表すための図表によるアプローチは、オントロジーの概念、あるいは内容、関係性および構造の象徴として定義され記述された。オントロジーに関しては、アセスメントのサンプリングと学習計画にとって重要なスキルを特定するための方法として、活用、経験、研究分野について議論されている。21世紀スキルのためのオントロジーにおいて必要なものについての簡潔な認識は、最近の仕事に関するトピックによって記述された。21世紀スキルと内容オントロジーをベースとしたアセスメントを作り出すためのアプローチ、「モデルベースのアセスメント」(Baker, 1997b, 2007a) は、高度なタスクを作り出し、コストを抑えたり、測定の技術的な質を向上させたりする中で、その有用性という点において述べられている。妥当性に関する手短な議論の中では、構成概念に無関係な相違を最小限に抑え、目的に関連づけた正しい推論を導き出す鍵となる考えを含んでいた。また、オントロジーのデータベース開発の拡張に関する議論が示された。それは、学習の明らかな側面と個人の関係のより複雑な体系である。つまり、スキル、内容、言語学、アセスメントフォーマット、経験そして発達曲線である。このストラテジーは、学生の経験についての自動化が系統的に遂行されるにつれて関連性を持ちそうである。学習とアセスメントのデザインの力学はすべて、最終的には、試験

において実際に能力を発揮する学生の数と質に依存する。出生率の変動、そしてグローバル化した世界での新しい可能性に直面しながら、試験制度の将来が変わることは確実である、ただし、変化の中にありつつも、制度は強固なまま残っている。



New Approaches to Measuring 21st Learning: Assessment in a Changing World

Eva L. Baker

CRESST National Center for Research on Evaluation, Standards, & Student Testing
 University of California, Los Angeles
 International Symposium of Organization for the Study of College Admissions
 National Center for University Entrance Examinations
 Tokyo
 18 November 2011

NCUEE-OSCA 1

© Regents of the University of California

Overview



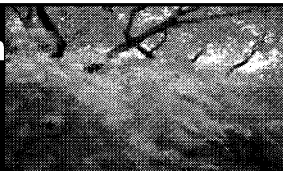
- Introduction
- The changing world and 21st century skills
- Measuring 21st century skills with academic & practical content in K-12 using four steps
- Finding the way forward



NCUEE-OSCA 2

© Regents of the University of California

Why Is the Season Ripe to Rethink Assessment?



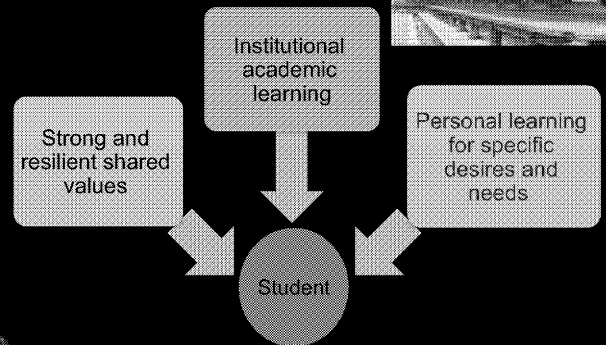
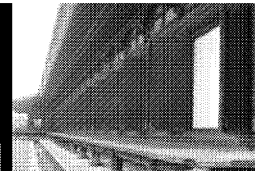
- Advent of 21st century skills linked to global standards for university and workplace
- Proliferation of technology and learners' expectations and expertise in use of technology
- New strategies for design, monitoring, and validating innovative tests



NCUEE-OSCA 3

© Regents of the University of California

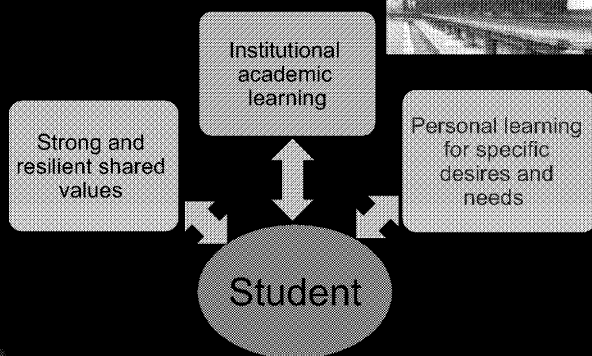
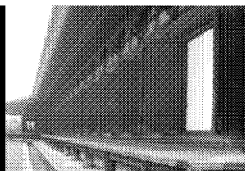
Goals for Assessment



NCUEE-OSCA 4

© Regents of the University of California

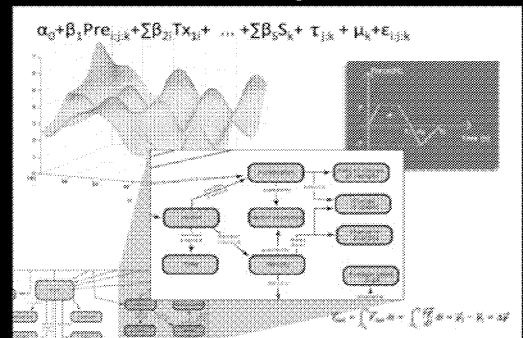
New Goals for Assessment



NCUEE-OSCA 5

© Regents of the University of California

Towards a New Paradigm of Assessment Design



NCUEE-OSCA 6

© Regents of the University of California

Traditional Purposes and Uses of Assessment:

STUDENTS

- Admissions
- Placement
- Communication
- Motivation
- Diagnosis
- Feedback
- Improvement
- Certification

INSTITUTIONS

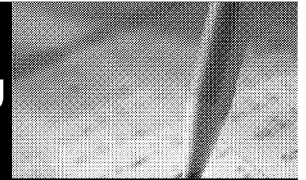
- Status
- Comparisons
- Improvement
- Personnel decisions
- Sanctions & rewards
- Public and policy estimates of quality



NCUEE-OSCA 7

© Regents of the University of California

Many Think of Tests as Differing in Format Only



- Problem sets
- Multiple-choice
- Paper-based
- Computer-administered assessments
- Projects, research studies, portfolios



NCUEE-OSCA 8

© Regents of the University of California

Some Innovations in Testing

- Computer-adaptive assessments
- Simulations & games integrating separate skills and challenging problems
- Automated assessment development and scoring
- External qualifications, badges and other verified performance



NCUEE-OSCA 9

© Regents of the University of California

Diving Deep Below the Surface



TO LEARNING!

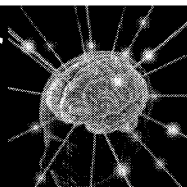
- *21st century skills*
- *Content*
- *Criteria to judge quality of performance*
- *A range of situations*
- *Tasks of increasing demand and complexity*
- *TRANSFER—new situations, new combinations of skills and content*
- *Disincentives to memorize*



NCUEE-OSCA 10

© Regents of the University of California

Assessment Design for 21st Century



Starts with key elements of performance—what and how students learn and what are a range of desirable attributes, accomplishments and skills

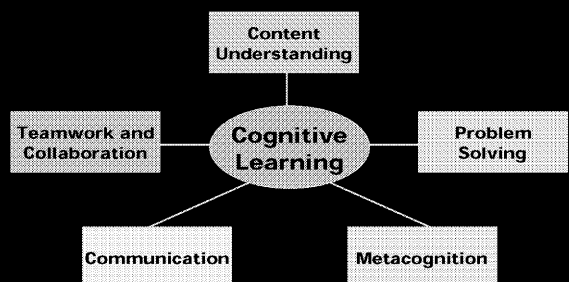
1. *21st Century Skills or cognitive demands*
2. *Content domain represented in an ontology*
3. *Blend skills and content together*
4. *Incorporate in tasks and tests*



NCUEE-OSCA 11

© Regents of the University of California

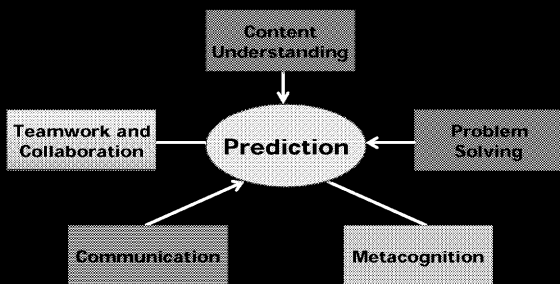
1: A Simplified Cognitive Model for Assessment Design



NCUEE-OSCA 12

© Regents of the University of California

1. Simplified Cognitive Model for Assessment Design—Admissions



NCUEE-OSCA 13

© Regents of the University of California

1. Measures of 21st Century Skills: Cognitive Traits or Achievement ?

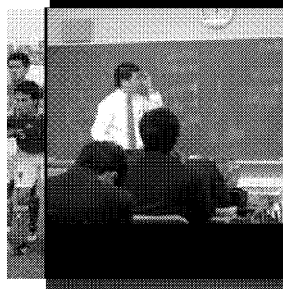
- Aptitude
 - Reasoning, creativity
- Learned
 - Key problem solving skills in domains
- Interaction between aptitude & learned skills
 - Ability to create new solutions, overcoming unexpected barriers in a range of new situations

NCUEE-OSCA 14

© Regents of the University of California

1. Expanded 21st Century Skills: Cognitive, Social, Intrapersonal

- Adaptive, complex problem solving
- Situation awareness and risk assessment
- Decision making
- Self-regulation
- Teamwork
- Learning to learn
- Communication
- Conceptual, procedural, and systemic learning of content



NCUEE-OSCA 15

© Regents of the University of California

1. Why Focus 21st C. Skill First?

- Ensure that important and challenging thinking skills in learning and assessment
- Reduce cost of expanded or subsequent test development
- Improve validity
- Communicate goals to secondary schools
- Help students ultimately to achieve transfer to a new domain

NCUEE-OSCA 16

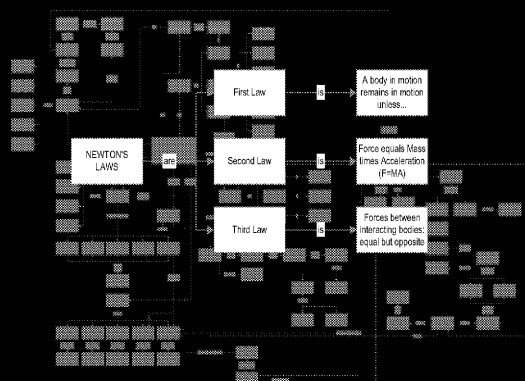
© Regents of the University of California

2: Content-Ontology-Based Architecture

- A content **ontology** represents graphically domains or standards and their components
 - Created by experts and expert materials
 - Use links, nodes in a domain(s) and their definitions commonly in a network
 - Depict relationships among the links and nodes, including direction and importance
 - Define a relational database of tasks, student performance, and subsequent achievement
 - Data allows CHANGES in content or relationships among tasks to be monitored and reflected
 - Guide redesign in a transparent way—evolving concepts of content

NCUEE-OSCA 17

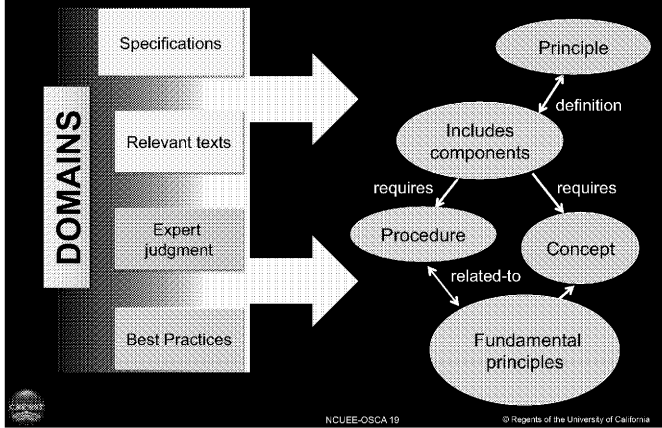
2. Physics Ontology



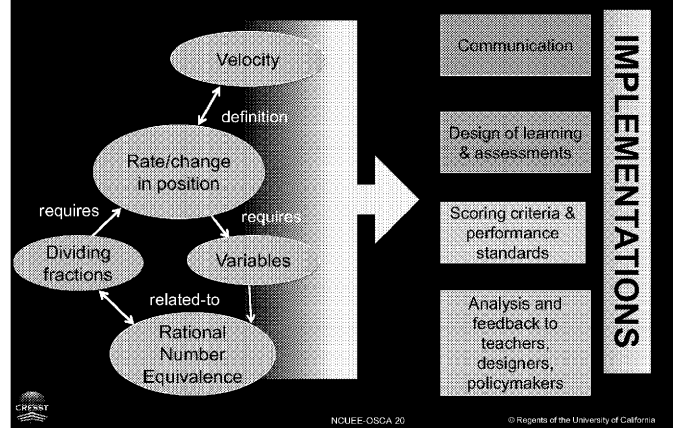
NCUEE-OSCA 18

© Regents of the University of California

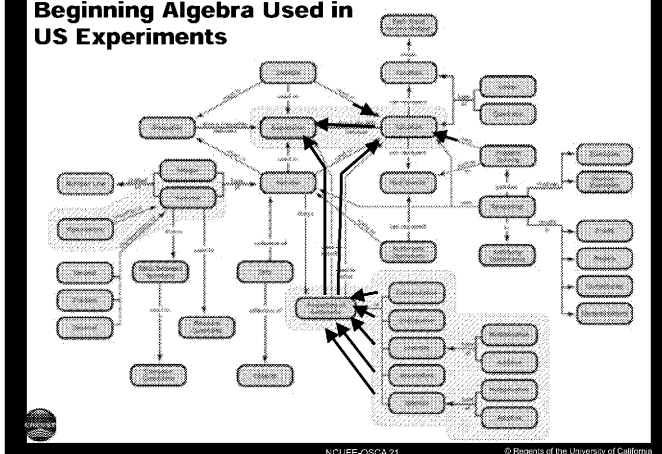
2. Ontology Design



2. Ontology Use



2. Content Ontology of Beginning Algebra Used in US Experiments



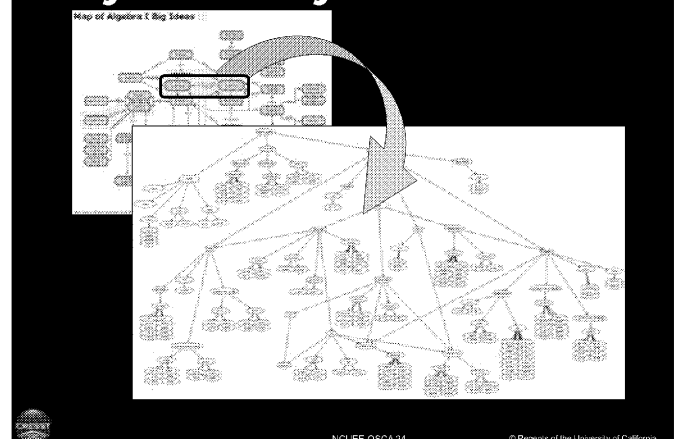
2. Uses of a Content Ontology

- Clarify content domain and relationships
 - First-cut at content validity
 - Guide design and review of learning and assessment
 - Modify by data, experts, computer natural language extraction or combination
 - Subject to evidence-based improvement
- NCUEE-OSCA 22 © Regents of the University of California

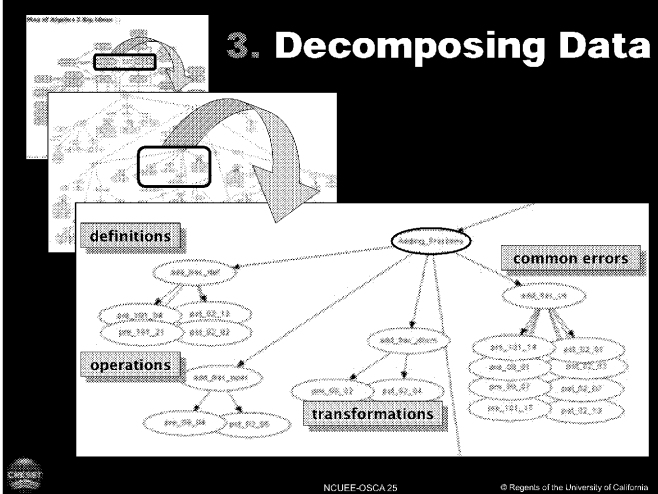
3: Blending cognitive-ontology architecture

- The ontology is an integrative frame - a level of a learning database
 - *21st Century Skills: problem-solving, reasoning, communication*
 - *Knowledge: declarative/conceptual/procedural/factual for content and age ranges of choice*
 - *Alternative formats and technical quality data*
 - *Affective behaviors*
 - *Alternative learning sequences, schemata*
 - *Situations for learning & transfer*
 - *Data accessed by separate or combined metatags*
- NCUEE-OSCA 23 © Regents of the University of California

3. Dynamic Bayesian Network



3. Decomposing Data



NCUEE-OSCA 25

© Regents of the University of California

3. Analysis of Performance

- Redefines ontology
- Redefines relationships within a content domain
- Changes content and 21st c. skill relationships
- Ontology adapts as new data patterns emerge

NCUEE-OSCA 26

4: Model-Based Task Design

- Sample from a defined universe of tasks and items based on skills, content, and on ways of measuring: models, templates, and scoring guides
- Reusable components result in time and cost savings, quality improvement
- Growing evidence base of technical quality
- Designed by computer or person, for either paper or computer
- Use for range of testing purposes

NCUEE-OSCA 27

© Regents of the University of California

4. Templates or Models

- Drawn from measurement experience and learning research
 - *Partially worked examples*
 - Focus on difficulty in procedural or problem solving (Sweller, Mayer)
 - Support the development of schema, key to expert performance, reduce working memory load
 - *Explanation of principles*
 - Why, not what, choices have been made
 - Supports metacognitive and review

NCUEE-OSCA 28

© Regents of the University of California

4. Models and Templates

- Contain empirically verified scoring criteria for essay or open problems, wrong answers/distractors, computer algorithms for scoring, weighting, and adapting learning sequence of tasks
- Mix of tasks, some long, some short should be included to reflect different difficulty of skills to be learned

NCUEE-OSCA 29

© Regents of the University of California

4. Current Technology Challenges

- Longer tasks or scenario-based assessments integrated
 - *Design and analysis rules to assess comparability or equivalence (fairness & reliability)*
 - *Sampling*
 - *Memorability*
 - *Automated-scoring focused on meaning*
 - *Hacking*

NCUEE-OSCA 30

© Regents of the University of California

4. Validity Minimums

Validation of measures

- *Expert review of alignment to the target ontologies*
- *Think-aloud protocols to determine whether expected learner processes are being applied in task examples*
- *Identification of critical paths of actual performance*
- *Psychometric studies of reliability, dimensionality and fairness (total scores and diagnostic subscales)*
- *Usable by targets: teachers, students, administrators*
- *Established before wide application*



NCUEE-OSCA 31

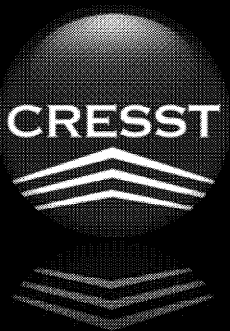
© Regents of the University of California

Summary: Future is Uncertain

- What has been predictable may not be
- Careers and content are exponentially changing
- New focus is on learning, in school and throughout life
- 21st century skills, content ontologies, and new methods of assessment design support learning of unpredictable requirements

NCUEE-OSCA 32

<http://www.cse.ucla.edu>



Eva L. Baker

email baker@cse.ucla.edu

NCUEE-OSCA 33

© Regents of the University of California

我が国の初中等教育政策と大学入試

銭谷眞美

(東京国立博物館長、元文部科学事務次官)

皆様、こんにちは。ただいまご紹介いただきました東京国立博物館長の銭谷でございます。

きょうは、教育行政を経験した立場から、我が国の大学入試についてお話をさせていただきたく存じます。

概要は37ページに記載をしておりますので、ご参照いただければと思います。

大きく2つのお話をしたいと思っております。1つは、近年の我が国の大学入試の変遷と現状をどういうふうにとらえるかということでございます。第2は、初等中等教育の立場から見た大学入試をめぐる今後の課題といったようなことでお話をさせていただきたいと思っております。

限られた時間でございますが、どうぞよろしくお願いを申し上げます。

さて、第1点の我が国の大学入試の変遷と現状についてであります。

ご案内のように我が国は、1868年の明治維新によりまして、国の近代化を進め、約140年前の明治5年、1872年に学制の発布を行いまして、近代的な学校教育制度を構想いたしました。

実は私が現在勤務をいたしております東京国立博物館も、同じ1872年、明治5年に創設をされております。来年が140周年ということになります。いろいろ楽しい催しを用意しておりますので、お暇がございましたら、ぜひ上野の東京国立博物館にお越しをいただければと思います。せっかくの機会ですのでPRをさせていただきました。

さて、この学制というのは、「邑に不学の戸なく、家に不学の人なし」という言葉をキャッチフレーズにしたわけでございます。教育の機会均等ということを目指に、学校の整備を進めてまいりました。

我が国では、学校教育が急速に普及いたしまして、出身の階級、階層、門閥にかかわらず、どの学校を出たかという、いわゆる学歴を重視する風潮が生まれまして、受験競争も激しくなってきたわけでございます。

戦前は、義務教育は小学校の6年まででしたので、受験競争は、まず非義務制の最初の学校である旧制中学校受験という形であられました。次いで、その後続きます一高、二高、三高といった旧制高等学校の受験が激しくなったわけでありまして。大学のほうは、旧制高校を出ていれば、おおむね比較的容易に入ることができたという実情だったように聞いております。

旧制高校の受験が大変激しかったというのは、夏目漱石の弟子としても知られております久米正雄が大正7年、1918年に書きました小説「受験生の手記」という小説の中にも書かれております。主人公は、一高を受験するわけですがけれども、失敗をし、さらに失恋もして、自殺をしてしまうという物語ですが、その中には当時の旧制高等学校の受験競争の厳しさというのがよく書かれております。

1945年、昭和20年日本は敗戦を迎えます。さまざまな社会改革が行われ、その中で学校教育も大きく改革されました。1947年、昭和22年新たに3年制の中学校ができて、小学校、中学校合わせた9年間で義務教育ということになりました。旧制中学は新制高校に移行いたします。それから、旧制高校は大学に移行しまして、旧制大学と一緒に新制大学ということになったわけです。したがって、現在の高等学校は、講学上は、高等教育機関ではなくて、中等教育機関という位置づけがなされております。

いずれにしても、戦後この6・3・3・4の単線型の学校制度が導入をされたわけでありまして。したがって、受験競争はまず最初の非義務制学校である高等学校受験をめぐる生じたわけでありまして。

新制高校が発足いたしました昭和22年、当時の高等学校の進学率は40%でした。それが20年後の昭和40年代、1965年ごろになりますと、80%の進学率に達しておりました。さらに10年後の昭和50年、1975年ごろには90%を超える進学率になったわけでありまして。

特に高校受験が非常に社会的な大きな問題になりましたのは、ベビーブーマー、戦後の昭和二十一、二年から二十四、五年ぐらいの間に生まれた世代、実は私もそのひとりですが、このベビーブーマーが高校進学に差しかかった昭和三十五、六年から昭和40年ごろ、1960年ごろから1965年ごろ、このとき高校の受験競争が大変大きな社会問題になりました。当時の合い言葉は、「15の春を泣かせるな」という言葉でした。この15歳の春を泣かせないように、高等学校の新設、増設が相次いで行われたわけでありまして。

また、有名進学高等学校への集中を避けるために、学校群制度、総合選抜制度、いわゆる数校まとめて、その入学定員までを合格させて、各学校に割り振るという学校群制度、総合選抜制度も導入をされたわけでありまして。

高校に少しおくれまして、大学受験も大きな社会問題となりました。これもやはりベビーブーマーが大学進学を迎えます昭和40年代、1965年ごろから大変大きな社会問題になったわけでございます。大学受験対策に偏重した高校教育の弊害、あるいは大学浪人の大量出現、予備校の隆盛、予備校なしでは大学合格は難しいという時代になりました。さらに入試における難問、奇問の出題ということが見られるようになったわけでありまして。

ご記憶の方もいらっしゃるかもしれませんが、フォークソング歌手の高石知也が歌いました「受験

生ブルース」、これは大ヒットした歌ですけれども、これがはやりましたのは、ちょうど昭和43年、1968年のことでした。その翌年、昭和44年には、大学紛争が高じまして、東京大学の入試の中止と
いった事態にもなったわけでありす。

このころ、よくアメリカやヨーロッパから教育使節団、教育調査団が来日しておりました。この欧米からの教育調査団の報告書には、私が記憶している限りでは、「日本は人生の代理戦争を18歳の少年・少女に行わせている」とありました。日本では18歳でどの大学に入ったかで、その後の人生が決まるといったような趣旨のレポートがあったと
思っております。18歳の大学受験というのは、大変大きな社会問題になったわけでありす。

いずれにいたしましても、今申し上げました戦後の高等学校入試、大学入試を考えるとときには、幾つかのポイントがあろうかと思ひます。

1つは、受験に当たって、それまでの学習、下級の学校の教育の成果や下級の学校教育への影響というものをどのように考えるかということでありす。例えば調査書を重視いたしまして、大学受験であれば、高等学校教育の成果というものを評価する
という方向にいくのか、いわゆる学力調査による一発入試を重視するの
かという大きな対立線があったやに思ひます。

これもご年配の方はご記憶があろうかと思ひますけれども、例えばある時期まで高等学校の入学試験というのは、中学校教育に配慮いたしまして、中学校で学ぶ9教科、音楽、美術、体育を含めて9教科全部についてペーパーテストを行った
といったようなこともありました。つまり、音楽、美術、体育、技術家庭、といった芸能、体育、技能的教科を入試科目から外しますと、高校受験のためにそういう教科の学習がおろそかになるということから、9教科すべてをペーパー試験にする
といったようなこともかつてはあったわけ
でございます。それまでの学習状況をどう評価するの
かというの
が一つの入試の考え方としてあろうかと思ひます。

2つ目は、受験者のどこを評価するの
かという問題です。つまり、知識量を評価する、それまでの学習内容に則した学力調査でいくのか、あるいは論文とか面接ということ
で、人間性といったようなものも含めて評価をするのか。あるいは潜在能力といひましようか、素質といひましようか、適性検査、能力判定といったようなことを中心に評価をするのか。受験者のどこを評価するの
かということも、入試を考えるとときには一つのポイントとしてあろうかと思ひ
しております。

さらに、3点目としては、資格試験的に考えていくのか、あるいは選抜重視でいくのかということ
があります。ある時期から日本の高等学校は、受験者すべてを収容するに足るだけの規模を持つよう
になりました。もう受験をしなくても、例えばある地域で100人の志願者がいれば、高校は十分すべての子供を受け入れることができる。ただ、それでもやはり試験をやりま
した。それは高等学校教育を受けるに足る能力がない、適格性を欠くということになったら、合格をさせない、定員内であって

も合格させないといったような考え方もありました。いわゆる資格試験的な方向というのもあったし、単に定員をオーバーした分の選抜だけではないという考え方もありました。ただ、高等学校では、比較的早くから、できるだけ定員までは入れようというのが動きであったと思います。

このほか、受験機会を1回にするのか、2回にするのかなど、さまざまな考え方があろうと思えますけれども、入試をめぐるまは、ある意味ではエンドレスの改革が戦後行われてきたということが言えるかと思えます。

さて、昭和46年、1971年、こういった状況を見て、中央教育審議会は、その答申の中で、大学入試について言及いたしました。いわゆる四六答申と呼ばれる有名な中央教育審議会の答申です。この中で大学入学者選抜制度が我が国の学校教育全般に及ぼす重大な影響にかんがみて、高等学校の学習成果が公正に評価され、選抜に合格することだけを目的とした、特別な学習をしないでも、能力、適性に応じて大学に入学できるようにするという方針が打ち出されて、調査書を重視し、広域的な共通テストを開発することの必要性がうたわれております。

この1971年、昭和46年の中央教育審議会の答申を受けまして、1977年、昭和52年に、国立大学の共同利用の機関として大学入試センターが設置をされ、1979年、昭和54年、全国の国公立大学と大学入試センターが共同して行う第一次共通学力試験、いわゆる全国的な共通学力試験というものが実施をされることになりました。この共通一次試験というのは、良質な出題が評価される一方で、5教科、7科目を一律に課したことによりまして、大学の序列化と偏差値輪切りを生んだといったような批判もございまして、1985年、昭和60年に臨時教育審議会が新たな共通テストの創設を提案いたしまして、平成2年、1990年に国立、公立だけでなく、私立大学、つまり各大学と大学入試センターが共同して行う大学入試センター試験というものが実施をされることになりました。この平成2年の大学入試センター試験が今日まで続いているわけであります。

大学入試センター試験の概要は、入学志願者の高等学校段階における基礎的な学習の達成の程度を判定するということでもあります。高等学校における基礎的な学習の達成の程度を判定するということが主たる目的といたしております。

その結果、この入試センター試験と個別試験との適切な組み合わせによる入試の個性化、多様化が進展しまして、国公立大学ではなくて、私立大学を含めた入試の改革が進んできておりますし、良問を提出することによりまして、難問、奇問が減少したと言われております。

現在の大学入試センター試験の状況ですが、これまで22回実施をされ、平成18年度からは英語のリスニングテストも導入され、これまで6回行われております。この春の実績では、志願者が55万8,984人、利用大学が665大学、163短期大学が参加しているという状況です。約20年間、日本の大学入試は、基本的には大学入試センター試験プラス個別試験という形で推移をしてきたわけであります。

ただ、この20年の大学の状況を見ますと、随分やはり変わってきているわけであります。大学入試センター試験が始まりました平成2年から2年後の平成4年、これが実は18歳人口の第2のピークでございましたが、18歳人口は205万人おりました。平成22年、おおむね20年たったこの平成22年は、205万人が120万人に減っております。大学、短期大学の志願者に対するいわゆる収容力、入学定員の受け入れ可能人数は、平成4年ごろは60%でしたけれども、平成22年では92%になっています。大学を選ばないと、ほとんど入れる状態になっているということであります。大学、短大の志願率も、平成4年の50%から、平成22年は60%に上昇しております。

また、大学入学者の現役・浪人の別を見ますと、4年制大学では、平成4年当時は、現役が3に対して浪人が2という状況でしたが、現在では、現役5に対して浪人は1という状況に変わってきております。

入試の方法も、いわゆる学力試験によります一般選抜、一般入試、これが中心ではあるわけですが、私立大学では、約半数がAO入試、推薦入試を経由して大学に入った学生ということになっております。

ちなみに、私立大学の細かい数字を申し上げますと、一般入試で大学に入った学生は、全体の入学者の48%にすぎません。推薦が40%、AO入試が10%という状態で、今、大学入試といっても、私立大学の場合は、いわゆる一般入試よりは、推薦、AO入試が中心になっているという状況が見られるわけであります。

さらに、私立大学では、全大学の約4割が定員割れを起こしておりまして、定員割れの大学ほど、志願倍率が低く、推薦、AO入試の実施率が高く、選抜が多様化しているという実態があります。

ただし、私立大学は、大学によって入試状況のばらつきが大変大きくなっておりまして、志願倍率は、2倍にいかない大学がかなりあるわけですが、一方で9倍以上の大学の割合も非常に高くなっています。これに対して国立大学は、大体3倍、4倍のところにとどまっているということで、私立大学は、大学によって入試状況のばらつきが大変大きいという結果が出ております。

以上、整理をいたしますと、AO入試、推薦入試の実施規模が拡大をし、入試の多様化が進む一方で、学力を問わない選抜になっているという指摘があります。つまり、大学入試の選抜機能が低下し、入試によって大学入学者の学力水準の担保ということが非常に難しくなっており、大学入試の存在が、進学希望者である高校生の学習意欲を喚起することにはなっていないという指摘もあるわけがあります。大学入試には、選抜という機能があるわけですが、それが働いていない。また、一方で、入試には副次的な効果として、高校生の学力保障ということに貢献するという側面もあるわけですが、それも働いていないという指摘が今なされております。

なお、高校生卒業生、つまり、大学受験者の格差の拡大という問題もあります。例えば、入試セン

ター試験におきましては、上位校の大学では、ほとんどセンター試験で差がつかないという指摘もされております。一方、下位校の大学では、センター試験自体が難し過ぎる。その結果、下位校の大学でも本当に選抜に役に立つのかという指摘もあります。したがってセンター試験も各教科・科目1種類でいいのかという意見も一部にあるわけであります。

次に、2つ目の初等中等教育の側から見た大学入試について、私なりの考え方をお話をさせていただきたいと思えます。

我が国では、大学入学資格は、ほかの試験を通過したとか、別の試験を一定点数とれば、入学資格が与えられるということではなくて、高等学校卒業をもって大学入学資格が与えられるというシステムになっております。いわゆるバカロレアをとるといったシステムではないわけであります。高校卒業イコール大学入学資格を持つというのが基本的な考え方でございます。

したがって、高等学校の側から見ますと、高等学校卒業ということを大変重視をするわけであります。もちろん高校を卒業していない人が大学入学資格を得る道としては、「高等学校卒業認定試験」という国が行っている試験があるわけですが、それは数としてはまだ少ないわけであります。高等学校としては、3年間、全人教育を行うことによりまして、高等学校を卒業させる。そのことがイコール大学入学資格を得るということになるわけであります。ですから、高校生活というものを高等学校は大変重視いたします。私は当然のことだと思います。通常、高等学校の卒業認定は、学校長が教育課程の履修、習得の状況を見て判断をするわけであります。教育課程といいますと、知育、徳育、体育、そういうものをすべて含んで高等学校3年間の全人的な教育の成果を学校長が判断する。それがイコール大学入学資格になるというシステムであります。

特に、いわゆる教科の面について言いますと、高校の教育課程の基準がありまして、我が国では、「学習指導要領」という形で示されております。この学習指導要領は、ほぼ10年に一度改訂をされておりまして、新しい高等学校の指導要領も既に改訂をされて、再来年、平成25年度から実施することになっております。高等学校側としては、この新しい指導要領に沿った大学入試の問題であってほしいと考えるわけであります。

もちろん高校の学習指導要領自体も、高校生の現状にかんがみまして、質的にはかなり幅があるものですし、必修だけでなく、選択的な内容科目も多くなっているわけであります。したがって、教育課程を編成、実施する高校にとりましても、高校生の学力の担保、保障ということは、大きな課題であります。高校卒業資格、高校を卒業したということは、どういう学力を身につけた人なのかということについての保障措置ということは、これは高校にとっても大きな課題であります。国語、数学、英語、理科、地歴科、公民科、いろいろな教科がございますけれども、こういう各教科の学力をどのように評価をして保障するようにすればいいのか、どのような測定の仕方があるのか、これは高校に

とつても大きな課題ですし、大学にとつても大きな課題だと思います。

今研究が進んでおります、いわゆる高大接続テストも、そういった観点からさらなる検討が必要かなと思っております。高大接続テストについては、大学入試に依存しないで、高等学校段階での学力を客観的に把握をするということを目的として、今いろいろな検討が行われているわけですが、どのような教科科目構成とするか。問題の難易度をどうするのか。だれがやるのか。受験回数はどうするのか。評価は1点刻みにするのか。Aランク、Bランクといったようなランクにするのか。課題はまだまだありまして、もう少し検討が必要かなと思っております。いずれにしても、高校生の学力保障、これはイコール大学入学者の学力保障につながるわけですが、どういう仕掛けがいいのか、これは私ども教育行政に関係する者にとっては、一つの大きな課題であると認識をいたしております。

もう一つ、初等中等教育のサイドから見た課題として、「新しい学力観」に立った教育、これを大学入試でどう評価していただけるのかという課題があります。初等中等教育では、平成元年の学習指導要領、平成元年ですから、もう22年前になりますけれども、平成元年の学習指導要領から新しい学力観に立つ教育ということを提唱してまいりました。この新しい学力観に立つ学力というのは、学力というのは、単なる知識、技能のみならず、学ぼうとする意欲や態度、あるいは思考力、判断力、表現力、応用力、活用力、こういったものを総括して学力と考えたらどうだろうかという考え方があります。この考え方は、平成10年の学習指導要領、そして小学校では平成23年、中学校では平成24年から実施をされます平成20年改訂の今回の学習指導要領でも、引継がれて踏襲されております。現在では、この「新しい学力観」は、豊かな心情：道徳性や健康な体力を育てることまでも含めて「生きる力」ということで総称されております。

平成19年の学校教育法の改正によりまして、第30条に学校においては、基礎的な知識及び技能を習得させるとともに、これらを活用して課題を解決するために必要な思考力、判断力、表現力、その他の能力をはぐくみ、主体的に学習に取り組む態度を養うことに特に意を用いなければならない旨の規定も設けまして、この新しい学力観に立つ教育を推進しているところであります。

子供たちの学力をどうはかるかというのは、大変大きな問題であります。従来、日本は、小学生、中学生の学力は、国際的な学力調査では、I E A（国際教育到達度評価学会）が実施をしておりますティムス（T I M S S）の調査で測定をいたしておりました。この調査には日本は昭和30年代後半から参加をしておりました。理科と算数、数学について、小学校4年生と中学校2年生を対象に実施をする調査でして、このティムスの調査では、日本は、常に上位に位置してございまして、現在でも、かつてよりは少し下がっておりますが、なお上位に位置をいたしております。

一方、2000年からは、O E C Dが実施をいたしますピサ（P I S A）の調査にも参加いたしました。いわゆるリテラシーをはかるピサの調査では、日本は第1回目は大変すばらしい成績をおさめま

した。数学的リテラシーは、32カ国中第1位、科学的リテラシーも、32カ国中第2位でした。読解力は若干低かったんですが、32カ国中第8位。そのときは、OECDのピサ調査は、余り話題にはなりませんでした。

ところが、2003年の調査、これは2005年に発表になったわけですがけれども、これが大変悪い結果でございました。数学リテラシーが41カ国中6位、かつて、3年前は1位だったものが6位。科学的リテラシー、これは41カ国中2位で、順位に変わりはありませんでした。読解力が大変問題でして、2000年が8位だったのが、2003年の調査では、41カ国中14位。このときは日本は中位、上位ではなく中位ということになりまして、我が国に「ピサショック」をもたらしました。日本の子供たちの学力は大丈夫か。特にピサは、「新しい学力観」がねらっておりました思考力、判断力、表現力、応用力といったものを調べる調査でしたので、愕然としたわけでありました。

成績のよかったフィンランド、2000年に成績が悪くてショックを受けたドイツ、あるいは全国的な学力調査を展開していたイギリスなどに手分けをして、私どもは調査に参りました。

その年、平成17年、2005年に、私どもは2つの方針を決めました。1つは、全国的な学力調査を日本もやろうということでありました。もう一つは、学習指導要領の改訂に着手しようということでありました。

この全国的な学力調査は、悉皆、全小学生、全中学生を対象にやろうということによって準備に入りました。

学習指導要領の改訂につきましては、特に日本が弱かった言語能力、これを重視した改訂をしよう。それから、今の日本における子供たちの状況を考えて、体験活動、こういったものを重視する改訂をしようということによって準備に入りました。指導要領については、平成20年に改訂をして、小学校はことしから実施をされているということによってございます。

学力調査につきましては、平成19年から実施をすることにいたしました。この2007年、平成19年に40年ぶりに全国学力量学習状況調査を行ったわけでありました。対象は、小学校6年生の全員、中学校3年の全員。教科は、国語と算数、数学ということにいたしました。問題は2種類。1種類、これをAと呼びますと、A問題はIEA、ティムスの調査に則したような、いわゆる知識、技能というものはかる問題にする。もう1種類、B問題は、ピサ(PISA)に則した問題にする。全体として、そういう調査にしようということにいたしました。

かつて40年前の全国学力調査では、都道府県別の結果は公表しませんでした。今回は、都道府県というのは、教職員人事を担当しておりますので、その県内の先生は大体同じように採用され、研修を受けて人事を行われておりますので、都道府県別の結果を公表するということにいたしました。

19年度、20年度、21年度、と3年続けて、悉皆調査をやりましたが、秋田県が、いつやってもト

ップのほうにくるといので、大変不思議がられております。吉本理事長も私も秋田県の出身です。それはさておきまして、多分、学力調査の結果の背景には先生の指導力の問題、あるいは家庭環境の問題、地域と学校の関係の問題、あるいは少人数学級の問題、いろいろな要素が背景にあると思っております。

この結果、今、日本の子供たちは、I E Aの調査で小学校4年生、中学校2年生が理科、数学を調査を受け、ピサの調査で高校1年生が受け、文部科学省の全国学力学習状況調査で小学校6年生と中学校3年生が受けているという状況になっております。

こういったピサ型の新しい学力観に立つ学力をどう大学入試で評価をするか。大学入試というのが、知識のテストのままでいいのか、偏差値の高い秀才を育てる教育のままでいいのか。このことへの検討が必要ではないかということでもあります。

3点目は、きょうの韓国のご発表にもございましたが、グローバル人材を育てるという観点から、英語のテストの在り方も含め大学入試の改善、充実の方策の検討が要るのではないかと思います。今年政府のグローバル人材の育成の調査会でも、そういった方向が出ておりますが、グローバルな人材を育てる観点からの入試のあり方の検討ということも、課題だと思っております。

いろいろ申し上げましたが、今、大学入試自体が、かつてのような大きな社会問題になっているということではございません。しかし、問題ははらんでいるわけでありまして。高校と大学、双方の教育保障、質保障に入試、あるいは高校、大学の接続がどういう貢献ができるのか。逆に言いますと、大学がそれぞれ個性を発揮し、オンリーワンの大学、受験生が行きたい大学、こういう大学を目指すということが、実は大学入試の問題を考えるときの大切なポイントではないかと私は思います。

あの大学で学んだ人はきっとこんな人と、こういうイメージができる大学づくりが必要になるのかなと思っております。

今、私がおります東京国立博物館には、さまざまな大学出身の方が働いてくださっております。また、キャンパスメンバーズになっております大学生の方がたくさんお越しをいただいております。また、大学生の方がインターンシップでもよくお見えになります。よく教育をされている大学の学生は気持ちがいいなとよく話しに出ます。このことを大学の先生方にお伝えをいたしまして、私の話を終わりにさせていただきます。

ご清聴いただきましてありがとうございます。

第2部
指 定 討 論

Discussion: Validity Issues for Next Generation NUEE

Joan L. Herman

National Center for Research on Evaluation, Standards and Student Testing
(CRESST)

UCLA Graduate School of Education and Information Studies

November 18, 2011

The conference presenters have shared new tests and products they're developing and raised some intriguing approaches and possibilities for University Entrance exams and particularly the National Center's Test:

- Dr. Kyung-Ae Jin has shared Korean Institute for Curriculum and Instruction's (KICE) plans and progress in developing an innovation new National English Ability Test (NEAT). NEAT recognizes the need to go deeper in assessing students' ability to apply their language skills in a variety of contexts and to incorporate performance assessment to assess students' speaking and writing skills. NEAT also is very innovative recognizing the different levels of competence students may need, Academic level II, or practical level III, depending on students future plans; in its use of technology and cloud computing, and in its score reporting.
- Dr. Deborah Harris has shared the rich portfolio of products and services that ACT offers to support students achieve success in education and the workplace. She described a series of products that backward chain from college readiness based on the ACT admission test to assessments and benchmarks to monitor and support students progress from elementary through middle and high school. She also described innovative ways in which ACT data are being used to benchmark students' proficiencies nationally and internationally and to support program evaluation and policy development.
- Dr. Eva Baker has outlined fundamental new ideas for a new generation of learning- and ontology based assessments of 21st century skills. She makes a strong case not only for the skills beyond traditional academic knowledge that students will need for future success, and lays out designs that start with specific types of 21st century skills and then embed academic content within them. She also raises important questions about the potential functions and future of admissions tests such as the NUEE.
- Dr. Masami Zeniya has provided the rich history of high school and university admissions testing system. Dr. Zeniya also has underscored the functions that different kinds of tests – for example international comparisons – play in education policy and noted the limitations of the current admissions tests. I believe his presentation too can call for the incorporation of 21st century skills into the NUEE and for leverage the test to support the development of a broader set of student capabilities.

Together, these presentations raise many rich ideas for creating a next generation NUEE

and create a rich menu of possibilities for subsequent National Universities Entrance Examinations. In discussing how NCUEE might evaluate the utility of these ideas and suggestions various ideas, I'd like to raise some very basic issues for consideration.

My discussion will address three major points:

1. *The meaning of validity*, the term the measurement community uses to characterize the quality of an assessment: Validity is the ultimate measure by which any new approach to the National Center Test needs to be judged.
2. *The motivation for changing NUEE*: Why change NUEE? In judging the adequacy of any of the approaches or ideas my colleagues have presented, I'd like to propose that NCUEE carefully consider what purpose each new approach is intended to serve. Does it address a real problem with the current admissions test? How will it improve the test, improve its use and consequences?
3. *The success criteria for any change*: How will NCUEE know whether any changes it makes are successful? I'd like to encourage NCUEE to look ahead NOW to how it will judge whether any new approach, any new instrument, any new system really works and is worth it. These questions bring us right back to issues of validity.

Validity and Consequences of NUEE

Validity, as I have mentioned, is the term that the measurement community uses to denote quality in educational assessment. Based on the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), validity is derived from a variety of sources of evidence documenting the extent to which:

- the assessment measures what it is intended to measure; and
- the assessment enables appropriate and accurate inferences for intended decision making purposes. That is, the scores well serve the purpose(s) for which the test is intended.

By this definition, an assessment itself is neither valid nor invalid, but rather validity is established for particular test purposes. An assessment may have a high degree of validity for one purpose but have little justification for another. For example, scores from admissions test may do a great job of identifying the most highly gifted students for admissions into a premier institution, but be of little use in identifying the floor for those who are minimally competent to succeed in 2-year colleges or in diagnosing and placing students to respond to their learning needs.

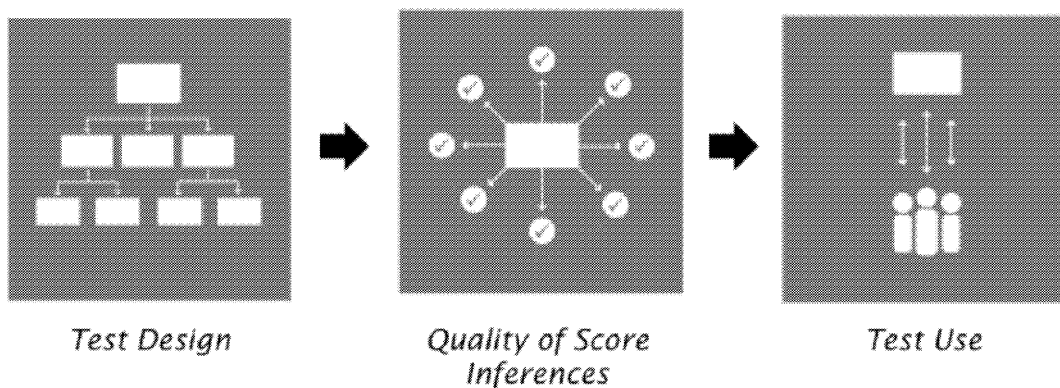
Further, in modern theory, validity is established not only through the accumulation of a variety of evidence sources, but moreover by an explicit argument that lays out and substantiate the chain of reasoning and specific evidence that justifies the use of the measure (See, for example, Kane, 2006).

Validity argument for admissions tests. In thinking about the claims and chain of reasoning supporting the use of a test, consider the building blocks for a measure serving any specific purpose (See Figure 1):

- you must design a test and the items on it so that they will likely
- yield accurate and appropriate score inferences
- that will support the decisions and uses for which the test is intended.

[INSERT figure 1 about here]

Figure 1: Validity Argument Components



Or, starting at the end point of Figure 1, you can think of the validity argument as backward chaining from:

- the inference(s) you need to draw from a test's results to enable the decision or use the test is intended to support, to
- the qualities that the scores must possess to enable the intended inferences, to
- the characteristics of the test and item design that can enable the intended score inferences.

Similarly, one can consider a validity argument as a series of if-then statements that need to be satisfied – throughout the test development and validation process – to justify the specific use of the test.

What might a general validity argument look like in the case of an admissions test? First, the test should be designed to accurately and fairly measure knowledge and skills relevant to college admissions, presumably the knowledge and skills that reflect college readiness and ability to succeed in college. The perspective is not only looking back toward how well students have learned the content of their high school courses, but also may look forward to what the capabilities and dispositions that students need for subsequent success.

Ideally, what students learn in their courses prepares them for subsequent success, but, like the speakers, I will have more to say about this below. The test must

be designed measure knowledge, skills and dispositions that have been determined relevant to college admissions and college preparation, and the validity argument requires that there be evidence that the test actually addresses these constructs. Such evidence can be collected early in the test design and development process through alignment studies and used, if necessary, to strengthen the test’s representation of relevant constructs.

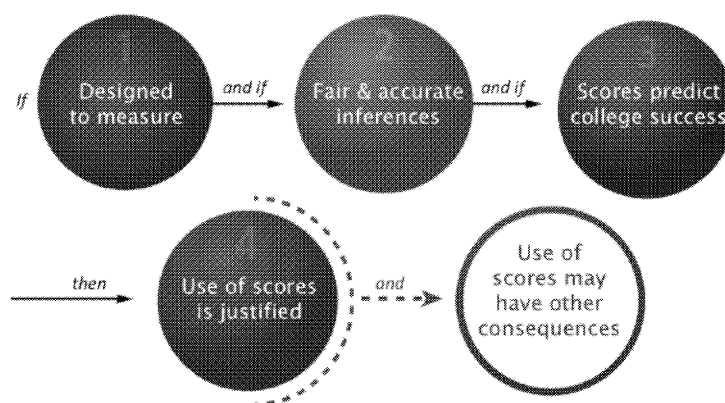
However, appropriate test design is necessary but not sufficient: student responses to the test must result in scores that are sufficiently reliable, precise, fair and appropriate to yield accurate inferences about students’ college readiness. Here standard psychometrics, differential item functioning, and relationships between convergent and divergent measures of the same construct provide evidence for such claims.

But because we not only want to measure college readiness, but to use scores for admissions decisions, students’ test performance on the test should relate to students’ subsequent college performance. The rationale for using admissions test as part of admissions decisions is that the scores provide objective data that enables colleges and universities to differentiate those who have relatively more and less merit and who are relatively more and less likely to be successful in college than those who score less well – or who score below a particular cut point. That is, the scores should predict success in college. Otherwise, why should the scores be part of admissions decisions? Relevant evidence to support these claims in the United States typically examines the relationship between admissions scores, subsequent college grades and/or college completion and examines the extent to which the admissions scores:

- Without bias, differentiate students who are likely to be more or less successful in college, in post college life, and/or
- Differentiate students who have more or less ability to succeed in particular disciplines (e.g., mathematics, physics, arts)

(see, for example, Bridgeman, McCamley-Jenkins & Ervin, 2000; Noble & Sawyer, 2002; Sackett et al., 2009)

[INSERT FIGURE 2 about here]



So in summary, the general claims in this scenario are:

- General claim 1: The test is designed to measure intended constructs for its intended use –in this case preparation for college (and perhaps life?) success.
- General claim 2: Scores on the test yield fair and accurate inferences about students' college preparedness.
- Claim #3: Scores are associated with success in college and/or predict who will be more and less successful in college (and life?)

When there is evidence supporting each of these claims, then the use of the scores as part of admissions decisions is justified.

Note that the evidence that is gathered to validate and justify the use of a test also may identify weaknesses that can have implications for subsequent test refinement and use. For example, if evidence suggests that some items are biased, those items would be modified or replaced. Findings that scores over or under predict college success for individuals from particular subgroups, e.g., girls relative to boys, could influence decision rules for individuals from these groups.

Further, in addition to the use for which the test is actually designed and intended, we know from research that high visibility tests are also likely to have other consequences (See, for example, Herman & Baker, 2009). Because admissions test performance is very important to students' future opportunities, the test is likely to influence what students study and what teachers teach, as I describe further below.

Other consequences. Beyond its ostensible purpose, what other consequences does UEE likely have? Based on research from around the globe (see, for example, Herman, 2010; Hamilton et al. 2007), we can say with some certainty that: admissions tests signal to K-12 teachers and students what is important to teach and learn, and, at least in the United States, the K-12 curriculum tends to focus on what's on important tests.

Moreover, again, based on data from the US, it is not only the content of the test that gets modeled, it is the test formats too: curriculum merges into test preparation exercises, which can be very narrow. We also know that text book publishers adapt their materials to what is on important tests and that other services pop up and adapt their methods and materials to prepare students for these tests. Japan's jukus and jobikos are a case in point.

Important tests clearly motivate both teachers and students to do well – or at least motivate those who believe they can do well and are interested in college. And because of these collective consequences, changes in important tests can leverage changes in the broader educational system. Indeed our speakers today have underscored the ways in which changes in NCUEE could support productive changes in K-12 education in Japan. For example, Dr. Jin has mentioned the role that NEAT is expected to play in supporting the transition to Korea's new language development goals and curriculum.

Core Questions for Next Generation NUUE

So, that brings us to core questions for any revisions that NCUEE might want to undertake: What is the purpose of any new approach to the NUUE? What are the

primary issues or problems that the redesign is supposed to address? To address this general question, I suggest NCUEE consider evidence of how well the NUEE currently is serving its intended purposes and to the consequences it currently is having. In speculating on these issues in light of the conference theme, I'll be drawing primarily on evidence from studies of testing in the United States.

Measuring the right constructs? A first question centers on the extent to which NUEE is addressing the right constructs: is it designed to measure the knowledge, skills and dispositions that students need to acquire to be successful in college and in life? Here I draw on research in the United States conducted by Dr. David Conley.

Based on a national survey of student engagement administered in more than 725 colleges and universities and an analysis of syllabus requirements for freshman courses, Dr. Conley documented students' need for four categories of knowledge and skills to be successful in college (see Conley, 2008):

1. Academic knowledge: the foundational content knowledge and skills that students need to acquire in core subjects;
2. Cognitive strategies: the ability to formulate problems, conduct research, analyze data, see patterns, find relationships, organize and communicate findings.
3. Academic behavior: largely meta-cognitive kinds of strategies, e.g., study skills, the ability to work independently and to self monitor and respond to one's progress
4. Contextual skills, such as collaboration and team work, social skills, knowing what's required for college.

In a separate, recent survey of college professor's reactions to the US's new Common Core State Standards in English Language Arts and Mathematics, Dr. Conley examined both the areas of agreement and gaps between freshman course expectations and the Standards (Conley et al., 2011). He found that faculty generally endorsed the new standards conception of academic knowledge and cognitive strategies, but called for more attention to students' speaking and listening skills.

Dr. Conley's general findings echo points raised by each of today's presenters. For example, NEAT, as described by Dr. Jin, has been developed to provide a balanced view of students reading, writing, listening and speaking skills and to address academic language skills in the context of real life applications and communication. Dr. Harris noted that it is not only academic knowledge that explains students' college success (see also Schmitt et al. 2009) and described ACT's work to develop measures of student engagement to address additional capabilities such as metacognition, social skills and collaboration. She also shared ACT research showing the impact of these skills on students' success. Similarly, Dr. Baker and Dr. Zeniya also especially noted the range of capabilities that constitute students readiness for college and life.

NCUEE, I suggest, should consider: Are these domains that NCUEE ought to be more comprehensively addressing?

Measuring skills students need for 21st century success? Similarly the deliberations of global businesses, the analyses of economists and labor market specialists, and the accords of national and international stakeholders across the world lead to a general consensus that academic knowledge alone will not prepare students for success in the 21st century. (See, for example, ATC 21, Partnership for 21st century skills, OECD, European and Asian nations). Businesses in the United States tend to agree on the kinds of skills they are looking to fuel their global competitiveness: initiative, innovation, ability to solve complex problems, work in teams, be adaptive problem solvers. Moreover economists note the kinds of living wage jobs that are increasing versus those that are diminishing in frequency (see, for example, Levy & Murnane, 2005). Jobs that call for routine, repetitive skills are being automated, but growing job categories feature those that require abstract thinking, adaptive problem solving, teamwork and communication. That workers can no longer expect to stay with the same company or in the same job for a lifetime means as well that individuals must be both adaptable and life-long learner, able to learn quickly and efficiently.

A recent US National Research Council that I chaired on assessing 21st Century Skills (Herman & Koenig, 2011) synthesized available lists of these skills into three, major inter-related categories:

- Cognitive skills, such as adaptive problem solving, critical thinking, systems thinking, innovation
- Interpersonal skills, such as communication, collaboration, cultural understanding
- Intra-personal skills, such as metacognition, executive functioning, and motivation.

To these, I also would add ICT literacy as a basic theme that crosses the above three and continues to radically change the way we live and work.

Like the skills needed to be prepared for college, all three speakers also noted the need to address these kinds of 21st century skills. Dr. Baker was most direct in laying out the categories and proposing a test design methodology that started first with defining the nature of these types of skills and then embedded content into their assessment.

Sending the right signal to teachers and students? In addition to measuring the right constructs, a second question I would ask is whether the NUEE is sending the right signal about what is important for students to know and be able to do. As Dr. Jin described, NEAT in large part is being developed to serve this function, to strongly communicate to teachers and students Korea's new English language development curriculum and the need to incorporate writing, speaking and other real-world applications into students' on-going teaching and learning. Because the existing language test did not integrate these features, teachers and students had little motivation to move to the new curriculum. Research, in fact, consistently shows that teachers and students focus on what is tested and tend to ignore or under-emphasize curriculum content or standards that are not tested (Stecher et al., 2000), so by changing what is tested, policymakers functionally are likely to change what is taught.

What and how does the current NUEE signal what is important for students to learn? Consider:

- Does NUEE focus on important capabilities, knowledge and skills that will transfer and support students' success in college, in life?
- Does it include a balance of academic and authentic applications, of academic content, cognitive strategies and behaviors?
- Is it transparent? Do teachers and students know what is being measured so that they can prepare for it?
- Does preparing for the NUEE encourage cramming or learning deeply?

In the United States, the issue of test format is very much bound up with what a test signals. Teachers tend to believe that multiple-choice items only measure lower level cognitive skills, even though such items can be designed to address complex thinking and convergent forms of problem solving. Because of these beliefs, when multiple choice formats predominate, teachers tend to focus classroom curriculum on lower level skills and drill students on the types of knowledge they think will be on the test. Teachers and students also spend time developing and practicing test-wiseness – so-called test taking skills directed solely at discerning the most efficient and effective strategies for approaching the specific item types found on the test and at how best to guess the right answer when you do not really know the right answer. These types of exercises do not build knowledge and skills that are transferable outside of responding to particular types of tests and therefore do not reflect meaningful learning.

The point is that it is not only what is tested that is important, it is how teachers and students understand what is being tested, what teachers and students think is being tested, that influences how they prepare.

Moving Ahead to A Next Generation System

And so I move from possible problems or challenges that any new approach or addition to the current NCUEE might want address, to a consideration of the specific approaches and issues my colleagues have discussed during this symposium. I suggest NCUEE might want to evaluate their applicability through another series of questions:

First, why change the NUEE? What is the problem the change is supposed to solve? NEAT, as described by Dr. Jin, is supporting the transition to a new curriculum and incorporating new item formats and a state-of-the-art technology platform. ACT, as described by Dr. Harris, seems to be moving from its original focus on college admission testing to a developmental testing and instructional support system to help prepare middle and high school students be better prepared for college; Dr. Baker highlights the importance of incorporating 21st century skills in admissions test, calls for broader roles for the test, and describes an innovative new approach to test design. Dr. Zeniya's history of admissions testing in Japan suggests that the country has moved from an elitist approach to higher education to a situation where there may be more places in universities and colleges than there are qualified individuals to fill them. Like Dr. Baker, he too emphasizes the need to incorporate 21st century skills more strongly into NUEE and to consider broader functions for the test.

Second, given that the purpose of any change(s) is clear, what is the chain of reasoning that connects the specific changes being proposed for NCUEE to the desired outcomes? There are at least two major chains of reasoning here. The first is the claims of the validity argument that need to be substantiated to justify the use of the test for admissions and/or additional proposed purposes. As I have discussed, in constructing the major claims, it often is useful to backward chain from the specific uses and purposes for which the test is intended, to the characteristics of the scores that will support that specific use(s), to the test design features and constraints that will need to be in place to produce the needed scores.

While there are many possible uses and consequences that NCUEE might want to foster, let's me suggest some of the claims that any NCUEE innovation might want to support:

- Better measures college readiness, better predicts college success
- Better measures preparation for success in the 21st century
- Fairness for diverse subgroups
- Enables better and fairer admissions decisions
- Improves K-12 curriculum and teaching and test preparation
- Improves alignment between K-12 and University expectations
- Informs University placement and teaching
- Improves/deepens student learning
- Improves students' preparation for college and life

In addition to chain of claims that validates and justifies the use of scores from any next generation NUEE for admissions and/or other purposes, there is another chain of logic that needs to be considered. This second one is more socio-political: it is the theory of action for how and why the new test will accomplish its goals and intended consequences – and avoid unintended consequences. NEAT provides one example: if the test is intended to motivate teachers' and students' transition to the new national curriculum, what is the chain of action that will accomplish this purpose? What changes are expected in classroom curriculum and instruction and in student learning opportunities? How are teachers and students expected to change what they do? Given that these changes are expected, what needs to be in place to support this change? For example, I mentioned earlier that it is not only what a test actually tests that is important, equally important is how teachers and students perceive the test and how they understand what is being tested. This suggests the importance of communication through various channels, the availability of new materials, professional development, score reports, etc. so that teachers can and will implement the new curriculum and that students will have opportunities to learn the new curriculum and that, as a consequence, students will acquire deeper capability in academic language and practical language applications.

These two chains of logic provide the basis for both planning for and evaluating the success of any proposed change to NUEEE. The chain of logic justifying the use of the new test undergirds the validity argument, and the chain of logic laying out the theory of action provides a framework for planning and evaluating the implementation and impact of the new tests.

Summary and Conclusion

In conclusion, here is a summary of my thoughts and advice about potential new approaches to NCUEE that symposium speakers have so well presented:

- Start with the purpose of any NCUEE redesign and what are its intended consequences. This sense of purpose and consequences is what should drive any test redesign.
- In evaluating whether a new approach will serve these purposes, lay out the chain of logic that links any proposed changes or innovation to these goals, to improvements in the performance and impact of the test and to its consequences. Just laying out the claims that the innovation should satisfy will help you to see whether there is a sound pathway to success and may help you discover and design in critical characteristics and/or components that will need to be in place to succeed.
- Each link of the chain subsumes a number of specific claims that require substantiation. The collection and analysis of evidence to substantiate these claims throughout test development and validation will help you to both evaluate and improve how well any new approach is working.

In closing, I started with the concept of validity and end with it. It is not too early to begin to consider the validity arguments and the evaluation data that will justify and support any proposed change to the NUEE. NCUEE will do well to incorporate validation and evaluation hand in hand with the design and development of any innovation.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade point average from the revised and recentered SAT I: Reasoning test* (College Board Rept. No. 2000–1). New York: College Entrance Examination Board.
- Conley, D. (2008) *College Knowledge*. San Francisco, CA: Jossey-Bass
- Conley, D., Drummond, K., de Gonzalez, A., Rooseboom, J., & Stout, O. (2011). *Reaching the Goal: The Applicability and Importance of the Common Core State Standards to College and Career Readiness*. Eugene, Oregon: Educational Policy Improvement Center.
- Hamilton, L.S., Stecher, B.M., Russell, J.L., Marsh, J.A., & Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. *Research in the Sociology of Education*, 16, 31-66.
- Herman, J. & Koenig, J. (2011) Assessing 21st century skills: A workshop summary. Washington, DC: National Academy.

- Herman, J. L. (2010) Impact of assessment on classroom practice. In P. Peterson, E. Baker, Barry McGaw, (Editors), *International Encyclopedia of Education*. Oxford: Elsevier.
- Herman, J. L. & Baker, E.L. (2009) Assessment policy: Making sense of the babel. In D. Plank, G. Sykes, & B. Schneider (eds) *AERA Handbook on Education Policy Research*. New York: Routledge
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.
- Levy, F., & Murnane, R.J. (2005). *The New Division of Labor: How Computers Are Creating the Next Job Market*. Princeton, N.J.: New York: Princeton University Press; Russell Sage Foundation.
- Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D.(2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin*, *135*, 1–22.
- Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T., Quinn, A., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact of demographic status on admitted students. *Journal of Applied Psychology*, *94*, 1479-1497.
- Stecher, B. M., Barron, S. I., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms* (CSE Tech. Rep. 525). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.
- Noble, J. & Sawyer, R. (2002) Predicting different levels of academic success in college using high school GPA and ACT composite score. ACT Research Report Series 2002-4. Iowa City: ACT. http://www.act.org/research/researchers/reports/pdf/ACT_RR2002-4.pdf

次世代の NUEE にむけての妥当性問題

Joan Herman
(UCLA=CRESST 副所長)

本シンポジウムの講演者は、彼らが開発している新しい試験や開発製品について講演し、大学入試と特にナショナルセンターの試験についての興味深いアプローチと可能性を提起した。

- ・キョン・エー・ジン博士は韓国教育課程評価院（KICE）が開発中の革新的な新しい試験である全国英語能力試験（National English Ability Test : NEAT）における計画とその進捗状況について報告した。NEAT では生徒の能力を評価する際に、様々な状況で語学スキルを使用し、生徒の話す・書く技術を評価するためにパフォーマンスアセスメントを取り入れて、より深める必要性を認識している。さらに、NEAT は、生徒の将来計画に従って、生徒が必要とするコンピテンスには異なるレベルがあること、つまり学術的なレベルであるレベル 2 と実践的なレベルであるレベル 3 があることを認識しており、非常に画期的である。テクノロジーを駆使しクラウドコンピューティングを使用することや得点の報告に関しても斬新である。
- ・デボラ・ハリス博士は、教育と職場において生徒が成功を獲得するのを支援するために ACT が提供する製品と教育サービスの豊かなポートフォリオを報告した。また、ACT のアドミッションテストは小学校から中学、高校までの生徒の成長の支援、モニターのための評価とベンチマークであるが、それに基づいたカレッジレディネスからさかのぼる一連の製品を発表した。また ACT データが生徒の能力を国内でまた国際的にもベンチマークするために使われ、プログラムの評価や理念設計を支えている状況を踏まえて躍進的な方法を発表した。
- ・エバ・ベーカー博士は、学びの新しい世代にむけて、また、オントロジーに基づく 21 世紀スキルの評価にむけて、基礎となる新しいアイデアの概要を述べた。彼女は、単に従来の将来の成功に必要なとされてきた学術的知識を越えたスキルのみを論拠するのではなく、まず 21 世紀スキルという具体的なタイプのデザインを形作り、そしてその中に学術的な内容を組み込んだ。彼女はまた、NUEE（National University Entrance Examinations）のような入学試験の潜在的機能と将来に関する重要な問題を提起した。
- ・銭谷眞美博士は日本における高校と大学入試の該博な歴史を提供した。彼はまた、国際的に比較できる異なる種類のテストが教育政策へと働く作用について注目し、そして、現在の入学試験の限界について言及した。彼の発表もまた、NUEE の中に 21 世紀スキルを取り込むことや、生徒の能力についての幅広い開発を支援するような試験へ影響を与えるだろう。

加えて、これらの発表は次世代の NUEE を作成するための多くの豊富なアイデアを提起し、後の NUEE のための将来性のある豊かなメニューを生み出す。NCUEE (National Center for University Entrance Examinations) が、これらのアイデアと提案の有用性をどのように評価するのか議論するために、考察にあたって私はここでいくつかのとても基本的な問題についてとりあげたい。

私の議論は主に以下の 3 つのポイントについて触れる

1. 妥当性の意味

その用語は、測定家が評価の質を特徴づけるために使用する。

妥当性は、大学入試センター試験におけるどのような新しい方法に対しても、これは必ず判断されなければならない根源的な測定手段である。

2. NUEE 改革への動機

なぜ、NUEE を変えるのか。我々が示したアイデアや方法の妥当性を判断する際に、NCUEE は、それぞれの新しい方法が何の目的にかなうか慎重に考えることを提案している。その方法は現行の入試の実質的な問題点に対処しているのか、一体どのようにしてテストを改良するのか、どのように使用法や結果を改善するのかといった問題がある。

3. 任意の変化の成功基準

NCUEE はどのようにしてそれらがもたらす改変が成功したかどうかを知りえるのか。

私は、新しい方法を判断するには、NCUEE に対して「今」よりもっと先を見ることを伝えたい。

そうすることにより、新しい設備、システムが本当に機能し価値のあるものか見えてくるからである。これらの問いは、妥当性の問題に我々を引き戻している。

NUEE の妥当性と結果 (意義、重要性)

述べてきたように、妥当性とは、測定家が教育評価の質を示すために使用する用語である。「Standards for Educational and Psychological Testing」(AERA、APA、NCME 1999)に基づき、妥当性は以下のような範囲を文書化するさまざまな証拠から導き出される。

- ・その評価は、測りたいものと思っているものを測定する
 - ・その評価は、意図された意思決定の目的のために適切で正確な推定を可能にする。
- つまり、得点はその試験が測ろうとしている目的に役立つ。

この定義によって、評価そのものは有効でも無効でもない、むしろ、特定のテスト目的のために妥当性が形成される。評価はある目的のために高い妥当性をもつかもしいないが、ほかのものに対してはほとんど根拠をもたない。例えば、入学試験の得点は、ランクの高い高等教育機関に入るような優れた能

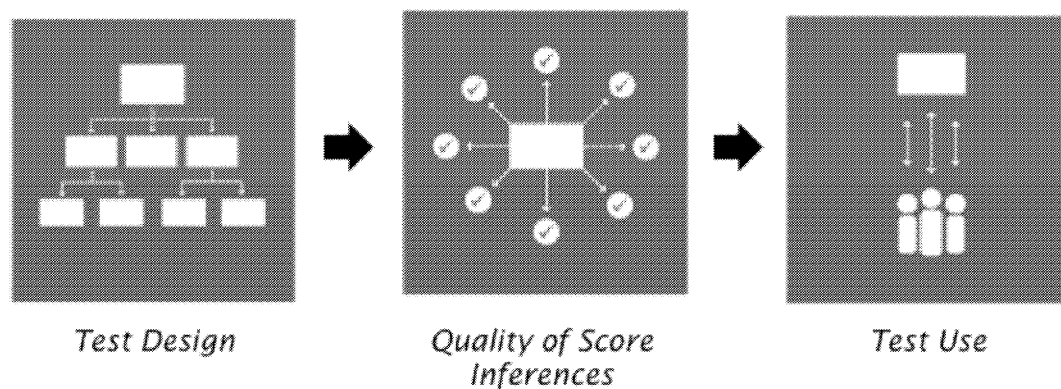
力を持つ生徒を特定する場合に大きな働きとなるかもしれないが、2年制の大学へ入学する生徒を特定することや、そのような生徒の学習ニーズに応じて各大学へ配置することにはほとんど使えないだろう。さらに、現在の議論においては、妥当性というのは、様々な証拠資料の蓄積だけではなく、その測定をしようするための理由となる証拠や論拠について立証するような明確な議論によっても確立される。(Kane,2006 参照)

入試に関する妥当性の議論

試験の使用を手助けする一連の論拠と主張について考える際には、具体的な目的に対して役に立つ測定のために構成要素に注意を払う。(図1参照)

- ・試験とその項目をデザインする
- ・適切で正確な得点推定を行う
- ・その試験が意味するものについて判断し、その使用をサポートする

Figure 1: Validity Argument Components



また図1を後ろの部分から見てみると、妥当性の議論は、以下のような逆への連鎖とみなすことができる

- ・判断を可能にするために試験の結果、あるいはサポートすることを目的と試験の使用から導く必要のある推定
- ・スコアが、意図された推論を可能にするようきちんと処理する質
- ・意図したスコアが導かれるようなテストの特徴付けと項目デザイン

同様に、妥当性の議論は、具体的な試験の使用についての正しい条件についてテスト開発と確認プロ

セスを通して満たすために必要な一連の if-then ステートメントとして考えることもできる。

一般的な妥当性の議論は、入学試験のケースの場合ほどどのように見えるのだろうか。まず、試験は大学入学に関連する知識と技術を公平に測定し、正確に設計されなければならない。おそらく、その知識と技術はカレッジレディネスと大学で成功する能力を反映させたものである。生徒が高校でどれだけよく学んできたかだけに目を向けるのではなく、今後、学生が成功にむけて必要だと考える能力と資質について見通すことが大切である。その展望は、生徒が高校コースの内容をどのように学習したかを振り返るだけではなく、また、その生徒が今後の成功のために必要とする性質や能力についての期待を寄せるかもしれない。

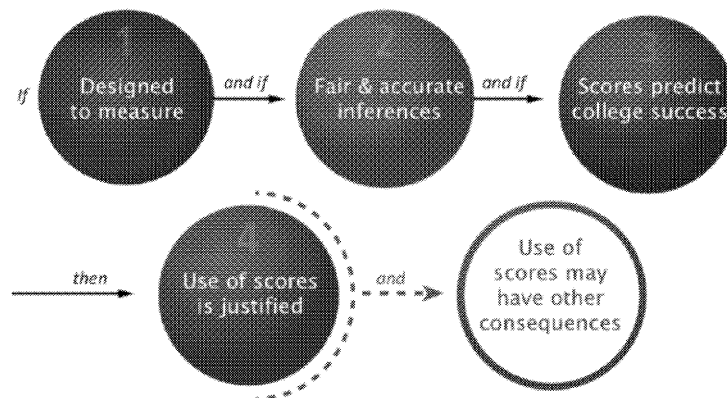
理想的には、生徒がコースで学んだことはその後の成功の準備になることだが、講演者の述べるように、私もこのことについて以下で述べる。テストは、大学入試やカレッジレディネスに関連して決定された測定知識、スキル、性質を設計しなければならず、妥当性の議論は、テストがこれらの構成概念について実際に取り掛かる証拠を必要とする。もしも関連する構成概念のテストの表現を強くする必要があるのであれば、そのような証拠は、アライメント研究を通してテスト設計・開発プロセスの早期に集められて、そして使用される。

しかし、適切なテスト設計は必要ではあるが十分ではない。生徒のテストに対する反応は得点であり、それは十分に信頼でき、正確で、かつ公平で、適切に、生徒のカレッジレディネスに関しての正確な推論をもたらすものでなければならない。ここでは標準的な心理測定、特異項目機能 (DIF)、そして同じ構成概念の収束的で互いに異なる測定の関係が、証拠をそのような主張に対して提供する。

我々は単にカレッジレディネスを測定するだけではなく、入学を決定するためにスコアを使い、テストにおける生徒のテストパフォーマンスが今後の大学生活にまで関係すると判断しているのである。入学を決定する一部として入学試験を行う論理的根拠としては、その得点が大学で成功するもしくはしないといったことについて大学が区別することのできる客観的なデータを提供していることにある。つまり、得点は、大学における成功を予測しなければならないのである。そうでなければ、なぜ得点が入学決定の一部でありえるのだろうか。アメリカではこれらの主張を支持する関連した証拠として、入試での得点と大学での成績、もしくは卒業時との関連性を調査し、入試得点は以下の範囲まで見据えているとしている。

- ・ 大学生活の中で、大学で多かれ少なかれ成功する学生を公平に区別する
- ・ 特定の分野（たとえば、数学、物理、芸術）で成功する能力がある学生を区別する

（たとえば、Bridgeman, McCamley-Jenkins & Ervin, 2000; Noble & Sawyer, 2002; Sackett et al., 2009 を参照）



つまり本論文における一般的な主張は以下の通りである

・その1

試験は、大学や（もしくは人生の）成功に対する準備が意図された構成概念を測定するようにデザインされる。

・その2

試験の得点は、生徒の大学への準備が万端かどうか、公平で正確な推測をもたらす。

・その3

得点は大学における成功と関係しており、そして、だれが成功するかしないか（大学だけでなく人生も）を予想するものである。

これらの主張について支持する証拠がある場合、入学の決定の一部に得点を使用することは正しいと判断されることになる。

テストの使用を有効にし、正当化するために集められる証拠が、その後の試験の改善や使用について影響を及ぼすような弱点を識別するかもしれないことに注目すべきである。たとえば、もし項目にバイアスがかかっていると論証されれば、それらの項目は修正されるか、入れ替えられることになるだろう。得点の上下が特定のサブグループ（たとえば男性と女性の比較など）から個人を大学での成功に近いか遠いかということを予測したとすると、その結果は、特定のサブグループからの個人のための決定基準に影響を及ぼすかもしれない。

さらに、テストが実際に意図してデザインされたもののために使用されるという点に加え、多くの注目度を持つテストはほかの結果を生み出すことが研究からわかっている（たとえば Herman & Baker, 2009 を参照）。入学テストのパフォーマンスが、生徒の将来の機会にとって大変重要であるため、以下で示すように、試験は、生徒が学ぶことと教師が教えることに影響を及ぼすだろう。

他の結果

表向きの目的をこえて、UEE がどんな結果が他にあるだろうか。世界中からの研究に基づいて、必然的に次のことが言えるだろう(Herman,2010、Hamilton et al.2007 参照)。入学試験は、K-12 (初等中等教育) の教師や生徒に対して、教え、学ぶために何が大事かを示しており、少なくともアメリカでは K-12 カリキュラムが重要度の高いテストに焦点を絞る傾向にある。

またさらに、アメリカのデータに基づき、モデル化されるのはテストの内容だけでなく、テストの形式 (フォーマット) まだがモデルとなるのである。カリキュラムはテスト対策となり、非常に狭い分野の学習となるであろう。教科書の出版社が、重要度の高いテストに対応するように教材を作り、新しいサービスを宣伝し、テスト対策のためにメソッドや教材づくりを進めていることを我々も認識している。日本の予備校や塾はまさしくその典型である。

重要度の高い試験というのは、明らかに教師と生徒にやる気を起こさせる。少なくとも、頑張ることができ、大学に興味があると思っている人たちを動機づけている。そして、これらの総体的な結果から、重要度の高いテストにおける変革は、幅広い教育システムの変化にも影響をおよぼすことができる。確かに、今日の講演者は、NCUEE の変化が日本における K-12 の教育の有効な変革を支えることができる方法について強調している。たとえば、ジン先生は、NEAT の果たすと予測される役割が、韓国の新しい語学教育向上の目標とカリキュラムへの移行を支援することであるとしている。

次世代 NUEE にむけた中心となる課題

NCUEE が試みたいあらゆる改訂について中心となる疑問について導いてみると、NUEE への新しいアプローチの目的は何か、デザインの変更で対処されるだろう主要な主張や問題は何かということである。この一般的な問題に取り組むために、私は、NCUEE に対して、現在 NUEE が意図した目的についてどれくらい役立っているのか、その結果どうなっているかということについてそのエビデンスを考慮することを提案したい。本シンポジウムのテーマに照らしてこれらの問題について思索するために、私は、主にアメリカでのテスト研究からのエビデンスに基づいて描き出していく。

適切な構成概念を測定できているか？

最初の課題は、NUEE が適切な構成概念について述べている範囲に集中している。それは、大学や人生での成功を収めるために、生徒が獲得しなければならない知識、スキル、性質を測定できるようにデザインされているのか。ここで、David Conley によって導かれたアメリカでの研究を利用する。

725 以上の大学で行われた学生のエンゲージメントに関する全国調査、および新入生のコースのための syllabus requirements を分析した結果に基づき、大学での成功のため獲得すべき知識とスキルについて 4 つのカテゴリーを実証した (Conley, 2008 参照)。

1. アカデミックな知識

学生が、主要科目で身につける必要のある基礎的な内容知識とスキル

2. 認知的方略

問題を明確に述べ練り、研究を行い、データを分析し、パターンを見て関連性を探り、組織化し、調査結果を伝える能力

3. アカデミックな振る舞い

大部分はメタ認知についての方略（たとえば、研究のスキル、独立して研究を行え、セルフモニタリングでき、進捗に反映させる能力）

4. 文脈上のスキル共同作業やチームワーク、ソーシャルスキル、そして、大学の利益のために何が必要なのか知っていること

またそれとは別に、大学教授のアメリカの新たな **Common Core State Standards in English Language Arts and Mathematics** への反応に関する最近の調査では、Conley は、新入生のコースの期待とスタンダードとのギャップ、そして一致する領域についての両方を調べた (Conley et al., 2011)。彼によると、教授陣は一般的にアカデミック知識と認知的方略についての新しい標準概念を支持するが、学生の話す、聞くスキルに多くの注意を求めているという。

Conley の全体的な調査結果は、今日の講演者によって述べられた点を繰り返している。たとえば、ジン先生によって述べられた NEAT は、生徒の読む、書く、聞く、話すスキルをバランスのよい見方を提供し、そして、実際の生活での適用とコミュニケーションの文脈の中でアカデミックな言語スキルを扱うために開発された。ハリス先生は、それが学生の大学での成功を説明するただのアカデミックな知識だけでなく、メタ認知、ソーシャルスキルおよび共同作業のような追加された能力について述べるため学生のエンゲージメントの測定を開発する仕事に ACT が取り組んでいることに注目した。さらに、これらのスキルが生徒の成功への影響を示す ACT の研究についても述べた。同様に銭谷先生とベーカー先生も、大学生活やその後の人生に向けての生徒レディネスを構成する能力の幅について特に注目した。

NCUEE は、以下のことについて考慮すべきであると提案する。NCUEE は、より包括的に扱うべきである領域に対応しているかということである。

21 世紀での成功のために生徒が必要なスキルを測れているか

同様に、世界的な企業での議論や、経済学者と労働市場の専門家の分析、および世界中の国家的および国際的なステイクホルダーは、みな一般的なコンセンサスを得ている。それは、アカデミック知識はそれだけでは学生の 21 世紀での成功への準備に十分ではないということである(たとえば ATC 21, Partnership for 21st century skills, OECD, European and Asian nations 参照)。アメリカのビジネスは、国際競争力に求められている動力の供給に同意する傾向がある。それらは、イニシアチブ、イノベーション、複雑な問題を解決する能力、チームワーク力、適応的な問題解決である。さらに、経済学者

は、頻度において減少しているものと比較して、増加しているある種の生計を立てるための仕事の種類について注目している（たとえば Levy & Murnane, 2005 参照）。型通りで繰り返しのスキルが必要とされる仕事は自動化されているが、成長が必要な仕事の 카테고리では、抽象的な考え方や、適応的な問題解決、チームワークそしてコミュニケーション力が求められるものを特色とする。成長が必要な仕事を得ている労働者が、同じ会社、あるいは同じ仕事にとどまるとは思っていない。同様に、その個人は、生涯を通して学ぶ姿勢と適応力の両方を持っているにちがいない。そうした人間は効率的に素早く学ぶことができる。

21 世紀スキルの評価について、私が議長を務めた最近の全米研究評議会（US National Research Council）（Herman & Koenig, 2011）では、相互関係のある主要なカテゴリにこれらのスキルが利用可能な3つのリストを統合している。

- ・ 認知スキル

 - 適応的な問題解決、批判的思考法、システムシンキング、イノベーションなど

- ・ 人と人におけるスキル

 - コミュニケーション、共同作業、文化的理解など

- ・ 個人内のスキル

 - メタ認知、実行機能、動機づけなど

さらに上記のスキルに、ICT のリテラシーを加えたい。ICT リテラシーは、上記のスキルと交差し、急進的に我々の生活と仕事の方法を変え続ける基本的なテーマであるからである。

3 人の講演者は全員、大学のために準備が必要とされるスキルのように、21 世紀スキルについて取り組む必要性に注目している。特にベーカー先生は、カテゴリをレイアウトし、この種のスキルの本質を定義するところからはじめて、そして内容をアセスメントの中に盛り込ませるテストデザイン方法を提案した。

教師と生徒に適切なシグナルを送れているか

適切な構成概念の測定という第一の質問に加えて、第二の質問は、NUEE が、生徒たちにとって何を知ることが重要なのか、何ができることが重要なのかということに関して適切なシグナルを送っているかどうかということである。ジン先生が述べたように、NEAT は広い範囲ではこの機能を果たすように開発されている。その機能というのは、教師と生徒に韓国の新しく開発された英語のカリキュラムを強く伝達する機能のことであり、そして書くこと話すことそして他の実生活での実用性を、学生たちの普段の教え・学びに組み入れる必要性を伝える機能を果たしている。これまでの言語試験が、これらの特性を統合していなかったために、教師と生徒は、新しいカリキュラムへ移行するための動機づけがあまりなかった。研究報告では、教師も生徒も何がテストされるかという点に重きをおき、テストに出ない

カリキュラム内容やそのスタンダードを過小評価する傾向があるということを絶えず述べている (Stecher et al,2000)。したがって、政策担当者は、何がテストされるのかという点を変えることによって、教えるものを機能的に変更するだろう。

現在の NUEE は、学生が学ぶために何がどう重要なのかを示しているのだろうか。以下のように考えてみる。

- ・ NUEE は、大学や人生における生徒の成功を支援するための重要な能力、知識およびスキルに焦点を当てているか。
- ・ それは、学術的な内容、認知的な方略および振る舞いに対して、実生活での応用と学術のバランスを含んでいるか。
- ・ 透明性はあるか。教師と生徒は、何が測られているか知ったうえでそれに備えているか。
- ・ NUEE の準備をすることが、猛烈に勉強したり深く学ぶことを促すか。

アメリカではテストの形式の議論は、そのテストが何を意図し、何を発信しているかということにとっても密接に結びついている。たとえその項目が、問題解決についての複雑な思考と思考を収束させる形として作られていても、教師は、多肢選択式の項目は、低いレベルの認知スキルを測るだけであると思う傾向にある。こうした信条のために、多肢選択式の形式が主流になると、教師は低レベルのスキルにクラスのカリキュラムを集中させ、そして生徒たちにテストに出ると思われる知識のタイプを勉強させる傾向にある。また、教師も生徒も、多くの時間をテストを攻略するための知恵を発達させ練習することに費やす。つまり、テストの特定の項目に特化した最も効率的で最も有効な解き方、そして正解を知らない場合にいかに推理し正答率を上げるか、その見分け方などのいわゆる受験テクニックに費やしている。この種の練習は、特定のタイプの試験以外の場面では、転移可能な知識やスキルを構築しにくいいため、意味のある学習とは言えない。

ポイントは、テストされることだけが重要なのではないということである。教師と生徒が、何がテストされているのかをどのように理解するかであり、どのように試験の準備をするかということに影響を及ぼすのである。

次世代システムへの移行

これまでは起こる可能性のある問題や新しいアプローチといえる挑戦、近年の NCUEE が取り組みたい課題について述べてきたが、一步進んで、本シンポジウムにおいて講演者が述べた議論や特定のアプローチについて考察していきたい。私は、NCUEE はまた別の一連の質問を通して、その適応可能性があるかどうかを評価したいのかもしれないと思う。

第一に、なぜ NUEE を変えるのか。変革によって解決されると思われる問題とは何か。NEAT は、ジン先生によって述べられたように、新しいカリキュラムへの移行を支援しており、新しい項目形式と最新技術のテクノロジープラットフォームを取り入れている。ACT は、ハリス先生によって述べられたように、大学入学試験という最初の関心から、中学生と高校生が、大学のための準備を万端にできるよう手助けするような発達のテストを行い教育サポートシステムへと移行しているように見える。ベーカー先生は、21 世紀スキルを入試に取り入れる重要性を強調しており、テストの幅広い役割を求め、テスト設計に関して新しい革新的なアプローチについて述べた。銭谷先生による日本における入試の歴史は、高等教育に対して国のエリート主義的なアプローチから、大学は様々に個人の質を高めるような場所へと移行していることが示された。ベーカー先生のように、銭谷先生もまた NUEE に 21 世紀型のスキルを取り入れる、そして試験のためにより広い機能を考慮する必要性を強調した。

次に、それぞれの変革の目的が明らかになったとすると、NCUEE に提案されている変化と期待される結果とのつながりは何だろうか。ここでの推論には、少なくとも 2 つの大きなつながりがある。一つ目は妥当性の議論である。それは、入試や加えて提案された目的のためにテストの使用を正当化するために立証される必要がある。これまで私が述べてきたように、主要な主張を構築するときは、意図されたテストの目的と使用から、その特定の使用と目的をサポートする得点の特性、またテストデザインの特長、得点を生み出す場面に必要な制限にまでさかのぼって関連を調べるのが有効である。

NCUEE が促進したいと思う多くの結果や使用についての可能性がある際に、変革についてどのようなサポートができるかを以下にいくつか示唆する。

- ・カレッジレディネスのよりよい測定と、大学での成功をよりよい予測
- ・21 世紀における成功のための準備のよりよい測定
- ・様々なサブグループにむけての公平さ
- ・入学決定をより明瞭に、より有効に
- ・K-12 カリキュラムおよび教育、テスト準備の改善
- ・K-12 と大学予測との間の調整の改善
- ・大学のプレースメントと学習の通知
- ・生徒の学習の改善と深化
- ・生徒の大学と人生への準備の改善

入学やまたそれ以外の目的があるにせよ、次世代 NUEE で算出した得点の使用を有効にし、正当化する主張の連鎖にくわえて、さらに、考察される必要がある論理の別の連鎖がある。2 番目のこれは、より社会政治学的である。それは、新しいテストがどのようにして、またなぜその目標を遂行するのか、結果を意図した行動に関する理論である。そのとき意図しない結果は回避する。NEAT は 1 つの例を提供する。もしそのテストが教師と生徒の新しいナショナルカリキュラムへの移行を動機づけるために意図される場合、その目的を達成するための行動の連鎖は何だろうか。どのような変化が、教室でのカリ

キュラムや授業において、そして機会を学習する生徒に予測されているのだろうか。教師と生徒がそれらをどのように変更すると予測されるだろうか。これらの変化が予測されるとすると、何がこの変化を支援するために必要なのか。たとえば、私が先に述べたように、あるテストが実際に何をテストするのかということも重要だが、同時に、教師と生徒がどのようにそのテストを認識し、何がテストされるのかということもどのように理解するのかということも重要である。これは、あらゆるチャンネルを通してコミュニケーションすることの重要性、新しい教材有用性、専門の開発、得点報告などを通して提案される。それゆえ教師は新しいカリキュラムを実行に移すことが可能になり、生徒たちは新しいカリキュラムを学ぶ機会が得られる。結果として、生徒は学術的な言語、そして実生活に即した言語におけるより深い能力を身につける機会を得るのである。

上記二つの論理は、NCUEE に課せられた変革の成功を計画し、評価するための基礎を提供する。新しいテストの使用を正当化している論理の連鎖は、妥当性の議論を強化し、また、活動に関する理論を展開している論理の連鎖は、新しいテストの実行および影響について評価し、そして計画するための枠組みを提供する。

要約と結論

結論として、本シンポジウムの講演者が詳しく示した NCUEE への潜在的な新しいアプローチに関する私の考えとアドバイスがここにある。

- ・ NCUEE のデザイン変更の目的とその意図された結果についてまず始めた。その目的意識と結果は、どんなテストの変更でも意図されるべきである。
- ・ 新しいアプローチがそれらの目的にかなうかどうかの評価する際には、その目的のために期待される変化と改革に関連した一連の論理を展開してみる。それがテストのパフォーマンスと影響における改善やその結果を結ぶ論理の配置となる。改革が満たされなければならないという主張をすることは、通る道が成功にあるかどうかを確かめることを助け、成功のために必要な重要な特性および構成概念を発見し、デザインすることを助けるかもしれない。
- ・ 連鎖についてのそれぞれの関連については、実証を必要とするいくつかの特定の主張を包含する。テスト開発と確認を通して得たこれらの主張を実証する証拠の収集および分析は、新しいアプローチがどのように効果があるのかを評価し改善するのに役立つだろう。

最後に、妥当性の概念について述べて終えたい。妥当性議論を考察するのに早すぎるということはなく、評価データは NCUEE に関しての期待される変革を正当化し、支持するだろう。NCUEE はいかなる改革においても、確認と評価をデザインと開発にうまく取り入れ、上手くやっていただく。

Discussion: Validity Issues in Moving Ahead

Joan L. Herman

International Symposium of Organization for the Study of College Admissions
National Center for University Entrance Examinations

Tokyo, Japan
 November 18, 2011

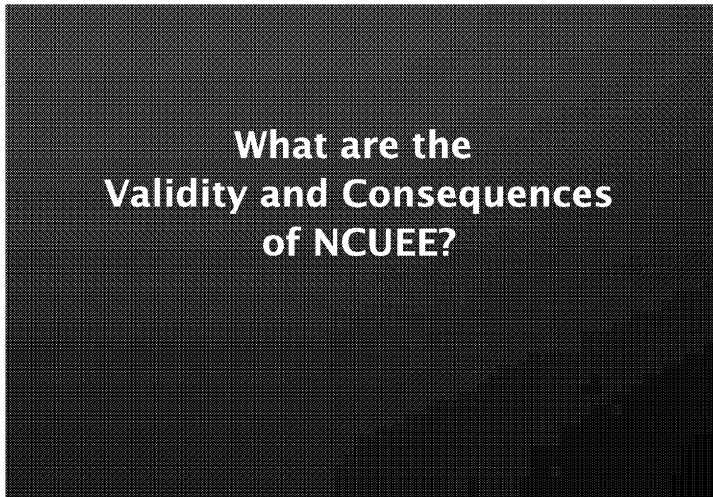


Overview

- Validity: Definition
- Why change NCUEE? What purpose(s) might a change be intended to serve?
- How will you know whether the change is successful?

2 / 19

National Center for Research on Evaluation, Standards, & Student Testing



Validity Defined

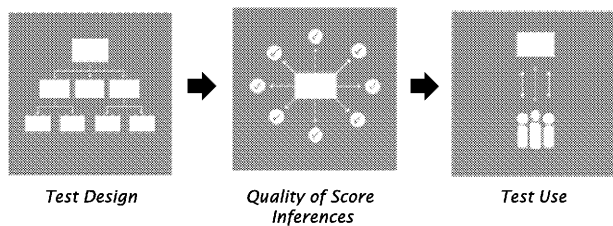
- Evidence that an assessment measures intended construct(s) *and* well-serves intended purpose
- An evidence-based argument that substantiates the chain of reasoning that links the measure to its intended purpose

4 / 19

National Center for Research on Evaluation, Standards, & Student Testing

Validity Defined cont.

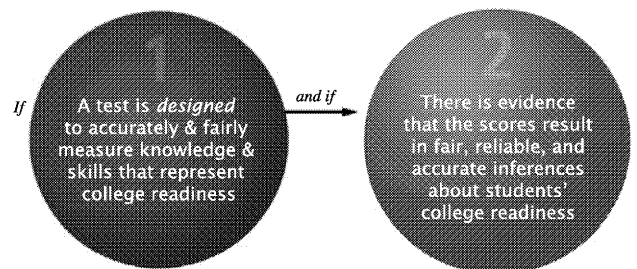
Chain of Reasoning



5 / 19

National Center for Research on Evaluation, Standards, & Student Testing

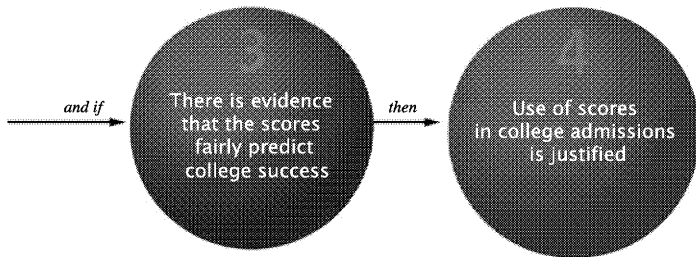
A Simplified Validity Argument: Admissions Test



6 / 19

National Center for Research on Evaluation, Standards, & Student Testing

A Simplified Validity Argument: Admissions Test cont.

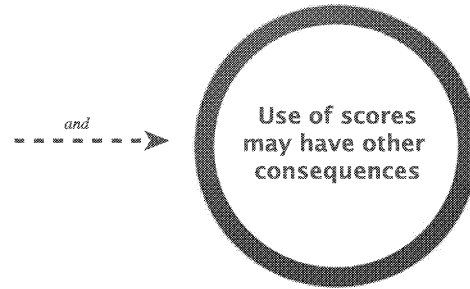


7 / 19

National Center for Research on Evaluation, Standards, & Student Testing



A Simplified Validity Argument: Admissions Test cont.

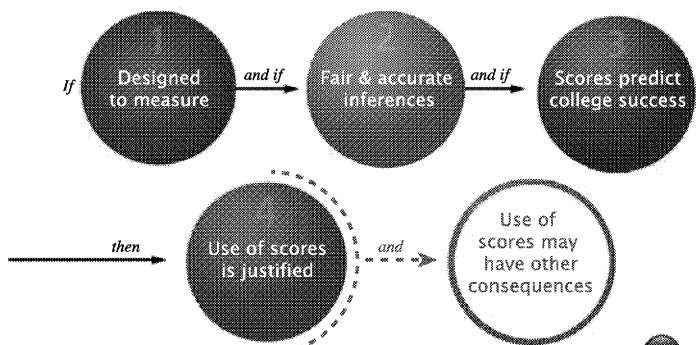


8 / 19

National Center for Research on Evaluation, Standards, & Student Testing



A Simplified Validity Argument: Admissions Test



9 / 19

National Center for Research on Evaluation, Standards, & Student Testing



What Are Other NCUEE Consequences (Intended or Not)?

- Communicates what is important in pre-college teaching and learning
- Serves as a model for:
 - ✓ *Pedagogical practice*
 - ✓ *Instructional materials and resources*
- Motivates behavior
- Changing NCUEE can leverage K-12 change

10 / 19

National Center for Research on Evaluation, Standards, & Student Testing

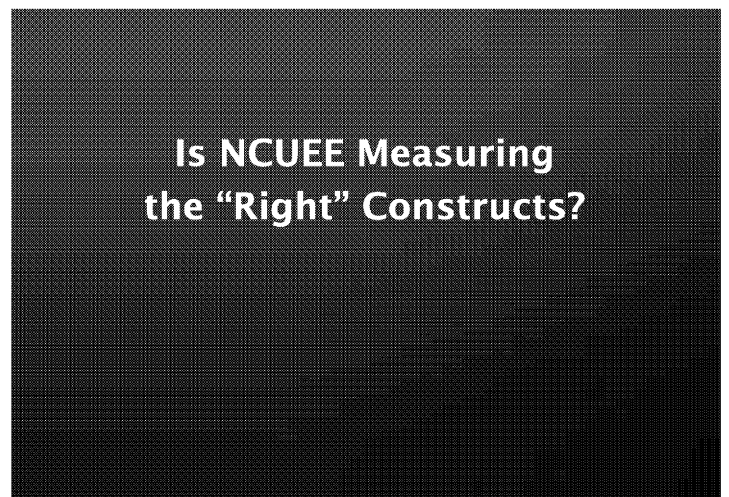


Core Questions for Any Revision

- What are the primary issues that any NCUEE redesign is supposed to address?
 - ✓ *How well is NCUEE serving its intended purpose?*
 - ✓ *What are other consequences of NCUEE?*
- Drawing on evidence from the United States testing

11 / 19

National Center for Research on Evaluation, Standards, & Student Testing



Measuring What Students Need to Be Prepared for College?

Based on Analysis of Course Requirements:

- ✓ Academic knowledge
- ✓ Cognitive strategies
- ✓ Academic behavior
- ✓ Contextual skills

Based on college professors' survey: What important skills are missing from *academic tests*?

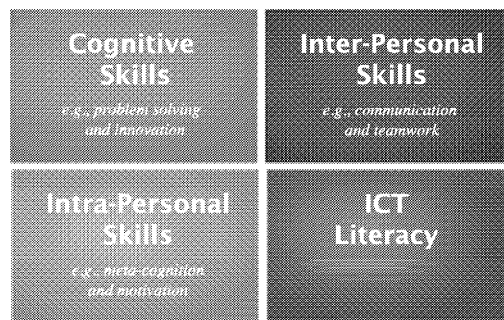
- ✓ Listening skills
- ✓ Speaking skills



13 / 19

National Center for Research on Evaluation, Standards, & Student Testing

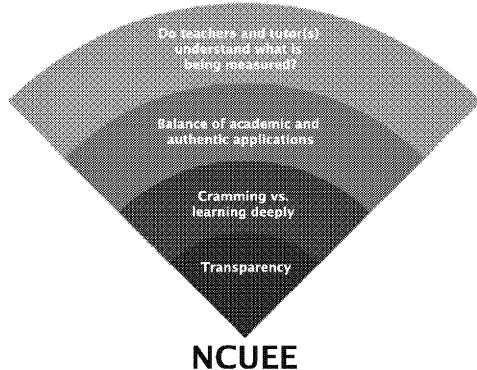
Is NCUEE Measuring What Students Need for Future Success?



14 / 19

National Center for Research on Evaluation, Standards, & Student Testing

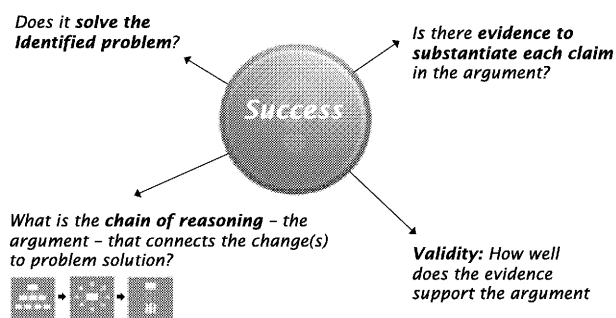
Is NCUEE Sending the Right Signal?



15 / 19

National Center for Research on Evaluation, Standards, & Student Testing

How Will You Know if NCUEE's "Revision" Is Successful?



16 / 19

National Center for Research on Evaluation, Standards, & Student Testing

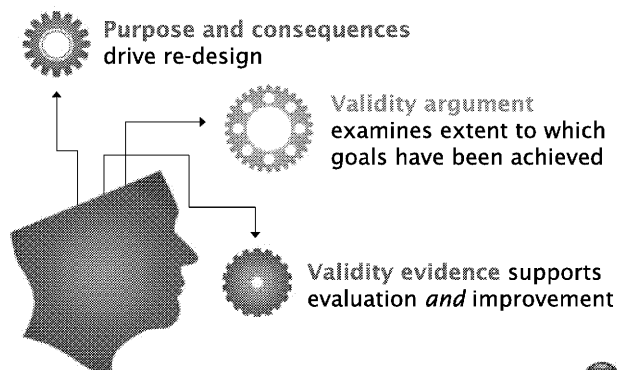
Possible Claims in the Argument

- #1 Better measures college readiness
- #2 Fair(er) for all students and diverse subgroups
- #3 Scores predict college success
- #4 Enables better admissions decisions
- #5 Improves teaching and test preparation K-12
- #6 Improves/deepens student learning

17 / 19

National Center for Research on Evaluation, Standards, & Student Testing

Concluding Thoughts



18 / 19

National Center for Research on Evaluation, Standards, & Student Testing



**National Center for Research
on Evaluation, Standards, & Student Testing**

UCLA | Graduate School of Education & Information Studies

VISIT US ON THE WEB
cresst.org

herman@cse.ucla.edu

パネリスト講演からの示唆

荒井克弘

(大学入試センター 入学者選抜研究機構長)

荒井機構長 ご紹介ありがとうございました。大学入試センターの荒井でございます。

主催者側の一人として、本日、たいへん充実した4本のご講演をお聴きできたことを深く感謝申し上げます。またコメンテーターのジョアン先生からは、大学入学テストの役割とその機能、実施組織のあり方に関して、示唆に富んだお話を聞くことができました。

私の役割は、主催者のひとりでありながら、このシンポジウムのコメンテーターのひとりということでもあります。同時に2つの役割は、いかにもやりにくいことで、どういうふうにコメントを構成したらよいか、じつは大変苦慮しておりました。

開会のところで、吉本理事長からシンポジウムの開催趣旨を含めたご挨拶をいただきましたが、それに関連してお話を始めさせていただこうかと思っております。

日本の大学入試の状況については、その歴史を含めて、銭谷先生からたいへん分かりやすいご説明を頂いておりますが、その定量的な部分で、私のほうでも整理したものもございますので、それをまずご紹介した上で、4つの講演についての私の感じたポイントを申し上げたいと存じます。

今回のシンポジウムは、「教育テストの可能性」というタイトルをつけました。昨年の、第1回の国際シンポジウムでは、“これからの大学入学者選抜を考える”というテーマで、朝日新聞社と共催で企画を進めさせていただきました。

大学入学者選抜研究機構がその年に大学入試センターに設置されましたので、昨年のシンポジウムは私たちの研究機構の発足記念でもありました。この発足にあたって研究機構に幾つかのミッションを担わされたわけですが、それは、大学入試研究の成果を社会に向けて積極的に発信していく、社会的な要請の高い課題に取り組む、さらに大学入試センターと社会のインターフェイスをつくり上げていくということでございます。自分たちの力に余るミッションでございますが、今回のシンポジウムもそのミッションを意識しながら、新しい議論を巻き起こしていくきっかけになればと考えております。

第1回目の国際シンポジウムでは、“これからの入学者選抜”をテーマに、大学・短大等の高等教育の側から現在の入学者選抜の状況を考え、これからどうするべきなのかということを議論させていただきました。そこではっきりしてきたのは、日本の現状はもはや選抜を中心に大学入学者選抜を考

えるという時代ではなくなった、次の段階に移ってきたようだ、ということでもあります。選抜よりも教育を中心に大学入学者選抜を考えなければいけない時代がやってきた、と言ってもよいかと思いません。

「教育テスト」という言葉は、日本ではそれほどなじみのある言葉ではありません。エデュケーション・テストという言葉は、今日の講演の中にも何度も出てきますが、これをどのように翻訳するのがよいのか考えまして、ここでは「教育を支援するためのテスト」という意味で「教育テスト」と呼ぶことにしました。大学入学者選抜のテストは“選抜テスト”から“教育テスト”へ変えていかなければいけない。そのコンテキストで、教育テストと正面から取り組んで行かなくてはならないというのが今回のシンポジウムの趣旨でございます。

グラフをご覧いただきたいと存じますが、青の実線のグラフが高校への進学率の推移です。1975年あたりで90%を超え、その後は、進学率の上昇は頭打ちとなり、高校進学が全入化に近づいていきます。赤の実線のグラフは、大学、短大への進学率の推移です。75～85年あたりで一旦停滞する時期がありますが、これは大学進学を量的拡大を政策的に抑制した時期にあたります。80年代の後半になりますと、年齢人口の急増に合わせて政策的な抑制は緩和され、第2期の大学進学を量的拡大が進みました。現在は、年齢人口の60%近くの人たちが大学、短大に進学する時代です。

グラフの背後に、18歳人口の推移が描いてあります。高校を卒業する人たちの年齢人口がどのように推移してきたのか、このグラフからお分かりいただけると思います。18歳人口は90年代のはじめにピークがあり、その後、急速に年齢人口が減少していきます。人口は急速に減少していきませんが、大学、短大の収容力は人口急増期に拡大され、それが大半維持されていますので、進学率は人口急減期に劇的に上昇することになります。それが、この間の変化の構造です。

そして、入学者選抜に何が起こったのか。我が国の高等教育（大学、短大等）は圧倒的に私学セクターで占められていることはご存じの通りです。入学者の8割近い学生たちが私立大学や私立短大に入学して行き、国公立大学・短大には2割を少し超える程度の学生たちが入学してくることになります。

次のグラフは四年制私立大学の入学事情を表しています。学生たちはどういうルートで大学、短大に入学してくるのか。改めてグラフにしてみました。学科試験を受けて入学してくる学生たちは、私立大学ではいまや48%になりました。残りのほとんどが推薦入試（ノミネーティッド・アドミッション）、AO入試（コンプリヘンシブ・アドミッション）になります。推薦入試とAO入試とは、学科試験を課さないことが選抜方法の原則でしたので、現在、私立大学に入ってくる入学者47万人の半数以上が学科試験を経ないで大学、短大に入学してくることになります。他方、国公立大学のほうは、まだ学科試験方式が健在であり、8割強が学科試験を受けて大学に入ってきます。推薦入試、AO入

試は、合わせても2割以下に止まっています。

さらに、短期大学（ジュニアカレッジ）のほうはどうなっているのか、私立短期大学での推薦入試、AO入試への偏りはより著しい傾向があります。そもそも短期大学は四年制大学以上に私学セクターが大きなシェアを占めており、割合は95%近くに達します。そしてその入学者の8割は推薦入試、AO入試で占められるという具合です。言い換えれば、短期大学に入ってくる入学者の8割は、学科試験を受けていないという実態です。国公立の短期大学はわずかで、短期大学全体の入学者数の5%を占めるにすぎませんが、入学者の半分は推薦入試、半分は学科試験を受けています。

20年ほど前までは、日本ほど学力選抜に熱心な国はなかったのだらうと思いますが、それが短期間に劇的な転換を遂げました。先ほど、銭谷先生が、社会的に大きな論点になっているわけではないけれども、静かにいろいろな変化が進行しているとお話になりました。静かな変化ではあるけれども、わが国の教育の歴史にとって劇的な変化が進んでいることは確かでしょう。

もうひとつ別のグラフをお目に掛けます。短大と学部、そして大学院の学生数全体の変化を示したものです。1990年のはじめから、この20年間に高等教育の規模はどのように変化してきたのか、それを示したグラフです。

先ほど申しあげましたように、この20年ほどの間に年齢人口は急速に減少してきました。90年代前半に205万人いた年齢人口は現在120万人まで減っています。10年後には110万人まで減ることが予想されています。しかし、少なくとも1990～2010年まで、日本の高等教育の規模はいささかも減らずに、一定の規模が保たれてきた。これが305万人ラインですが、この規模の高等教育が維持されてきたという事実があります。

これをどのように考えるのかということで、これからの大学入学者選抜を考えるスタンスが違ってまいります。多くの学生が大学に行きたい、ぜひともこういう分野で勉強したいという教育要求の増加があつて、年齢人口の減少にもかかわらず、これだけの進学率が達成されているのかどうか。事実はそうではなく、大学と短大、高等教育の規模を維持するために、学生の意志とはかかわらずに収容力の拡大がもたらした結果であるかもしれません。それは教育の問題というよりも、経営的な必要性、財政的な事情が優先された結果と解釈することができます。その観点に立ちますと、20年近く変わらなかったこの高等教育規模は、今後も維持される可能性が高い。年齢人口がさらに10万人減ったとしても、教員や事務職員の失業回避が最優先され、高等教育というマーケットは維持されるという選択はおおいにあります。それが日本の高等教育を考えるうえで無視できない要素です。だとすれば、日本の大学入学者選抜を教育的な見地だけから考えることは難しいのです。そのことを我々は切実な課題として考慮していかなければいけない。そういう状況に我々はいます。社会的なイシューにはなりにくい面があるかもしれませんが、高等教育の基本的条件としてはシリアスな状況です。私学セクタ

ーが大きいというだけで、経営的な制約は相当に厳しい。そうした中で何が可能かということが重要
です。

本日、4つの貴重なご報告をいただきました。それぞれが我々にとってこれから考えていかなければ
いけない、あるいは選択していかなければいけない方向をある程度指し示しているように思います。

第1に、ジン先生は、韓国におけるイングリッシュ・アビリティテスト（NEAT）の導入を紹介
されました。NEATが韓国の大学修学能力試験（CSAT）の英語のテストにとって替わるとい
うお話でした。この改革が我々にとってどのような意味を持っているのか。これは私の個人的理解に
過ぎませんが、韓国におけるNEATの導入は、旧来の大学入試という考えかたを突き崩すひとつの
ブレークスルーのようです。

韓国では、実施にあたって、レベル2とレベル3というふうに慎重に分けて、学術的な英語と運用
的な英語を峻別する考えかたをとっています。しかしいずれにしても、英語という学校の教科を前提
とした試験から、実際に使える英語、話す・聞く・書く・読むというその4機能を同時に備えた言語
能力（技量）試験に切り替えていくという方向性をはっきりと示していると思います。それを運用力
ベース（プロフィシエンシー・ベースト）と呼ぶのであれば、教科主義（サブジェクト・プリンシ
ブル）から運用力ベースに替わるという変化を意味します。これは、日本の入学試験が立脚してきた枠
組みとは異なる方向です。そういう方向を韓国がいま模索しはじめた。教科主義にとられない学力
試験の新しい実験が「英語」という外国語教科の分野で進められている。これは我々にとってたいへ
ん刺激的な事実です。

第2に、ACTのデボラ・ハリス先生からは、ACTが大学入学に向けてどのようなテストサー
ビスを展開しているのか、進路選択のためのエクスプローラー・プログラムなど、ACTAPという入学
テストを軸に、大学入学への準備プログラムがどのように整備されているのか、目の当たりに見せて
いただきました。

ACTは、ハリス先生のお話にもありましたように、1950年代の終わりに創設されたアメリカのテ
スト機関です。その創設者であるアイオワ大学のリンクスト教授は、テストというものは、選抜のツ
ールという以上に教育にとって重要なツールなのだ、ということを強調して言うておられます。

我々は、大学に進学する準備をするというと、入学試験のための勉強、その準備をあれこれ考えま
す。このところ、俄に注目されている大学と高校の接続という問題に関しても、入学者の学力の多様
性が拡大し、大学が求めている入学学力と高校での卒業水準の間に乖離が拡大しているという議論が
主流です。しかし、大学入学に必要な学力の達成は、ハリス先生の話聞いていますと、必要条件に
すぎない。高校から大学への進学を保障するものは、学力だけではなく、それ以外のさまざまな精神
的な成熟、個人の進路選択を実現するための準備が重要だということが分かります。入学試験に合格

するだけで、高大接続の条件がみたされるわけではない。それは条件の一部にすぎない。大学進学への必要条件と十分条件を満たすこと、その概念がカレッジ・レディネスであると伺いました。講演のなかでは、カレッジ・レディネスからさらに広がって、キャリア・レディネスという内容にも話が及んでいたと思います。

いずれにしても、ACTの事業の目的は、進学する大学でよりよい学びを可能にすること、そして社会に出て望ましい仕事を達成することです。単純に大学教育と高校教育をつなげるという必要条件のためだけに何かを考えるとということではない。このカレッジ・レディネスという概念は、我々にさまざまな見方を与えてくれたように思います。

第3に、エヴァ・ベーカー先生からは、たいへん大胆な教育の構想、新しい評価（アセスメント）のお話を伺いました。

ベーカー先生のお話の中で印象的であったのは、「未来」という言葉と「予測（プレディクション）」という言葉です。未来は不確定である、21世紀も予測不能な不確定な環境です。新しい環境に置かれたときに、我々はどう行動し、何をすべきか。何を学び、その学びとられた知識・能力はどのように測られるべきなのか。ベーカー先生は学んだ知識内容や学ぼうとする概念そのものを、存在論、オントロジーという表現で話をされていました。その概念にもとづく、知識の構成、またそれを測定する方法を含めて、オントロジーが21世紀に必要な知識、能力というものを探っていくための新しいツールだということを我々に伝えてくれたと思います。

教育の評価も従来の試験概念をこえて、オントロジーをベースにして組み立てられる。それが新しいアセスメントのデザインであるということも伺うことができました。ベーカー先生のお話は、現時点から先の話、未来を展望した入学試験、これからの教育アセスメントのビジョンを示していただいたと思います。

そして最後に、銭谷先生からは、豊富な行政経験をもとに、我々の世代にとってたいへん懐かしい時代の話から、現在の教育課題全般にいたる話題までご呈示をいただきました。知識だけではなく、思考力、判断力、応用力の必要、それから90年代から新しい学力観として注目された、意欲・関心・態度といったもの、そういう非認知的（ノンコグニティブ）なものをどのように測るのか。またさらに、近年、急速に普及した学科試験以外の入学者選抜方法である推薦入試、AO入試というものをいかに“実質化”していくのかということも課題としてご指摘いただきました。社会が急速にグローバル化していく。その中で活躍できる人材育成というものをどのように考えていくのか。こういった問題に関連させて入学者選抜の新たな構想の必要性を示唆していただきました。

銭谷先生が課題として示されたものと、その前にお話しいただいた3つの講演は、深くつながっております。高等教育の拡大がそれぞれの国で進む中で、高校と大学の接続の困難さは、各国共通の問

題になっています。たいへん似た状況、同じような問題がそこに析出しています。その点で、我々がいま取り組んでいる問題は、きわめてグローバルな問題であるといえます。

私の話はコメントという言葉で括るには、余りふさわしくない内容だったかもしれませんが、本日の貴重な講演、そしてジョアン先生のコメントから我々が学べるものはとても多くのものがありました。それらを今度は現実の社会に、そして我々自身の組織に持ち帰って議論を深めてまいりたいと考えます。その点で、本日、ご参加いただいた方々からも、アンケート等によりいろいろなご意見をいただきたいと思いますと考えております。

ご静聴、ありがとうございました。

パネリスト講演からの示唆

荒井克弘
独立行政法人大学入試センター・大学入学者選抜研究機構

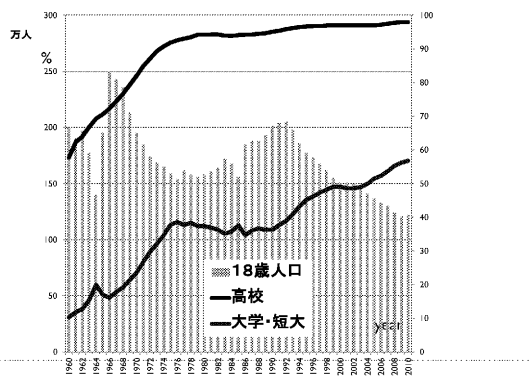
シンポジウム・テーマの趣旨

「これからの大学入学者選抜」を考える(2010国際シンポジウム)

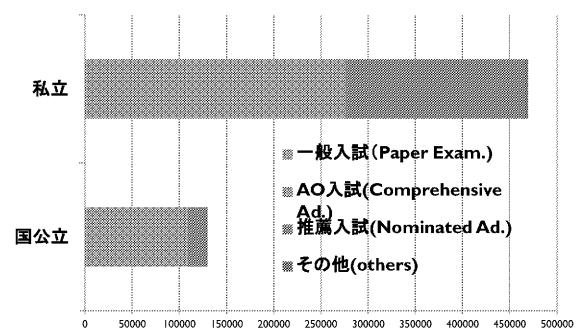


選抜テストの時代から教育テスト
(Educational Testing)の時代へ
:教育を支援するテストについて考える

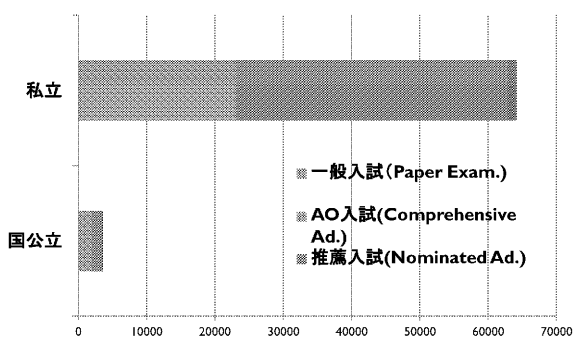
進学率の推移 (高校・大学&短大)



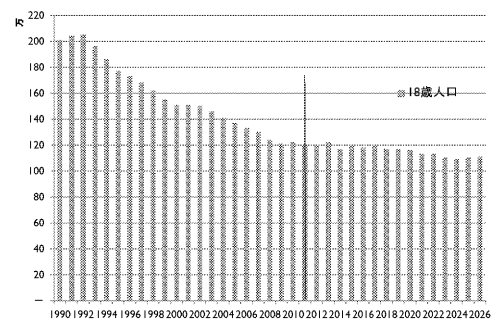
4年制大学の入学者数 (選抜方法別)



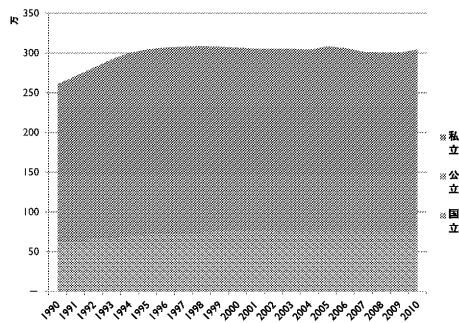
短期大学の入学者数 (選抜方法別)



18歳人口の推移



高等教育の学生数(短大・学部・大学院)



“New Approaches to University Entrance Examinations in Korea -NEAT(New English Ability Test) and CSAT(College Scholastic Ability Test)”

By Kyung-Ae Jin (KICE)

新しい英語検定システム
(NEAT)の導入
教科主義(subject principle)から運用力
重視(proficiency-based)へ

“New Approaches to Educational Testing and ACT”

By Deborah Harris (ACT Inc.)

カレッジレディネス

College Readiness

試験準備から進学準備へ
学力診断、興味検査、職業選択、
進学経費の見積もり等

“New Approaches to Measuring 21st Learning”

By Eva L. Baker (UCLA)

21世紀に必要な知識能力とは何か？
何を学び、学んだものをどのように測るか？
Ontology-Basedの知識内容の構成
知識内容を測る評価方法の再デザイン

我が国の初中等教育政策と大学入試

By 銭谷真美 (東京国立博物館長・
文部科学事務次官)

- 現状
- ・ 受験競争の過熱から大学入試の多様化へ
 - ・ AO入試、推薦入試の普及・・・選抜機能の低下
 - ・ 進学者の学習意欲、高校教育の質保証
- 課題
- ・ 思考力、判断力、応用力、意欲・関心・態度等の評価
 - ・ AO入試、推薦入試の実質化
 - ・ グローバル人材の育成

WRAPPING UP

Eva L. Baker
CRESST National Center for Research on Evaluation, Standards, & Student Testing
University of California, Los Angeles

International Symposium of Organization for the Study of College Admissions
National Center for University Entrance Examinations
Tokyo
18 November 2011

NCUEE-OSCA 1

Wrapping up a great day

Conclusions
Research Options
Policy Speculation

NCUEE-OSCA 2

Impact of Uncertainty on Admissions Testing

- Achievement is more than selection
- Learning a continuous process
- Students falling below prior expectations should be helped - joint responsibility
- Admissions tests may transform into placement examinations or markers addressing 21st century skills in evolving content domains for use in learning



NCUEE-OSCA 3

© Regents of the University of California

Admissions Exams Competing Goals

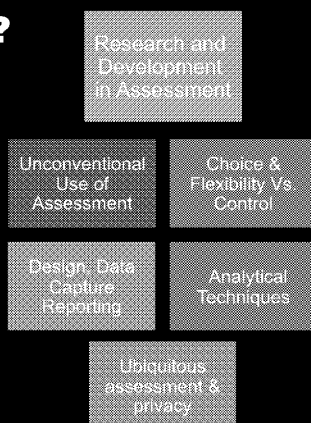
- Matching individuals to best available programs
- Identifying and supporting undiscovered sources of talent
- Meeting varied institutional missions
- Energizing educational pipeline to improve achievement
- Reconceptualizing role of learning and tests



NCUEE-OSCA 4

© Regents of the University of California

Futures?



NCUEE-OSCA 5

© Regents of the University of California

独立行政法人大学入試センター 入学者選抜研究機構国際シンポジウム報告書
「教育テストの可能性ー21世紀型能力の育成と高大接続ー」

発行 平成 24 年 3 月 31 日

編集・発行 独立行政法人大学入試センター入学者選抜研究機構
〒153-8501 東京都目黒区駒場 2-19-23
電話 : 03-3468-3311 (代)

印刷 株式会社 コームラ

