The Gordon Commission
on the Future of Assessment in Education

Assessment in the Service of Teaching and Learning

Clifford Hill

Arthur I. Gates Professor of Language and Education Emeritus

Teachers College, Columbia University

Introduction

As the Gordon Commission undertakes its task of considering the future of educational assessment, I would like at the outset to examine the varied forms that testing has taken in different societies, with particular attention to the range of purposes it has served. A comparative perspective can shed light on the complex ways in which cultural values and available technologies have shaped assessment practices and thus help us imagine new directions for the future.

I would also like to clarify how I will be using certain terms. Throughout this document *assessment* will be used as a generic descriptor of any activity designed to evaluate human performance, whereas *testing* will be restricted to an activity conducted under the following conditions:

stable tasks
a limited time frame
no external support[1]
a pre-established evaluation scheme
institutional sanction[2]

As the Commission explores future directions, the degree to which the strict conditions of testing are to be preserved is a fundamental question to be considered, especially in the light of digital technologies that can tell us not just the 'what' but also the 'how' of an individual response to a test.

Origins of Testing in China

It is generally accepted that testing originated in Imperial China in the 3rd century when Emperor Wu, seeking to find a counterweight to the increasing power of the noble class,

[1] In the most restricted approach to testing, individuals are allowed no external tools at all. But a more liberal approach is, at times, adopted in which individuals are allowed to use certain tools—for example, a calculator or a dictionary—that they are accustomed to working with in the real world (this policy allows test takers to work more rapidly and thus complete more test items). But even with this approach, social interaction, even though it is fundamental to work in the real world, is not allowed, since it, unlike an external tool, cannot be strictly controlled.

[2] This last condition is often not specified, but if it is not present, then the activity is not really a test, even though all the other conditions are present—as in the case of, say, a practice test.

explored ways to recruit promising young men from the provinces to serve in the central government. This limited experiment in testing eventually led to its large-scale application at the national level under Emperor Yang at the beginning of the 7th century.

Although *selecting* was the most obvious purpose of the imperial examinations, they also served another important purpose: *certifying* that anyone who passed them had mastered not only the **四書** ‘Four Books’ of Confucian classics that delineated the basic principles of a harmonious society but also the specific bodies of knowledge that were useful in government service: civil law, geography, revenue and taxation, agriculture, and military strategy. The many years of arduous study preparing for the exams also nourished personal qualities of discipline and patience needed to be a reliable civil servant. Hence, the imperial examinations facilitated the kind of learning and character development that was useful in the work for which people were being selected.

Although the imperial examinations were administered at a national level, the number of individuals who sat for them was relatively limited in a country as vast as China. Unless an individual showed exceptional promise, he—and it was certainly a ‘he’ since women did not sit for these exams—would not undertake the many years of preparation that the exams required. Given the rigorous standards used by the court scholars who evaluated the exams, only about five percent of the candidates actually passed.

These scholars were demanding with respect not only to the content of an individual response but also to its form. As the imperial exams continued to evolve, candidates were increasingly expected to conform to a highly structured rhetorical template, known as **八股文** ‘the eight-legged essay,’ in which words and expressions that were considered offensive or that revealed the candidate’s identity could not be used. This template also specified the total number of sentences as well as the number that could be used in each of the eight sections. If a response did not sufficiently reflect this template, it did not receive high marks, even when its content was considered to have merit.

It is interesting that what we now think of as test prep was evidenced even in this early use of testing. Sample essays that successfully used the eight-legged model were widely disseminated with the imperial imprimatur. A concern with fair play was also evidenced, since as early as the tenth century the written responses of candidates were recopied in order to preserve their anonymity and thus reduce the possibility of biased evaluation. Hence, from the very beginning of testing we can observe certain problems that we still face when evaluating written work. As we will see, these problems led, in the first half of the 20th century, to the total removal of writing from the testing enterprise in this country.

Testing in European Countries

As Europeans began to visit China, they brought back news of the imperial examinations, which sparked the imagination of social reformers such as Jean-Jacques Rousseau. During the political uprisings of the 18th century, these reformers claimed that introducing testing into education would strengthen the emerging democratic movements, since it rewarded individuals on the basis of individual merit. No longer would an individual be admitted to a university simply because he was born into a family that belonged to the noble class or possessed substantial wealth.

Abitur in Germany. As popular political movements gained momentum, they led to innovative legislation that established testing in the educational systems of European countries. In 1788, a set of written examinations known as the *Abitur* was set up as the official university entrance requirement in Prussia, thus taking away the patronage of individual universities that were accustomed to admitting those who possessed favorable connections with the ruling powers. By 1812 the *Abitur* had spread to secondary schools throughout Prussia, and with the formation of the German Empire in 1871, it was introduced in German-speaking areas beyond Prussia. The *Abitur* is still functioning in Germany, albeit with substantial modifications that ensure its relevance to education in a modern state.

***Baccalauréat* in France.** In 1808, the *Baccalauréat* was established by Napoleon I in France as the diploma required for university admission, and it, too, was based on a system of exams. Although this system has had numerous modifications, it has endured for more than 200 years as the foundation of secondary education in France.

The *Baccalauréat* diploma is generally thought of as certifying the completion of studies at a *lycée*, but it is possible for individuals who have not attended a *lycée* to receive it if they pass the required exams in one of the following areas:[3]

> *Scientifique* (heavily weighted toward mathematics as well as physics, chemistry, and biology but also includes foreign languages, history, and geography)

[3] Students in all three areas must pass exams in philosophy and the French language.

Sciences économiques et sociales (heavily weighted toward the social and economic sciences but also includes the humanities as well as limited mathematics and natural sciences)

Littéraire (heavily weighted toward literature, history, geography, and foreign languages, but also includes limited mathematics and natural sciences)

For the most part, these exams consist of written essays, although in the natural sciences some laboratory work can be required.

In nearly all European countries, an examination system based on written essays is widely used at the completion of secondary school, although it goes under a variety of names, most commonly some form of *Matura*. This is the name used in Austria and Switzerland, whereas *Maturità* is used in Italy and Slovakia. In all these countries, students take a predetermined set of exams based on their course of study, and the results are used to certify that they have completed secondary studies. On the basis of this certification, students are then selected for further study at the university level. Those who pass at the highest level are admitted to the most prestigious universities: in France, for example, *les grandes écoles* such as *Science Polytechnique*. Students can also be selected for professional internships in designated fields that require them eventually to take highly specialized exams.

Professional education is especially well developed in Germany, where students enroll in *Berufsschule* that support domains as diverse as banking and plumbing. After nearly four years of salaried apprenticeships, they take exit exams that qualify them to work in a professional setting. These exams have achieved a reputation for effective integration of theory and practice, and they could offer a model for the Gordon Commission if it decides to consider assessment in professional domains. Even if it does not, it still may wish to explore professional assessment for exemplary practices not only in European countries but also in this country.

Testing in the United States: Developments in the Early 20th Century

The American tradition of testing can be traced to an egalitarian model of education: however inconsistently applied, the American ideal since the time of Thomas Jefferson has been of a public education system available to all citizens. As changing patterns of immigration produced a larger and more diverse student population in the early 20th century, educators enthusiastically embraced a new kind of testing that reflected the scientific ethos of the day. As Edward Thorndike put it, this model provides indices of merit that are "fair and objective, standardized, competitive—and quantified" (cited in Jonçich, 1968, p. 295). For Thorndike, the model was especially important for a culturally diverse student population, since he was concerned that teachers are often biased in the judgments they make about these students.

Given his view that "reading is reasoning," Thorndike (1915) believed that a reading test and an intelligence test could be understood as measuring much the same thing. He thus developed scaled tests of reading comprehension for which he claimed a number of advantages: they were easily scored, produced numerical scores rather than individual judgments, and minimized dependence on students' powers of written expression (Hill & Parry, 1994).

The multiple-choice format, which Thorndike used only in a limited way, was first introduced on a large scale by Arthur Otis and Lewis Terman who designed an intelligence test for assigning army recruits to specialized units during World War I. During the 1920s this format was used in large-scale administration of reading tests, which eventually came to be viewed as the final arbiter of a student's literacy knowledge and skills—and this is a position these tests still hold in American education.

The development of machine-scoring in the 1930s further increased the appeal of the multiple-choice format. After World War II, multiple-choice testing developed rapidly in this country and became a major American export as educators in many countries around the world sought to deal with rapidly expanding student populations. In the People's Republic of China, for example, the multiple-choice format is used in the university entrance exam, which was taken by more than twenty million students in 2011.[4]

As multiple-choice testing has been growing in popularity in many countries, educators in this country have become increasingly skeptical about its use (Engel, 2011; Johnson, 2009; McNeil, 2000; Ravitch, 2010). Given the high stakes associated with such testing, it tends to distort the curriculum and lead teachers to devote inordinate amounts of classroom time to test preparation, which is often based on commercially produced materials of dubious value (see Hill, 2000, for an op-ed piece on how insufficiently vetted material was rushed to the market and then used extensively at schools in poor neighborhoods in the Bronx and Harlem).

[4] From the perspective of Chinese academics who respect the ancient tradition of testing based on slow mastery of classical learning, multiple-choice testing can be viewed as a kind of fast food within modern education (Lin Qingming, personal communication). They have, at times, compared it to other American exports such as McDonald's and Kentucky Fried Chicken that adversely affect classical cuisine in Chinese culture.

At Teachers College, Columbia University, colleagues and I carried out various studies that challenge the model of reading comprehension that test makers end up reifying as they work with the multiple-choice format.[5] This model arises inevitably from various constraints that they must work with. To begin with, test makers select individual passages for a reading test that are relatively short and lacking in context, since they must construct a test that fits into a limited time frame and yet contains different kinds of material, so that no one sample has undue weight.

In constructing the multiple-choice tasks that accompany the individual passages, test makers face two further demands. On the one hand, they must be able to defend the choices that they designate as the target response. Hence, these choices have to be limited to what psycholinguists (Trabasso, 1981; van Dijk & Kintsch, 1983) designate as the *text base* (i.e., information that is explicitly stated or can be automatically inferred from a test passage) as opposed to the *situation model* that a reader constructs.

On the other hand, test makers have to provide alternative choices (i.e., distractors) that have sufficient discriminatory power: a task does not make it onto the test unless a sufficient number of students select a distractor during field testing. In order to construct genuinely attractive distractors, test makers tend to be drawn to various kinds of inferences that the relatively short and decontextualized passages activate, but which, strictly speaking, cannot be justified by the text base. Once distractors built around inferences are included in a multiple-choice task, they take on a life of their own and can lead a test taker to construct a larger world of meaning. Ultimately the text that a test taker has to comprehend is not simply the reading passage, but rather a larger configuration that consists of the passage and the tasks that accompany it.

The highly constrained model of reading comprehension that test makers are backed into is clearly at variance with the constructivist model of reading. It is thus not surprising that teachers committed to constructivist pedagogy find that multiple-choice testing undermines what they are trying to accomplish in the classroom and have become strong advocates for alternative approaches to assessment.

[5] These studies were carried out on multiple-choice tests used with native speakers of English (Aronowitz, 1984; Coyle, 1992; Hill, 1997a, 1997b, 1992, 1995, 1999, 2000, 2001; Hill, Anderson, Ray, & Watt, 1989; Hill & Larsen, 1983, 2000; Nix & Schwartz, 1979; Sims-West, 1996) as well as non-native speakers of English (Adames, 1987; Bhasin, 1990; Chu, 1993; Hill & Parry, 1988, 1989, 1990, 1992, 1994; Hill & Wang, 2001; Ingulsrud, 1988; Parry, 1996; Yuan, 1997).

I would now like to turn to these alternative approaches and focus on two promising models that have been developed at Teachers College, Columbia University. The first will be described as a *digital testing model,* in that the strict conditions associated with testing are preserved in a digital environment. The second will be described as a *digital project model,* in that it draws on a familiar method of evaluating students in higher education—the course project. As we will see, however, the innovative use of digital technologies can transform this common approach into a powerful tool for learning. Indeed, in the case of both models, an appropriate use of digital technologies can foster a robust connection between assessment and classroom teaching and learning.

Digital Testing Model: Assessing Students in American High Schools

With support from a federal grant, colleagues and I developed a digital testing model for students in the Pacesetter Program,[6] a national program sponsored by the College Board that helps high school students prepare for higher education. This program is concentrated in urban high schools where many students come from culturally diverse backgrounds and do not speak standard English at home. In order to appeal to these students, the Pacesetter English course is built around culturally diverse material in three areas—literature, film, and media.

Within each of these areas, material can be found that reflects the cultural worlds of three major groups: African Americans, Latino Americans, and Asian Americans. At the same time, this curriculum emphasizes the importance of participating successfully in the larger society—hence the strong emphasis on technology, as evidenced by the digital testing model that we developed.

When students begin a Pacesetter assessment activity,[7] they are presented with *resources*, which may include a text, an audio clip, or a film clip, either in isolation or in various combinations. They then respond to three integrated tasks that focus on *factual*, *inferential*, and *experiential* aspects of comprehension.[8] In the first task—the *planning task*—students use digital

[6] The Pacesetter Program has evolved into a new program called the *College Success Initiative*, which includes middle school as well as high school and a more substantial digital component.

[7] Borrowing from Vygotsky (1962), I use the term *activity* rather than the traditional term *test item*.

[8] The overall design of our assessment activities can be traced to the model of comprehension developed by Bloom (1984). This model was previously used in developing assessment for early childhood education (Hill, 1992,

tools to interact with the resources and create a database of relevant material that they can draw on as they respond to the two tasks that follow: the *interpretation task,* which asks them to use the database in exploring more deeply an issue raised by the resources, and the *application task,* in which they draw on their own experience to deal with the issue in a broader context. In effect, students use digital tools in the planning task to gather information from the resources that they can then use in responding to the constructivist tasks that follow. Hence, we use the term *grounded constructivism* to describe the set of integrated tasks: the use of digital tools on the first task insures that students are appropriately grounded in the resources before they move on to respond to higher-level tasks (see Figure 1).

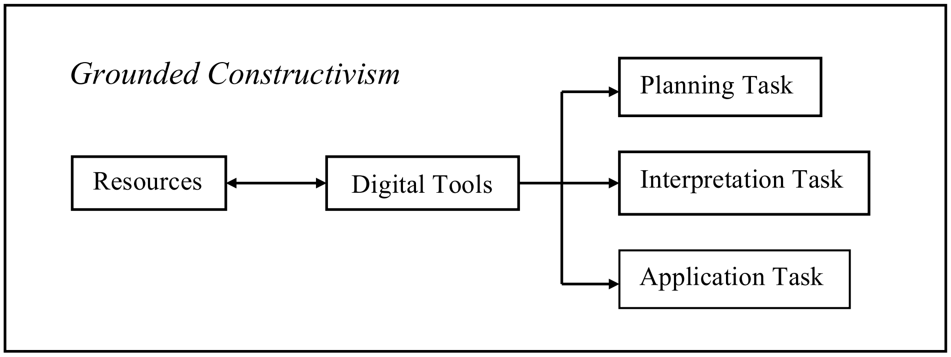


Figure 1

Before presenting a sample activity that shows how the model is implemented, I would like to illustrate briefly some of the ways that one digital tool—the search function—is used during the planning task to produce a database. In an activity based on "The Browning of America," a media essay by Richard Rodriguez (1998), students are asked to search for the different ways that he uses the word "brown" (he uses it as a noun and a verb as well as an adjective). They then use the copy/paste function to assemble these examples in a database. In the interpretation task, they use this database in discussing how the author's distinctive uses of "brown" help him convey a new vision of race in the U.S.

In an activity based on "Stranger in the Village," an essay written by James Baldwin in 1953, students are asked to think about how the terms "Negro" and "black man" are used. After searching for uses of the two terms and arranging them in a database, they can see that for

1994, 1999, 2000; Hill & Larsen, 2000) and adult education (Hill, Anderson, Ray, & Watt, 1989; Hill & Parry, 1988, 1989, 1990, 1992, 1994).

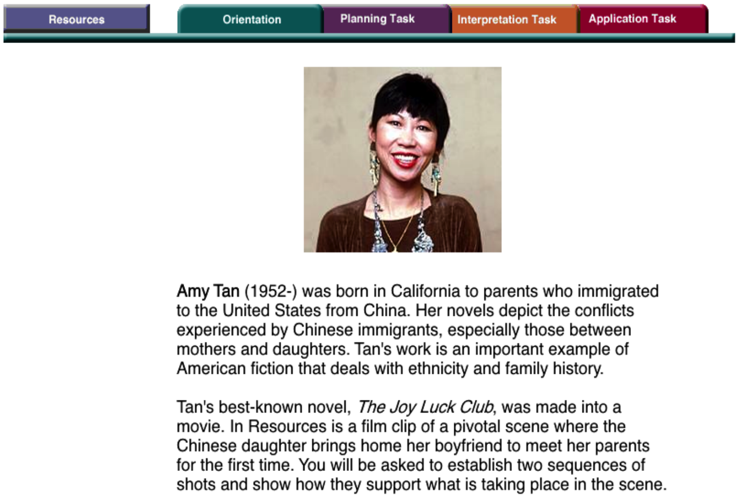
Baldwin “black man” is a general term for people of African descent, whereas “Negro” is used to refer only to African Americans. This distinction reflects the larger purpose of the essay, which is to contrast the experience of people of African descent in Europe and the United States. The application task focuses on the situation some 15 years later when the term “Negro” fell out of favor and was replaced by “Black.” And, of course, using “man” to refer to people in general was no longer acceptable.

Another activity is based on a news report about the Cartoon Network’s decision to stop showing Speedy Gonzalez cartoons because of the ethnic stereotyping they contain. The report reflects a familiar genre in which people are quoted on both sides of an issue. By searching for quotation marks, students are able to build a database that contrasts the arguments for the ban with those against it. In the interpretation task, they are asked to isolate the strongest arguments and then add at least one argument of their own on each side. Then in the application task, they write an email to the head of public relations at the Cartoon Network in which they not only take their own position on the issue but also address the arguments used to support the opposing position.

Throughout these activities students are using a search tool to rapidly build a database that would be tedious and time-consuming to produce in a traditional print environment. One of our major goals in building an assessment activity is to simulate strategic uses of the computer in the real world: in this instance the rapid assemblage of information through searching and copying/pasting allows more time for critical reflection on that information.

Sample Assessment Activity

To illustrate how our assessment activities work, let us examine an activity based on the film *The Joy Luck Club*, which plays an important role in the Pacesetter English curriculum. The activity begins with an *orientation* screen (see Figure 2) that provides background information about Amy Tan and the novel on which the film is based.



Amy Tan (1952-) was born in California to parents who immigrated to the United States from China. Her novels depict the conflicts experienced by Chinese immigrants, especially those between mothers and daughters. Tan's work is an important example of American fiction that deals with ethnicity and family history.

Tan's best-known novel, *The Joy Luck Club*, was made into a movie. In Resources is a film clip of a pivotal scene where the Chinese daughter brings home her boyfriend to meet her parents for the first time. You will be asked to establish two sequences of shots and show how they support what is taking place in the scene.

Figure 2

It then describes the film clip presented in the resources and outlines what students are to do. The tabs at the top provide a simple way for students to access the different parts of the activity.

The resource window is shown in Figure 3. Buttons enable students to play either the whole film clip or one of the segments that they will be working with.

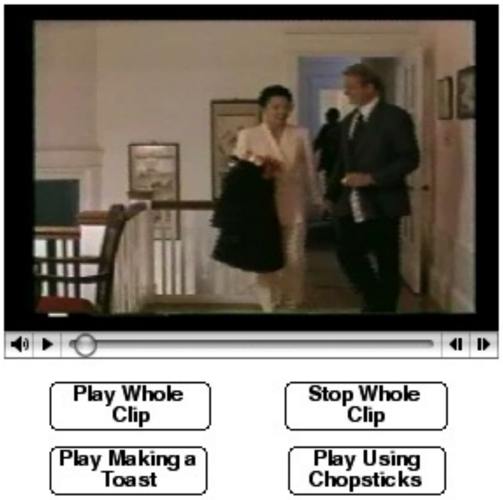
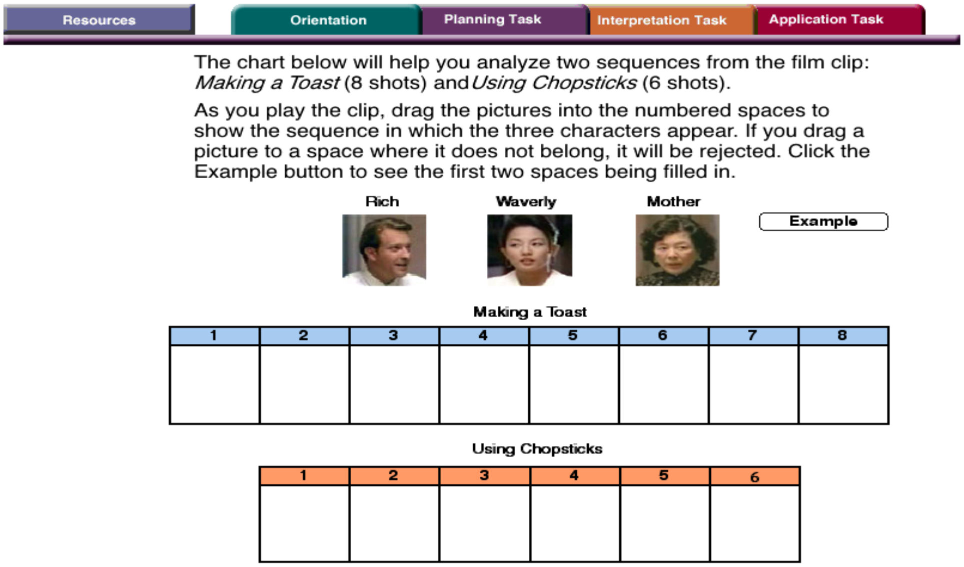


Figure 3

In the planning task (see Figure 4), students examine two segments of the film clip—Making a Toast and Using Chopsticks—to discover the sequence in which three characters—the mother, the daughter, and her boyfriend—are presented.



The chart below will help you analyze two sequences from the film clip: *Making a Toast* (8 shots) and *Using Chopsticks* (6 shots).

As you play the clip, drag the pictures into the numbered spaces to show the sequence in which the three characters appear. If you drag a picture to a space where it does not belong, it will be rejected. Click the Example button to see the first two spaces being filled in.

Making a Toast

1	2	3	4	5	6	7	8

Using Chopsticks

1	2	3	4	5	6

Figure 4

Students show the sequences they discover by dragging pictures of the characters into the appropriate boxes, as shown in Figure 5. If they drag a picture to a box where it does not belong, it will be rejected. In other words, the task is designed so that all students end up with the same sequences. In Making a Toast, the characters appear in the sequence boyfriend > mother > boyfriend > daughter, but in Using Chopsticks the sequence is boyfriend > mother > daughter.

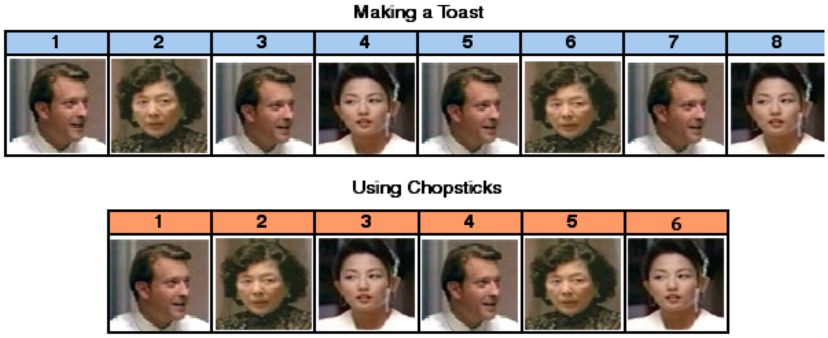
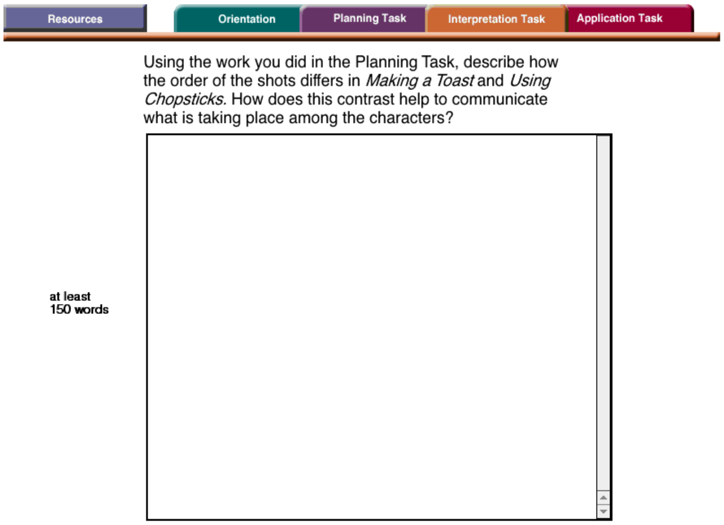


Figure 5

Figure 6 shows the interpretation task in which students work directly with the information assembled in the planning task. They are asked to make a descriptive generalization about the facts first and only then to interpret them. The rationale for this description/interpretation sequence is twofold: first, we want to assess whether students can clearly describe the information they have assembled in the database; and second, we want to encourage students to ground their interpretations in an accurate description.



Using the work you did in the Planning Task, describe how the order of the shots differs in *Making a Toast* and *Using Chopsticks*. How does this contrast help to communicate what is taking place among the characters?

at least 150 words

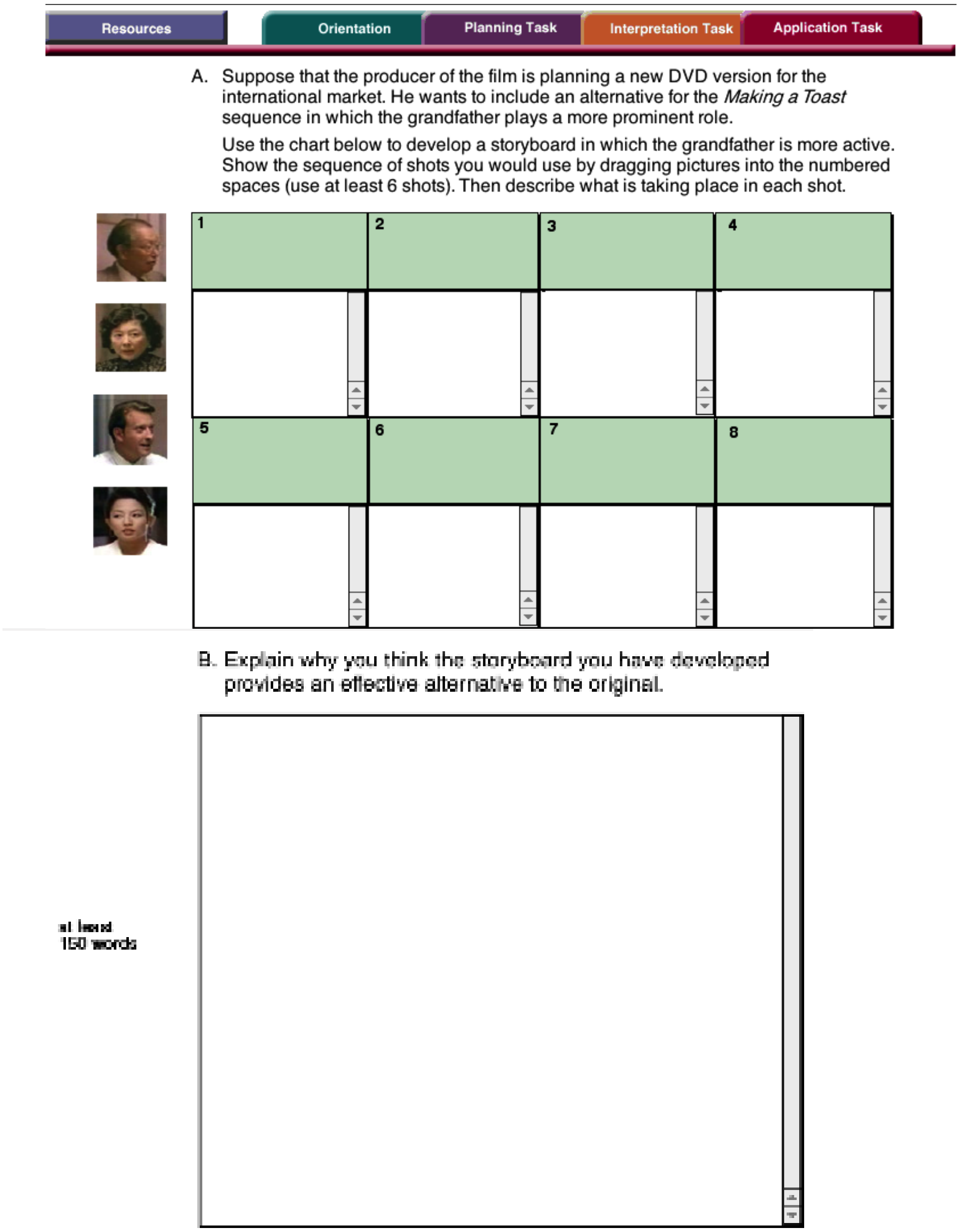
Figure 6

The Pacesetter curriculum is designed to develop students' film literacy. Hence it teaches them to analyze not only technical aspects of film shots—for example, the camera angle used—but also the sequences in which shots occur. As students carry out such analyses, they begin to understand how film technique affects meaning.

Field-testing revealed that this was a demanding task, although a number of students were able to provide insightful answers. For example, one girl pointed out that during the early part of the dinner the daughter is still hopeful that her mother will accept her boyfriend. As the dinner progresses, however, she becomes increasingly aware of her mother's negative reactions and begins "to filter her reactions to the boyfriend through the mother."

We found this phrasing particularly apt in characterizing what was happening as the dinner progressed. In *Making a Toast*, the daughter is shown responding independently to her boy friend, but in *Using Chopsticks*, the mother is shown as interposed between the boyfriend and the daughter, which carries the message that the mother is now controlling her daughter's response to the boyfriend.

In the application task (see Figure 7), Part A calls for students to carry out a familiar activity in the Pacesetter curriculum—creating a storyboard to represent an alternative version of a film scene. Part B asks them to reflect on this storyboard and to express why they think it would be effective. In this digital approach to storyboarding, students can drag a picture of one or more characters to a particular box and then use the space below it to represent what takes place.



A. Suppose that the producer of the film is planning a new DVD version for the international market. He wants to include an alternative for the *Making a Toast* sequence in which the grandfather plays a more prominent role.

Use the chart below to develop a storyboard in which the grandfather is more active. Show the sequence of shots you would use by dragging pictures into the numbered spaces (use at least 6 shots). Then describe what is taking place in each shot.

B. Explain why you think the storyboard you have developed provides an effective alternative to the original.

at least
150 words

Figure 7

During the field-testing, we discovered considerable variety in the storyboards that students created. A number of students produced a script in which the grandfather first models how the boyfriend should make a toast, the boyfriend makes a more acceptable toast (e.g., sipping the wine instead of gulping it), the mother responds more positively, and the daughter ends up in a considerably happier state. Not all students, however, opted for an optimistic outcome. There were a few who produced a script built around the battle of the sexes: in one storyboard, for example, the grandfather joins the boyfriend in excessive drinking, which sets off fierce opposition from the mother and daughter.

Sample Evaluation of a Student Response

Each response is read by two evaluators, and if they disagree, a third is consulted. During the reading, various points are noted for potential inclusion in the written feedback, which is drafted by one evaluator and edited by the other. The observations made are typically supported with references to what the student has written, often through direct quotation. And even when there is much to praise, time is still taken to point out missed opportunities. The pronoun *we* is freely used so that the comments will not seem arbitrary. The overall stance of the feedback is that it represents a consensus of engaged and honest readers.

Thus the evaluation of our assessment activities can be described as holistic. It is based on the reactions of real readers with a genuine interest in what a real student has to say, rather than a search through a student's responses for the presence or absence of predetermined characteristics.

The evaluators do use the rubric presented in Figure 8 as a framework to ensure that feedback to different students is reasonably consistent.

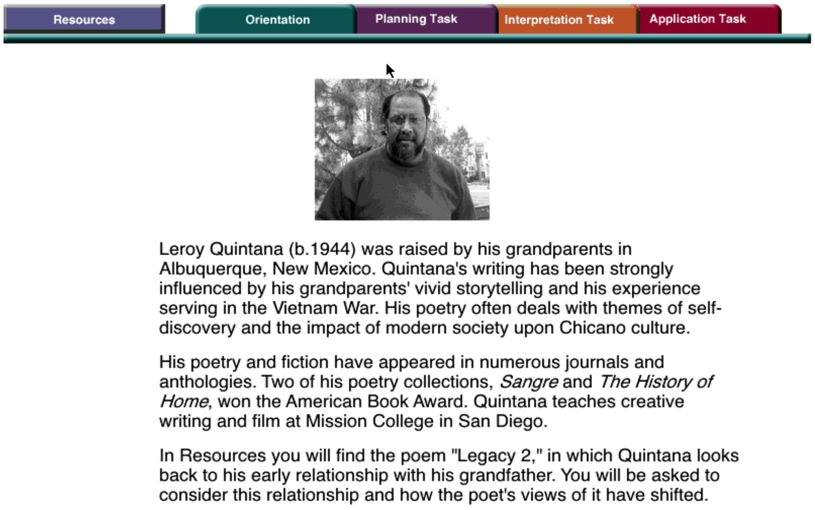
Content	accuracy	sufficient support
Clarity	an overall structure that fits the task	clearly expressed statements in an effective sequence
Critical/Creative Thinking	relevant connections	authentic interaction with resource material

Figure 8

The tripartite structure—*Content*, *Clarity*, and *Critical/Creative Thinking*—is adapted from a rubric developed for the International Baccalaureate (see Hill, 1996). It provides not only an easily understood way of structuring the feedback but also a grounded way to discuss with students—both before and after assessment—the standards of good writing that are being applied.

Within each category, two core attributes are provided as a frame of reference. With *Content,* for example, accuracy and the provision of adequate support for the statements one makes are regarded as key. With respect to *Clarity,* the attributes suggest the different levels that contribute to our perception of whether a text is clear**:** the overall structure and the phrasing and sequencing of individual sentences. *Critical/Creative Thinking* considers the extent to which a student "makes meaning," to use a common phrase. An important aspect of such construction is whether students make connections between the material in the resources and their knowledge of the world and personal experience. Another aspect is the genuineness of their communication, the sense that they have not only engaged the material but have attempted to say something important about the encounter. A major goal of the Pacesetter English course is to encourage students to develop an authentic voice.

To illustrate how we evaluate student responses, let us turn to another assessment activity, this one based on a poem. The orientation is presented in Figure 9 and the poem in Figure 10.



Leroy Quintana (b.1944) was raised by his grandparents in Albuquerque, New Mexico. Quintana's writing has been strongly influenced by his grandparents' vivid storytelling and his experience serving in the Vietnam War. His poetry often deals with themes of self-discovery and the impact of modern society upon Chicano culture.

His poetry and fiction have appeared in numerous journals and anthologies. Two of his poetry collections, *Sangre* and *The History of Home*, won the American Book Award. Quintana teaches creative writing and film at Mission College in San Diego.

In Resources you will find the poem "Legacy 2," in which Quintana looks back to his early relationship with his grandfather. You will be asked to consider this relationship and how the poet's views of it have shifted.

Figure 9

Legacy 2
Leroy Quintana

(1) Grandfather never went to school
spoke only a few words of English,
a quiet man; when he talked
talked about simple things

(2) planting corn or about the weather
sometimes about herding sheep as a child.
One day he pointed to the four directions
Taught me their names

(3) *el norte*
poniente *oriente*
el sud
he spoke their names as if they were
one of only a handful of things
a man needed to know

(4) now I look back
only two generations removed
realize I am nothing but a poor fool
who went to college

(5) trying to find my way back
to the center of the world
where grandfather stood
that day.

Figure 10

The application task for this activity provides an opportunity to demonstrate full use of the rubric. Figure 11 includes the prompt, one student's response, and our feedback. This student did not structure his response according to what is called for in the prompt. This is pointed out in the feedback under *Clarity*, where it is suggested that he should at least acknowledge the framework that has been suggested, even if he decides to go in another direction. Under *Content*, the evaluators point out several places where readers are likely to be puzzled by what he said. Under *Summary*, they observe that he finished early, using only about three quarters of the time available, and suggest that he could well have used the remaining time to improve his response.

On the positive side, the evaluators find a good deal to admire**:** in particular, the development of an interactive framework that is complex and effective. Taking all of this together, the evaluators gave this response a *promising+*, which is the second highest rating on our scale: *Accomplished*, *Promising*, *Developing*, *Beginning*.

Figure 11

Application Task

The poet writes about a contrast between his present life and that of his grandfather. What are some differences between the ways in which you and an older person you know look at things? Select one of these differences and give reasons to explain it. How does this difference affect your communication with this person?

I have an administrator who I have communicated with much during my time in high school named Bruce. Throughout high school, I have made it my goal to excel in certain areas of curriculum so as to stand out from the rest of my peers. Frustrated with my practice, Bruce, a much older man once in the armed forces, asked me, "Why are you trying to make things so much harder on yourself than they have to be?"

That struck me as odd. Throughout my entire juvenile life people have been telling me not to fall behind because no matter what I do there's always someone waiting to take my position and hold me back. So I replied, "There's so many people out there smarter and more qualified. I have to work hard now so I can make a better living later."

Bruce looked at me and sighed. Not a sigh of depression but one of sudden realization. "Look," he said, "you're going to find that you spend your entire life as a young man in search of money. You're going to work hard and maybe get it. But by the time you've earned enough to stop working so damned hard, you're going to be too old to enjoy it the way you wanted to."

I laughed at him when he told me this, years ago to date. After all, it made no sense to make money the primary incentive for staying and doing well in school and then turning around and deflating the importance of the correlation between money and happiness as Bruce had done.

During those years, I've continued to excel; hard work and constant struggle have defined my life long before and after my conversation with Bruce. Even now, I understand the message he was trying to convey, but I haven't really changed my course of action.

I don't and probably won't believe Bruce. That is, until and unless logic and experience prove him right. Although it seems feasible, it's not something I care to believe in. After all, I am a teenager, a young man in the prime of my life. I believe acting on Bruce's belief could ruin me; becoming so jaded, so weary of the struggles of life before I encounter even half of them seems foolish.

It is the word of an older generation against a younger generation, a person with the knowledge of experience against a person with the hope that naivety brings to the inexperienced. However, he was right about one thing, I will reach my goal. It's simply a question of, "When?"

Feedback

Content. We like the way in which your piece is structured as a real exchange between you and Bruce. This approach enables you to develop a lot of rich content. That said, we find your presentation of the contrast between "enjoying life while you can" and "working hard to ensure future success" sometimes hard to follow. For example, you suggest that Bruce contradicts himself: "It made no sense to make money the primary incentive for staying and doing well in school and then turning around and deflating the importance of the correlation between money and happiness." Yet in the preceding text it is you, not Bruce, who suggests that future success is the main reason for working hard in school.

Your presentation of the contrast is either-or, which is easy for the writer, but less straightforward for readers, who may have thoughts that don't fit neatly into the writer's scheme. For example, you don't allow for the possibility that hard work can be a deeply satisfying, even enjoyable, part of life. Also, we don't see how giving yourself more downtime might make you "so jaded, so weary of the struggles of life."

Clarity. You use paragraphing effectively—each of your seven paragraphs is a good length and is internally coherent. We especially liked the way in which you began a new paragraph by focusing on a reaction to what has just been said: "That struck me as odd," "Bruce looked at me and sighed," and "I laughed at him when he told me this."

Of course, you could have used paragraphing to respond to the task as it was structured: first, a presentation of various differences, next a focus on a single difference and its source, and finally a characterization of the ways in which this difference affects communication. It might have been good if you had found some way to acknowledge the structure of the task while approaching the material in your own way.

Critical/Creative Thinking. Your personal voice is heightened by the way you use quotation. We like the fact that you quoted yourself as well as Bruce. We especially like your technique of commenting on the quoted speech: for example, you end the first paragraph with speech attributed to Bruce and then start a new paragraph in which you first describe your internal reaction and then what you said. This combination of internal and external reaction helps to convey the complexity of your response to what Bruce said.

Summary. You managed to write a great deal in just over 37 minutes (336 words on the interpretation task and 432 words on the application task). And the process log provides evidence that you did some revision. Still it might have been helpful if you had used at least some of the remaining time to read your work carefully and do even further revision. We look forward to reading what you will write on the next assessment—your voice is strong and engaging on the page.

Sample Process Log

We also used the computer's capacity to keep track of time to help us monitor how students went about responding to our tasks. Hence, along with students' written responses, we collected data on the order in which they moved through the various sections of an assessment activity and how long they spent on a given section. Table 1 shows the amount of time that the same student spent on the various tasks built around Leroy Quintana's poem.

TABLE 1

Section	Time (min:sec)	Cumulative (min:sec)
Resource	37:01	37:01
Orientation	0:04	0:04
Planning Task	0:07	0:11
Interpretation Task	2:18	2:29
Planning Task	0:17	2:46
Orientation	0:01	2:47
Planning Task	1:14	4:01
Interpretation Task	10:16	14:17
Application Task	0:04	14:21
Interpretation Task	2:30	16:51
Application Task	18:44	35:35
Return to Work/Final Exit	0:03	35:38
Application Task	1:17	36:55
Return to Work/Final Exit	0:06	37:01
Application Task	0:09	37:10
Return to Work/Final Exit	0:05	37:15
Sent file to server		37:15

The top row shows that this student kept the Resource window (the text of the poem) open during the entire session. The remaining rows show his sequence in moving from one task to the other and how much time he spent before switching to another window.

Table 2 shows our analysis of the raw data in Table 1, using familiar terms for crucial phases in the writing process—*previewing*, *drafting*, and *revising*.

TABLE 2

Section	Function	Time	Total	Words	Words/min
Resource		37:01	37:01		
Orientation		0:04 0:03	0:07		
Planning	previewing previewing revising	0:07 0:17 1:14	1:42		
Interpretation	previewing drafting revising	2:18 10:16 2:30	15:04	336	22.3
Application	previewing drafting revising revising	0:04 18:44 1:17 0:09	20:14	432	21.3

As indicated by the last column labeled **Words/min**, this student maintained a comparable rate of production while responding to two quite different tasks: the Interpretation task and the Application task. In the case of certain students, the information in this column revealed that even though they spent considerable time on a particular task, they were not able to produce much writing, which allowed us to provide more tailored feedback.

We also discovered the words/minute ratio to be useful when we were evaluating the viability of a particular task during field-testing. If this ratio was consistently low for students, it provided evidence that the task itself was blocking them from showing what they can do and thus needed revision.

Digital Project Model: Assessing Students in Chinese Universities

The digital project model was developed at Teachers College, Columbia University, by Zhang Wei and was subsequently implemented in a popular course known as *Doing English Digital* at Beijing University. The educational goals of this model were first articulated in the William P. Fenn Lectures, *English in China: Educating for a Global Future*, which I presented at Beijing University and twelve other universities throughout the People's Republic of China in 1998. These lectures proposed that the traditional College English course, which is required for all students not majoring in English, should become more digitally oriented by requiring students to use the Internet to conduct research in their major field of study, such as biochemistry, geology, economics, or history.

In carrying out such research, students develop an array of skills in using digital tools:

searching for relevant information (search tools such as Google)

evaluating the information (online lists of reliable resources in an academic discipline)

organizing the information (digital tools such as Inspiration)

making an oral presentation (digital tools such as PowerPoint)

making a written presentation (digital tools integrating text and graphics)[9]

This research experience can play an important role in preparing Chinese students to participate in a global future, since it helps them integrate their knowledge of English as the lingua franca of the Internet with digital skills that are crucial in transnational communication (Hill, 1998).[10]

Structure of *Doing English Digital*

Fundamental to the success of this digital project model is that it is curriculum-embedded and instructionally-oriented. At its core are eleven modules that provide various kinds of scaffolding to support students as they complete their research projects. Each module consists of five components: tasks, guidelines, examples, tools, and readings. These components have been designed to anticipate various problems that students encounter in conducting their research projects. The guidelines provide step-by-step help for completing the tasks in each module. The examples mainly come from previous student work or links to other websites. The tools offer different kinds of procedural and conceptual scaffolding, such as search tools (e.g., keyword searches, topical indexes, search engines), organizing tools (e.g., Inspiration), software for graphic organization, publishing tools (e.g., Adobe), software for Webpage development, and assessment tools such as checklists and evaluation rubrics.

[9] In communicating with mass audiences throughout China, I deliberately used the traditional distinction between oral and written communication. In digital communication, however, this distinction is often blurred, since the oral can be present in a written document (e.g., a video link that the reader can activate) and the written can be present in an oral presentation (e.g., PowerPoint slides that present written text the audience can read). Such hybrid forms of communication have become normative, especially for those who are growing up with hand-held devices that they use in their daily lives.

[10] It is commonly assumed that such communication takes place with a native speaker of English, but it often takes place with another individual who is not a native speaker, especially in Asian countries where native speakers do not abound and English is increasingly used as a lingua franca for transnational communication in the domains of politics, business, and education (Wang & Hill, 2011). It is important to bear in mind that not just in Asia, but throughout the world there are now considerably more non-native speakers of English than native speakers.

Doing English Digital has now been offered seven times at Beijing University and so Zhang Wei makes available on the course website previous research projects, which are categorized according to major topics (see Figure 12). These sample projects are useful in helping students to select a topic in their major field and then review the existing research.

Physics	Psychology	Economics
Education	Law	Literature
Philosophy	Politics	Advertising
Cultural Studies	Women's Studies	Literacy Studies

Figure 12

As students attempt to select a topic, they post potential topics online and provide feedback to each other. Once they have selected a topic and reviewed the relevant research online, they move on to the next stage, in which they post potential ways of organizing their research online and, once again, provide feedback to each other. As they begin to develop a written presentation, they submit an initial draft to Zhang Wei, who provides feedback, but also uses this draft to help authenticate that the final draft is fundamentally the work of the individual student.

During the final stage of the course, students make oral and written presentations that are evaluated according to the rubric that Zhang Wei has adapted from the digital testing model (see Figure 13):

Content	focused controlling ideas	sufficient evidence	credible materials
Clarity	coherent patterns of organization	consistent control of language	effective use of graphics
Critical/Creative Thinking	thoughtful treatment of the topic	fair-minded evaluation of varied perspectives	active interaction with sources

Figure 13

Zhang Wei has added a third subcriterion under each of the major criteria that focuses on digital aspects of the student project. Under *Content*, *credible materials* is used to evaluate whether students have used Internet resources that are well vetted and thus reliable (they are required to include live links in the digital version of their projects so that evaluators can easily go online and check up on the resources they have used). Under *Clarity*, *effective use of graphics* is used to evaluate whether tables and figures are well integrated with the text (students—and not just in China—often laboriously repeat the information in tables and figures in the text itself rather than moving on to provide strategic commentary on what has been graphically presented). Under *Critical/Creative Thinking*, *active interaction with sources* is used to evaluate whether students have reshaped the online material so that it is effectively integrated into their project.

In applying the rubric to student projects, evaluators use three levels of scoring: *excellent*, *good*, *passing* (along with +/– for each level). Two evaluators respond to each project and a third is used when the first two do not agree. In order to insure a stable use of the rubric, Zhang Wei followed a procedure that I developed when working as a consultant for the International Baccalaureate Program. In the first stage, sample student projects are scored holistically by evaluators who are experienced in using the rubric. In the second stage, the rubric is used to conduct detailed textual analyses of projects that exemplify different levels of scoring in order to create an exemplar packet (see Hill, 1998, for a more detailed description of this process).

Zhang Wei then uses the exemplar packet to train evaluators. In her dissertation research (Wei, 2003), the consistency in scoring was especially high for the written presentations ($r = .82$, $p < .01$). Not surprisingly, it was somewhat lower for the oral presentations ($r = .74$, $p < .01$), given the inherent complexity of an oral presentation in which speech must be integrated with visual information on PowerPoint slides.[11]

Zhang Wei also distributes the exemplar packet to students during the early stages of Doing English Digital to help them internalize the rubric. One of the major benefits of this digital project model is that it teaches students to internalize basic criteria that they can use to evaluate their own writing not only in this course but in other courses as well—and, indeed, in their later careers beyond the university.

[11] Zhang Wei and I are currently working on a more detailed rubric for evaluating PowerPoint presentations. This rubric includes criteria for evaluating (1) the construction of slides—for example, whether the use of language is appropriate (i.e., key words and phrases rather than entire propositions), and (2) the use of slides—for example, whether the presenter's oral communication, including gestures, is effectively integrated with the visual information available in the slides.

Sample Research Project

To show how Doing English Digital works, I will examine a research project conducted by Yue Yu, a student majoring in psychology at Beijing University. This project takes on an important topic in education—how best to develop moral thinking in children.[12]

His written presentation and the PowerPoint slides for his oral presentation can be found in the Appendix.[13] The written project is relatively short (2,031 words including references) and the number of slides limited—there are only 12, including an introductory title page and a final slide in which he thanks his fellow students for their attention. Zhang Wei deliberately restricts the length of both presentations, since the students she teaches have virtually no experience in extended speaking or writing in English.

Yue Yu, like most of his fellow students, has not spent time in a country where English is spoken and has had relatively limited exposure to the language in his formal education. In China, most students study English for about five hours per week in secondary school, where teachers have traditionally lectured about grammar as a means of controlling large classes. It is thus not surprising that various kinds of infelicities can be found in their initial efforts to use English in spoken and written communication.

Given this context, Yue Yu's use of English to express the sophisticated thinking evidenced in his project is quite remarkable. He received a score of *excellent* on both the oral and written presentations. Let us briefly consider how Zhang Wei (2010) applied the rubric in evaluating his presentations, bearing in mind that her purpose was to develop sample materials for training evaluators. She moves systematically from one criteria to the next, each time focusing on features of the student essay in relation to the subcriteria.[14]

[12] One of the attractive features of Doing English Digital is that students take on topics that are relevant to the larger society. When I last observed the course at Beijing University, a student was exploring how best to approach the topic of comparing Chinese and American university students in their use of social media.

[13] Zhang Wei requires students to include an evaluation of the reliability of their sources. I have included one example of such an evaluation in the Appendix, which can be found after the references.

[14] Her use of the rubric contrasts with the one reported in the digital testing model, where the focus was on providing feedback to the student.

Evaluating the Written Presentation

Content

Focused controlling ideas

The writer[15] begins by observing Chinese elementary education is built around reciting slogans. After reviewing various alternative proposals, he introduces the possibility of using moral dilemma stories as an alternative. He then introduces a theoretical framework as well as practical classroom methods for using these stories. He ends by pointing out various benefits as well as potential difficulties in using these stories.

Sufficient evidence

The writer provides relevant detail to support his ideas: for example, the first 12 lines provide rich documentation of the current methods of teaching moral education.

Credible materials

The writer has used professional websites that provide reliable information.

Clarity

Coherent patterns of organization

The writer has effectively used headings, paragraph markers, and connective phrasing to signal the structure of his essay. Consider, for example, lines 41–64 that discuss the potential benefits of using moral dilemma stories. This section begins with a heading that consists of a rhetorical question: *In what way would moral dilemma discussion benefit Chinese moral instruction*? (This is the first of three rhetorical questions used as parallel subheadings.) In this section, the writer discusses three benefits, each signaled by connective material initiating a new paragraph:

> Line 44: “One of the major benefits…”
>
> Line 53: “Apart from its benefit of autonomous learning…”
>
> Line 58: “Moreover, moral dilemma discussion is…”

Consistent control of language

The writer, despite the occasional awkward locution or grammatical infelicity, maintains a firm control of language at both the macro-level (as indicated by the above examples) and the micro-

[15] In discussing the application of the rubric, Zhang Wei uses the generic phrase “the writer” rather than a personal name, since anonymity is preserved throughout the evaluation process.

level (as indicated by a consistent use of vocabulary items such as “autonomous” and “moral reasoning” that maintain an appropriate register for the topic.

Effective use of graphics

The writer constructs a table that succinctly presents Kohlberg’s developmental stages of moral reasoning: the first two are presented in a column labeled ‘Preconventional,’ the second two in a column labeled ‘Conventional,’ and the final two in a column labeled ‘Postconventional.’ He also constructs a flow chart that illustrates how a moral dilemma story can be effectively presented in an elementary classroom. In the case of these graphics, the text that follows does not laboriously recycle the information presented, but rather moves on to provide strategic commentary.

Critical/Creative Thinking

Thoughtful treatment of the topic

The writer takes on an important topic—the moral education of children—and begins by pointing out that the approach in Chinese elementary schools has not been sufficiently thoughtful. He immediately engages his fellow students by providing vivid examples of the slogans that they were forced to repeat as children in primary school.

Fair-minded evaluation of varied perspectives

The writer reviews various approaches to moral education by both Chinese and Western scholars, but ends up recommending the one that he thinks would be most appropriate for children. After outlining benefits that this approach could bring, he is careful to address problems that are likely to arise if it is implemented in Chinese elementary schools. In addressing these problems, he points out the importance of adjusting the approach so that it is more congruent with Chinese cultural norms (i.e., children expect their teachers to provide authoritative opinions).

Active interaction with sources

The writer is willing to think critically about the Western model and adjust it in the light of what Chinese children expect from a teacher. This pragmatic spirit is present throughout the essay: the writer provides little direct quotation but rather rethinks the source material so that it fits the particular topic under consideration.

Evaluating the Oral Presentation

It is difficult to apply the rubric to an oral presentation based on PowerPoint slides. As we are all painfully aware, such slides are often deficient in both design and use. It is not uncommon that a presenter simply reads lengthy propositions crowded onto a slide while the audience squirms impatiently.

Yue Yu managed to avoid both kinds of problems. If you turn to the Appendix, you will see that his slides are parsimoniously constructed. Consider, for example, slide 2 on page 45. He uses the single word 'Outline' as a title and then provides short phrases—or merely a single word in the case of 'Introduction' and 'Conclusion'—to describe the four sections of his presentation. He uses short questions to further break down the third section, which is the heart of his presentation (and he avoids the further use of bullet points for these subsections).

This spirit of parsimony is also evidenced in the slides that Yue Yu constructed to guide his presentation of the third section. Consider, for example, slide 9 on page 46, which deals with the potential benefits of using moral dilemma stories. Under the bulleted heading 'In what ways would it help?' he lists three short phrases to guide his presentation:

autonomous learning
easy to accept
pertinent to real-life issues

When speaking, he used these phrases as mnemonic devices to cue both himself and his audience as he moved through the discussion of potential benefits (a strategic use of a pointer reinforced the power of these visual cues). Their mere presence on the slide was a signal to the audience that he had carefully planned his presentation and was prepared to speak extemporaneously. Given this greater freedom, he was able to maintain eye contact with his audience instead of looking down at a text.

Concluding Reflections

I have examined in some detail a digital testing model designed for American high school students and a digital project model designed for Chinese university students. I would now like to highlight certain features of these models that show particular promise for the future. I would then like to propose that these two kinds of models are best viewed as complementary and hence should be integrated into a more comprehensive model, which will be described as a *digital* *assessment model*, that can be used not only to support classroom teaching and learning but also certify high school students and select them for further educational opportunities.

Digital Testing Model

The digital testing model is not limited to traditional print literacy, but rather provides students the resources—print, sound, image, and animation—that they are accustomed to working with in a digital age. Students are provided a range of tools that allow them to work efficiently with these resources: for example, they can copy and paste material from film as well as text, or they can conduct a search for crucial material and rapidly assemble it in a strategic database. Hence this model reflects greater authenticity, since it allows students to engage in the kind of work that they ordinarily do when using a computer.

This model also has the virtue of presenting students with a set of integrated tasks. The planning task provides grounding in the resources that students draw on in responding to constructivist tasks: first an interpretation task in which they respond critically to the resources and then an application task in which they place the resources in a broader context. In responding to these two tasks, students work with digital tools as well: a notepad for planning what they will write, a live word counter for monitoring how much they are writing, and a spell checker for correcting typos and misspellings.

As students revise what they have written, they have access to familiar tools: for example, cutting and pasting allows for material to be reordered easily. As one student pointed out, when she comes up with a good idea, she simply writes it out and keeps it "at the front of what I am writing" so that she can draw on it when an appropriate context emerges. She also observed that if her fingers are not on a keyboard, she is not able "to do any real thinking and get any words flowing onto the page." These words force us to consider whether assessment is fair when it requires students to respond in handwriting, thus depriving them of the tools they are accustomed to using. Of course the question of fairness is confounded by the fact that within our multicultural society students vary considerably in the degree to which they have access to computers. As we move more deeply into the digital age, this issue will become more prominent.

Finally, I would like to call attention to the process log, which allows us to analyze how students spend their time as they work with the resources and tasks. For example, we can determine whether they initially preview all three tasks, whether they use digital tools efficiently to assemble a database, and whether they spend sufficient time drafting and then revising their responses**.** Thus, the process log allows us to highlight time management along with *Content*, *Clarity*, and *Critical/Creative Thinking* in the feedback that we provide. The challenge we face in developing a digital testing model is to preserve broad values while providing students insights into how well they manage digital resources and tools.

Digital Project Model

The digital project model also leads to student work that is characterized by greater authenticity. In a digital age, using the Internet to find information about a particular topic is an essential activity. The course Doing English Digital is set up to teach students a comprehensive set of skills that they can use to find information online and then communicate it effectively.

This greater authenticity is reinforced by the social interaction that students engage in as they develop their research projects. During the early stages—identifying the topic, finding appropriate resources, developing a coherent plan for the presentation—they interact with each other and their teacher not only face-to-face but also through the course website. Once they begin to write what they plan to present, they continue to interact with the teacher, who provides feedback on early drafts.

Rubric Design

I would like to call attention to the rubric used in Doing English Digital to evaluate student work and provide feedback. It is adapted, as previously noted, from the one used in the digital testing model, which, in turn, was adapted from a rubric built for the International Baccalaureate. This rubric has the virtue of focusing on important values in writing while avoiding excessive detail. In using simple terms to identify three broad areas—*Content*, *Clarity*, *Critical/Creative Thinking*—it sets up a framework that is easy for evaluators to use and for students to internalize.

As the standards movement has developed in this country, rubrics have become increasingly complicated writing has come to play an important role in testing at both the state level (e.g., the New York State Tests used for certification) and the national level (e.g., the SAT used for selection). Unfortunately, this movement has spawned rubrics that reflect widely circulated standards that have value in the larger educational enterprise but are inappropriate for evaluating the kind of writing that can be done in a testing situation.

State Level. Inappropriately inflated rubrics are especially noticeable in tests designed for children at the state level. Figure 14 presents the rubric used to evaluate fourth graders' writing on the English Language Arts Test in New York State (2002). The first column lists general qualities, while the second provides descriptions of how these qualities are manifested in responses that receive the highest score (level 4).

Quality	Responses at Level 4
Meaning: The extent to which the response exhibits understanding and interpretation of the task and text(s)	*Taken as a whole:* • fulfill all or most requirements of the tasks • address the theme or key elements of the text • show an insightful interpretation of the text • make connections beyond the text
Development: The extent to which ideas are elaborated, using specific and relevant evidence from the text(s)	*Taken as a whole:* • develop ideas fully with thorough elaboration • make effective use of relevant and accurate examples from the text
Organization: The extent to which the response exhibits direction, shape, and coherence	*The extended response:* • establishes and maintains a clear focus • shows a logical sequence of ideas through the use of appropriate transitions or other devices
Language Use: The extent to which the response reveals an awareness of audience and purpose through effective use of words, sentence structure, and sentence variety	*The extended response:* • is fluent and easy to read, with vivid language and a sense of engagement and voice • is stylistically sophisticated, using varied sentence structure and challenging vocabulary

Figure 14

In a review of the New York State Test for fourth graders, I called attention to the mismatch between the criteria found in the rubric and the writing that children are able to do in the particular conditions that the test affords.

> Consider, for example, such descriptions of language use as "is fluent and easy to read, with vivid language and a sense of engagement or voice" and "is stylistically sophisticated, using varied sentence structure and challenging vocabulary." In state education departments throughout the country, phrases like these have been recycled in rubrics used to evaluate what children write on language arts tests. It is disconcerting that standards associated with the highly edited work of seasoned adult writers, working on familiar material over months or even years, is being applied to what children, working under the pressure of a high-stakes test, manage to get on the page when they have about 15 minutes to respond to three tasks about a story that they have just heard for the first time. (Hill, 2004, 1099–1101).[16]

[16] Since retirement, I have been mentoring a child in Harlem through a program for children who have a parent in prison. In helping him prepare for this test, I discovered that he ended up limiting his written responses because the large-size printing that he likes to use when writing for official purposes would extend well beyond the text

National Level. Since the SAT began to evaluate student writing in 2005, it has received a good deal of criticism for the approach it is using. The writing section includes not only a written essay but also multiple-choice tasks that require students to identify errors, complete sentences, and improve paragraphs. The total score is heavily weighted toward the multiple-choice component (75%).

The scoring of the essays is based on a rubric and carried out by two trained readers. If their scores differ by more than one point on a 6-point scale, a senior reader is called in to assign the final score. Since the essays are quite short (only 25 minutes is allowed to read the prompt and write a response), they can be rapidly scored (the average time used to score an essay is 3 minutes).

This brief writing and rapid scoring has led to fundamental questions about the value of including this kind of writing task on the SAT. Les Pearlman, who directs undergraduate writing at MIT, found that the length of the essay strongly correlates with the assigned score ($r < .9$): the shortest essays (about 100 words) tend to receive the lowest scores and the longest essays (about 400 words) the highest scores (Winerip, 2005).[17]

Pearlman also questions the official policy of ignoring factual errors when evaluating the essays. He argues that a key feature of undergraduate education at MIT is instilling in students a respect for the accurate use of factual information. From Pearlman's perspective, such respect is fundamental to the development of scientific thinking.[18]

boxes provided for answers. In explaining why his responses were so short, he said that he was afraid he would "lose points if his writing goes outside the box." When doing homework with him, I had discovered that his teacher subtracted points whenever his responses did not fit into the text boxes provided in his workbook.

I should further note that even if his printing had been small, the text boxes on the test were generally too small to accommodate the information called for by the tasks. If this child were to take the test on a computer—he is quite comfortable on a computer because of his love of videogames—the text box would automatically expand to accommodate whatever he writes.

[17] As far as I can ascertain, Pearlman did not carry out a multifactor analysis. I suspect that features such as effective sequencing of arguments would correlate positively with essay length. After all, one needs a certain amount of textual space in order to develop effective argumentation.

[18] Under Pearlman's guidance, an undergraduate at MIT took the SAT and wrote an essay about Franklin Delano Roosevelt and the Depression that was deliberately filled with factual errors. His essay was, however, the desired

Digital Assessment Model

Given the problems that attend evaluation of student writing in a testing situation, I recommend the development of a comprehensive model that includes a project component as well as a testing component. The term *digital assessment model* can be used to refer to this more balanced approach, which maintains the positive benefits associated with traditional testing while allowing for a more responsible appraisal of student writing. As illustrated by Doing English Digital, when students are allowed a more extended time frame and provided scaffolding that supports the writing process, they are able to produce writing that can be evaluated fairly with rigorous standards.

An extended time frame does allow for the possibility of a student receiving help from others, which the traditional approach to testing is designed to prohibit. Doing English Digital is designed so that students receive responses to their writing not only from the instructor but also from other students in the course. From the vantage point of Zhang Wei, these responses are fundamental to what goes on in writing projects in the real world, and hence assessment should take account of the degree to which an individual student can make effective use of feedback. At the same time, Zhang Wei relies on the various drafts that individual students produce as a means of verifying that the final draft is essentially their own work. Given that these drafts are digitally stored, she is able to examine them closely to detect both the changes that signal an effective response to feedback and the continuities that signal the authentic voice of an individual writer.[19]

Once an extended time frame is introduced, the cost of evaluating student writing increases dramatically. As already observed, the average time for a reader to evaluate the relatively short

length and contained vocabulary items such as “plethora” that are used in essays that receive a high score. His essay received a score of 10: the maximum number of points is 12, since each rater can assign up to 6 points (Jaschik, 2007). I should note that in the rubric developed for the International Baccalaureate and adapted for both the digital testing model and the digital project model, *accuracy* is included as a subcriterion under *Content*.

[19] Digital technologies will increasingly be able to authenticate the work of individual students by analyzing samples of their writing for stylistic features. But even if such technologies are perfected, collecting valid samples for individual students might turn out to be too difficult and hence the judgment of thoughtful readers will, no doubt, still be needed.

written essay on the SAT is three minutes. The sample essay by Yue Yu is 2,031 words, which is about five times longer than the lengthier essays written for the SAT.

In the future, automated scoring will play an increasingly prominent role in holding down cost. As Bennett (2011) has observed, automated scoring is built around surface features, ranging from spelling and punctuation to discourse markers of cohesion, which are markedly different from the features represented in the rubrics that we have been discussing. As the field of artificial intelligence continues to develop, we can anticipate increasingly reliable ways of using surface features as indices to deeper levels of structure. It is important to bear in mind that in our ordinary acts of reading we have access to deeper levels of structure through a judicious sampling of surface features.[20]

As Bennett further points out, it is difficult to know just how trained evaluators make use of a rubric when evaluating student writing. Anecdotal evidence suggests that they are able to use a complex array of sampling techniques to arrive at holistic judgments that are not the result of a mechanical application of the rubric (although these judgments can still be reasonably consistent with its values). Bennett makes a tantalizing suggestion that I plan to test out in a forthcoming research project: use two independent systems of automated scoring and bring in a human rater only when they disagree.[21]

[20] Automated scoring of student writing is called for in the consortium *Partnership for Readiness in College and Career*s (PARCC), one of the two major consortia that are being funded by the U.S. Department of Education to develop assessment systems for the Common Core Standards. According to knowledgeable sources, the complexity of developing reliable automated scoring systems has been considerably underestimated, and it will be difficult to meet the deadline set for the academic year 2014-15.

[21] I will collaborate with Wang Haixiao, who chairs the Department of Applied Foreign Languages at Nanjing University, on a research project in which we will use both China-based automated scoring and Western-based automated scoring to evaluate research projects that were scored by the rubric in the Doing English Digital course at Beijing University. We hypothesize that the two methods of automated scoring will produce scores more similar than those based on the rubric, given that they both are oriented toward a presumably comparable set of surface features. The leading vendor of automated scoring in China, like many vendors in this country, has not made public the system it uses. As Bennett points out, the development of automated scoring is handicapped by the widespread lack of transparency.

Digital Archives

A comprehensive digital assessment model could be built around an archival system used to carry out the basic functions of certification and selection. Given the capacity of digital technologies to efficiently store and retrieve information, constructing a unified system at the national level is technically feasible at the present time. Before one can be put in place, however, formidable political obstacles, most notably those having to do with the strong tradition of states' rights, will have to be overcome. The movement to adopt Common Core Standards is clearly a step toward developing a more unified system, and the fact that 43 states have already approved these standards provides grounds for cautious optimism.

How might digital archives for individual students be constructed so as to strengthen the relations between assessment and classroom teaching and learning? Let us consider, in turn, the testing component and the project component.

Testing Component. This component would consist of carefully designed classroom activities carried out under the strict conditions associated with testing. In order to insure comparability, these activities could be based on certain strands in a common curriculum, perhaps those in American history that have to do with developing civic responsibilities (see the Common Core Standards for material that would be widely accepted across the political spectrum and thus not be opposed on partisan grounds).

These classroom activities would take place on a regular basis throughout the academic year and all the student work would be digitally archived. For the purposes of accountability, there would be externally appointed evaluators, working with classroom teachers, who would use methods of random sampling to evaluate student responses to a limited number of classroom activities in selected areas of the curriculum.

Since these classroom activities would take place on a regular basis, students would be less likely to think of them as tests, especially since teachers would use process logs to provide helpful feedback on matters such as time management and the writing process. In effect, everyday activities of the classroom would come to function as tests only as they are selected by a process of random sampling.

Given the demands of accountability that accompany any assessment system, this testing component would necessarily carry substantial weight. The fact that this component would be digitally administered could play an important role in accountability: for example, real-time records would be available in the digital archives and could be used to document that the strict time limits associated with testing are properly observed.

Project Component. Here, too, random sampling methods would be used to select samples of student work from the digital archives that would be evaluated by the team of external evaluators and classroom teachers. Since the number of projects that an individual student can carry out is limited, the team would evaluate only one or two projects for each student in a given subject matter (e.g., English Language Arts). The system might be designed to allow students to select one project that they would submit to the evaluation team. They would submit not only the project, but also a statement that explains why they have selected it as representing their best work.

In any evaluation of student projects, it is imperative that teachers be included so they can deepen their experience in using the rubric and reinforce its standards in their daily interactions with students. The ultimate goal of any assessment system should be to insure that teachers continuously circulate high standards in the classroom so that students bring them to bear on all the work that they do, not only in school but also in the larger society.[22]

Certification. There are different ways in which a state education agency could use information based on digital archives in granting students a high school diploma. For purposes of efficiency, it could use only a numerical score based on the testing component if it is sufficiently high. Hence, information from the project component would be introduced only when the score from the testing component is marginal and needs to be supplemented.

Another approach would be to use information from both components simultaneously, with the possibility that the state agency might provide greater weight to one of the components (presumably the testing component). Ultimately, decisions about the use of archival data would depend upon policy decisions at the state level. Given the resistance to centralized authority in this country, it is important that as much autonomy as possible be maintained at the state level.

[22] As Linda Darling-Hammond (2004, 2010) has observed, as assessment activities become integrated into the everyday classroom, they come to play an important role in the professional development of teachers. See Bennett (2010) for discussion of the ways in which the CBAL project that is underway at ETS contributes to teacher development.

Selection. Admissions offices in American colleges and universities would develop individual policies about how to use information from digital archives. Since these offices are already committed to using samples of student writing, they would welcome the opportunity to receive randomly selected student writing evaluated within the project component. Such writing could be supplemented by a sample that individual students select from digital archives and submit along with a statement of why they value this particular writing. As for the quantitative component of an admissions dossier, the score generated by the testing component of the digital assessment model could be used in place of an SAT or ACT score.

A final thought—a comprehensive digital assessment model, if properly designed and administered, could lead to a diminished use—or even a gradual withering away—of externally mandated tests based on the multiple-choice format. Such a change could lead to greater integrity in the American classroom: It would free students from the debilitating anxieties they often experience in preparing for these tests and teachers from the burden of devoting an inordinate amount of class time to test-prep activities.

As the larger world explores the ways digital technologies can transform educational assessment, countries such as France and China with a strong central government are in a position to act boldly and use digital technologies to create systems that more effectively integrate assessment with teaching and learning. Despite all the political obstacles to developing an integrated assessment system in this country, the Gordon Commission could fulfill its mandate on the future of assessment by articulating a vision in which all the information to be used in evaluating students is generated in carefully designed classroom activities that are stored in digital archives. This country is now entertaining proposals to build digital archives for personal health information, so why not formulate a proposal for a comprehensive assessment model that would use digital archives in certifying high school students and selecting them for further educational opportunities?

Appendix

Applying Moral Dilemma Discussion in Chinese Elementary Classroom

Yue Yu
Yuanpei Pilot Program and Department of Psychology
Peking University

"Love the motherland, the people, labor, and science and take good care of public property," this is the famous "Five Love" slogan used in Chinese elementary school. In a typical moral education class, pupils would echo slogans like this again and again, sometimes with the words go in from one ear and out of another. It is natural to ask the question when viewing such scenarios: is this kind of instruction effective?

In fact, researchers have raised the same question. Current Chinese moral education in elementary school mainly teaches social moral concepts as the "Five Love", and behavior norm such as "respect teachers" and "follow the disciplines" (Xia, et al., 2005). Patriotism and collectivism are highly emphasized while little concern has been put on pupil's character and moral reasoning. As for teaching method, the dominant classroom instruction is exhortation, which is sometimes referred as "bag of virtues" or "values clarification" (Zhang, 2002; Xia, et al., 2005). This kind of instruction has shaped pupils who can only recite slogans without knowing how to apply them, and school is disconnected with family and community education (Yang, 1997). As Zhang (2002) pointed out, Chinese moral education is facing quandaries.

Former researchers have given various suggestions to improve Chinese moral education. These include changing virtual-situation education into real-situation education (Fu, 2005), paying more attention to inherent moral culture (Liao, 2000), corresponding the instruction with pupils' moral reality (Zhang, 2002), and combining school education with family and community education (Yang, 1997). However, these concepts seem too theoretical to be taken into classroom practice. In this paper, I would introduce a distinctive teaching approach used in western countries—moral dilemma discussion, and discussed its possible appliance in Chinese elementary classroom.

Moral Dilemma Discussion

Moral dilemma discussion approach, or New-Socratic approach, is a teaching technique derived from Kohlberg's theory of moral development (Kohlberg, 1981). According to his framework, the life-long moral development can be divided into six stages (see table 1), each representing a unique type of thinking defined by how we process moral-ethical and value questions. Higher stages of moral development take account of broader perspective, represent more complex and abstract thought, contain more personal empathy, and provide more principle-based solutions to social problems.

Table 1: *Kohlberg's Stages of Moral Reasoning*

I. Preconventional Level	**II. Conventional Level**	**III. Postconventional Level**
Stage 1: Punishment and Obedience Orientation. Physical consequences of action determine its goodness or badness.	**Stage 3: "Good Boy-Good Girl" Orientation.** Good behavior is whatever pleases or helps others and is approved of by them. One earns approval by being "nice."	**Stage 5: Social Contract Orientation.** What is right is defined in terms of general individual rights and in terms of standards that have been agreed on by the whole society.
Stage 2: Instrumental Relativist Orientation. What is right is whatever satisfies one's own needs and occasionally the needs of others. Elements of fairness and reciprocity are present, but they are mostly interpreted in a "you scratch my back, I'll scratch yours" fashion. Individual adopts rules and will sometimes subordinate own needs to those of the group.	**Stage 4: "Law and Order" Orientation.** Right is doing one's duty, showing respect for authority and maintaining the given social order for its own sake. People define own value in terms of ethical principles they have chosen to follow.	**Stage 6: Universal Ethical Principle Orientation.** What is right is defined by decision of conscience according to self-chosen ethical principles. These principles are abstract and ethical, not specific moral prescriptions.

Note. Adapted from "*Educational Psychology: Theory and Practice (7th ed.)*" by R. E. Slavin, 2003, Boston: Pearson Education. p. 55.

His framework emerged from interviews with his research subjects using moral dilemma stories, which describe troublesome moral situations.

Kohlberg's story-telling approach has been used in the moral classroom to increase the level of the pupil's moral maturity (Kohlberg, 1981; Berkowitz, 1984). Refined by Berkowitz (1984), such discussion typically contains three successive phases: starting, continuing, and reaching closure, as illustrated in Figure 1. During discussion teacher should figure out the current moral status of individual pupil and treat them respectively, and they should also hold supportive view to pupils' discussion (Berkowitz, 1984). The application of dilemma-discussion approach in the instructional practices of moral education proves to be consistently effective for children's moral development (Thoma, 1984). For example, in an experiment servicing institutionalized delinquent and predelinquent students, moral discussion group showed significantly higher impact on moral reasoning compared to values clarification group and control group (Niles, 1985). In fact, there appeared many successful projects of dilemma-discussion approach all over the U.S. since 1980 (Slavin, 2003).

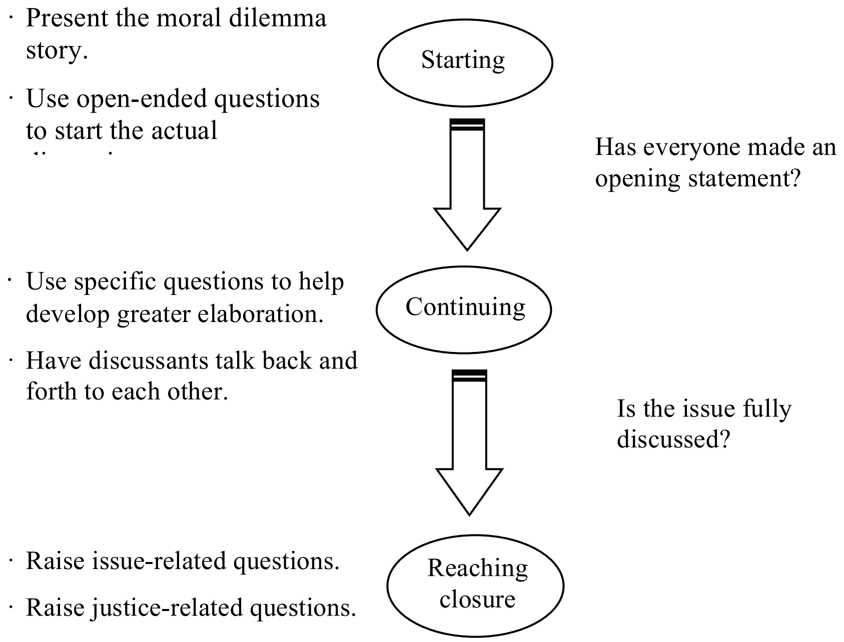

· Present the moral dilemma story.
· Use open-ended questions to start the actual
Starting
Has everyone made an opening statement?
· Use specific questions to help develop greater elaboration.
· Have discussants talk back and forth to each other.
Continuing
Is the issue fully discussed?
· Raise issue-related questions.
· Raise justice-related questions.
Reaching closure

Figure 1. Leading moral dilemma discussions

Applying Moral Dilemma Discussion in Chinese Elementary Classroom

In what way would moral dilemma discussion benefit Chinese moral instruction?

Compared with Chinese traditional moral instructions, dilemma-discussion approach has remarkable benefits in its autonomous learning, easy acceptance and pertinence to real-life issues.

One of the major benefits of moral dilemma discussion is that it helps foster moral reasoning and promote autonomous learning of pupils. Kohlberg (1981) has claimed that virtue judgment is just putting a label on certain people or behavior, while moral decision always relies on other thinking process. In fact, the base of real-life moral actions is not right-wrong judgments, but the way to analyze and reason these situations. Experiments showed that students in lower moral reasoning stages cheat in tests four times as much as their classmates in higher stages, while their judgment on propriety of cheating remains the same (Kohlberg, 1981). Moral dilemma discussion focuses on improving children's reasoning skills on moral issues, rather than simply identifying something as right or wrong. Since moral reasoning in Chinese students is less critical and logical (Xie, 2001), this kind of training is especially valuable.

Apart from its benefit of autonomous learning, moral dilemma discussion approach is also easier for pupils to accept. Dilemma-discussion approach is based on the identification of children's current moral developmental stage, and aimed at pushing them forward within their capacity. In contrast, traditional moral education in China often talks about abstract virtue criteria, which are sometimes far beyond the pupil's level of comprehension. And only through understanding can moral judgments be digested and applied to action.

Moreover, moral dilemma discussion is highly pertinent to issues in real life. Materials of moral dilemma can be selected from headline stories from newspapers, everyday incidents, popular moral issues, and incidents from movies or readings. The advantage is two-fold: on the one hand, real issues can arouse much interest, thus increasing the motivation and involvement of pupils; on the other hand, the judgments of real moral situations can be applied to real life behavior more easily. Since students taught by traditional Chinese moral instructions often show dissociation between their moral judgments and moral action (Lu, 2002; Fu, 2005), real issues can bridge principles and reality.

What are probable challenges in the application?

When borrowing a teaching approach from one country to another, cultural differences are of primary concern for its feasibility. As for moral development, literature has revealed a cross-culture character of Kohlberg's stages, which lends support to the usage of moral dilemma in different countries. Snarey (1985) has reported two studies, conducted in Kenyan and Turkish respectively, where moral dilemma material was used with properly transposed contents. The result showed a similar sequence of moral development from the less complex to the more complex levels, and the main cross-cultural difference showed up was between urban and tribal societies rather than western versus nonwestern societies. A later survey in both Taiwan and U.S. also found the important dimensions identical in the response to the dilemmas, regardless of the subjects' language and religion differences (O'Neill, 1998). Hence it is fairly reasonable to assume that Chinese children may go through a similar development in moral maturity.

Although the universality of moral development is guaranteed, other problems might occur in teaching practice due to cultural difference. The role of teacher, for example, is quite different between China and western countries. According to dilemma-discussion approach, teachers should withhold their own views, and let students reason by themselves (Berkowitz, 1984). Whether this point is suitable for Chinese moral education is open for discussions. Since moral education in China is often combined with ideological and legal education (Zhang, 2002), it bears responsibility of setting up social convention and political standpoint for pupils. Moral dilemma discussion itself cannot achieve all of the aims, exhortation and illustration of objective moral standard is still needed for children's civic development.

What are practical suggestions in applying moral dilemma discussion?

Classroom discussion is the core of dilemma-discussion approach. With concern to the reality of China, I suggest that moral dilemma discussions replace some (not all) of the ideological and moral classes in higher grades of elementary school. This grade period is chosen because children move from preconventional morality stage into conventional stage between age 9 and age 12 (Kohlberg, 1981; refer to table 1), so it is critical time for their moral development. During a discussion class, the material of moral dilemma can be searched from many ways as mentioned, and an atmosphere should be established that every pupil can raise his or her opinion no matter it is right or wrong. In the closure part, slightly different from Kohlberg's view, I think the teacher should give an instructive answer to the dilemma, which represents the stage of moral growth that is a little higher than the pupils'. This kind of combination between dilemma discussion and values clarification can better fit the multi-aimed moral class in China.

Besides classroom instruction, another scenario for moral dilemma discussion can be peer interaction. According to William Damon (1988), children at elementary-school age would discuss concepts of fairness and justice when confronted with difficult group decisions, which would enable them “to become a full citizen in society, with all of a citizen’s leadership prerogatives and fellowship responsibilities” (p. 86). Hence, moral dilemma discussion in peer group should be encouraged though establishing a supportive and democratic campus atmosphere.

Conclusion

What is the most effective way to cultivate virtuous citizens? This question is widely concerned and long debated in China. In this paper I suggest applying a western education approach—moral dilemma discussion—to moral education system in China after reasonable modifications. The dilemma-discussion approach emphasizes autonomous learning compared to traditional approach, and is easier to accept and more pertinent to real-life issues. Moral dilemma discussion has been proved to be effective in fostering children’s moral reasoning, and is widely used in western countries. With evidence of cross-cultural character of moral development, it’s reasonable to anticipate that dilemma-discussion approach would improve moral education in Chinese elementary school. Together with traditional value clarification, this new approach would broaden and vivify moral instruction in elementary classroom, thus fostering pupils’ internal moral development in a more effective way. A further step can be combining moral dilemma discussion in school with pupils’ behavior in family and community, so as to establish an all-around moral educational system to produce righteous citizens.

References

Berkowitz, M. (1984). Process analysis and the future of moral education. Paper presented at the annual meeting of American Educational Research Association, New Orleans, April 23.

Damon, W. (1988). *The moral child.* New York: Free Press.

Fu, W. (2005). Moral conflict of authenticity and students' moral development. *Educational Research, 12(3),* 13-16. Retrieved March 2007, from *CNKI* database.

Kohlberg, L. (1981). *Essays on moral development, Volume I & II.* New York: Harper and Row.

Niles, W. J. (1985). Effects of a moral development discussion droup on delinquent and predelinquent boys. *Journal of Counseling Psychology, 33(1),* 45-51. Retrieved May 2007, from *Sciencedirect* database.

O'Neill, P. & Petrinovich, L. (1998). A preliminary cross-cultural study of moral intuitions. *Evolution and Human Behavior, 19,* 349-367. Retrieved March 2007, from *Sciencedirect* database.

Slavin, R. E. (2003). *Educational Psychology: Theory and Practice (7th ed.).* Boston: Pearson Education.

Snarey, J. (1985). Cross-cultural universality of social-moral development. *Psychological Bulletin, 97(2),* 202-232.

Thoma, S. (1984). Do moral education programs facilitate moral judgment? A meta-analysis of studies using the Defining Issues Test. *Moral Education Forum, 9(4),* 20-25.

Bibliographical information & Source evaluation

Article 1

Search strategy: searched on "ScienceDirect" using "Moral dilemma" & "education" as key words.

Bibliographical information:

Author name: O, Neill Patricia; Petrinovich, Lewis

Date of publication: 1998

Title of the article: A Preliminary Cross-Cultural Study of Moral Intuitions

Name of the journal: *Evolution and Human Behavior*

Name of the publisher/database: Elsevier

Date of retrieval: 07.3.29

URL: http://www.sciencedirect.com/science/article/B6T6H-3VB35PS-1/2/807ae734f67d388cf2bb4c334db0116e

Source evaluation:

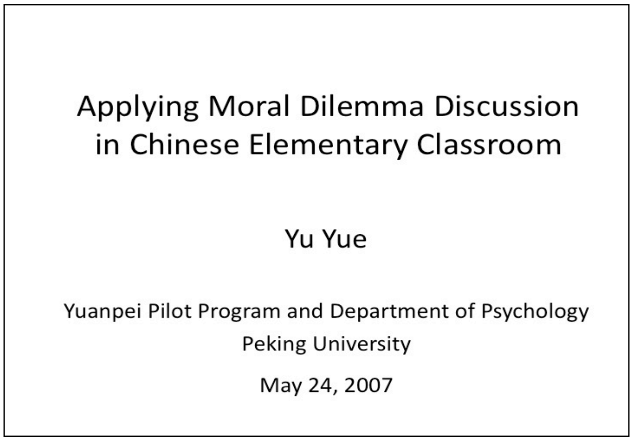
Relevance: The article discusses the universality of moral beliefs and is thus relevant to my topic.

Authority: As I've never heard of the journal *Evolution and Human Behavior* nor the authors, it is not easy to justify the authority of this article.

Accuracy: The article is supported by abundant theoretical and empirical evidence and is thus reliable.

Currency: The article is a little behind the times since it's published in 1998.

PowerPoint Slides Used in Yue Yu's Oral Presentation

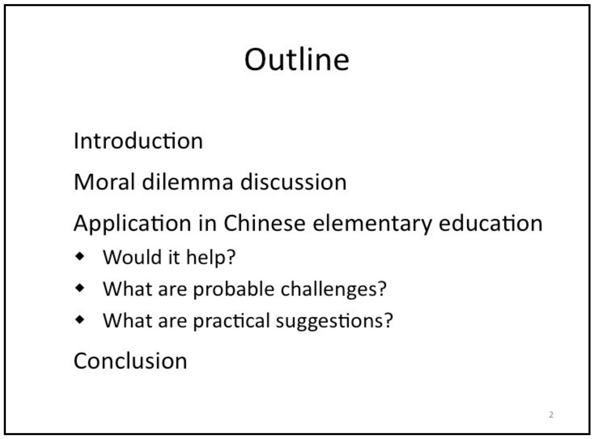


Applying Moral Dilemma Discussion in Chinese Elementary Classroom

Yu Yue

Yuanpei Pilot Program and Department of Psychology
Peking University

May 24, 2007



Outline

Introduction

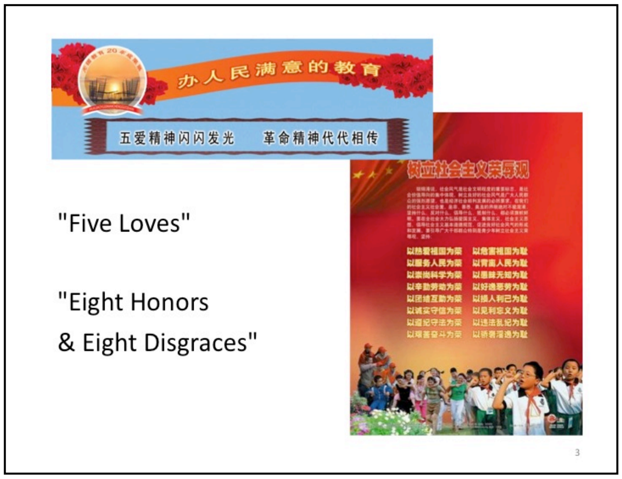
Moral dilemma discussion

Application in Chinese elementary education

- Would it help?
- What are probable challenges?
- What are practical suggestions?

Conclusion

2



办人民满意的教育
五爱精神闪闪发光 革命精神代代相传
树立社会主义荣辱观
以热爱祖国为荣 以危害祖国为耻
以服务人民为荣 以背离人民为耻
以崇尚科学为荣 以愚昧无知为耻
以辛勤劳动为荣 以好逸恶劳为耻
以团结互助为荣 以损人利己为耻
以诚实守信为荣 以见利忘义为耻
以遵纪守法为荣 以违法乱纪为耻
以艰苦奋斗为荣 以骄奢淫逸为耻

"Five Loves"

"Eight Honors
& Eight Disgraces"

3

Introduction

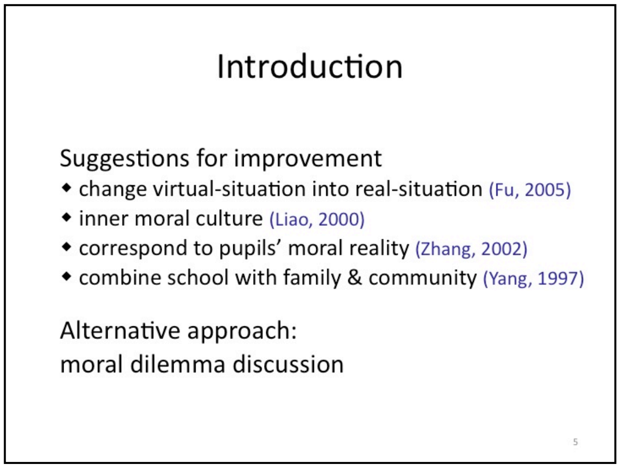
Current moral education in China

- social moral concepts & behavior norm
- patriotism & collectivism
- "bag of virtues" (Zhang, 2002; Xia, et al., 2005)

Is it effective?

- slogans vs. behavior
- school vs. family and community (Yang, 1997)

4



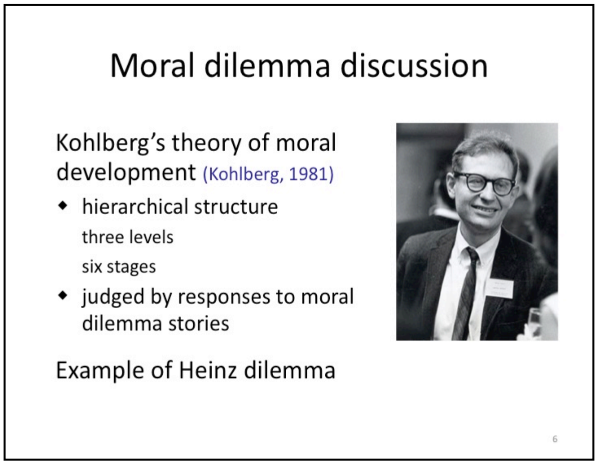
Introduction

Suggestions for improvement

- change virtual-situation into real-situation (Fu, 2005)
- inner moral culture (Liao, 2000)
- correspond to pupils' moral reality (Zhang, 2002)
- combine school with family & community (Yang, 1997)

Alternative approach:
moral dilemma discussion

5



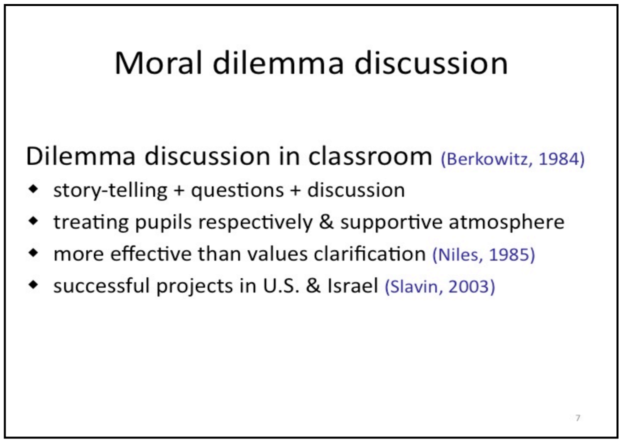
Moral dilemma discussion

Kohlberg's theory of moral development (Kohlberg, 1981)

- hierarchical structure
  three levels
  six stages
- judged by responses to moral dilemma stories

Example of Heinz dilemma

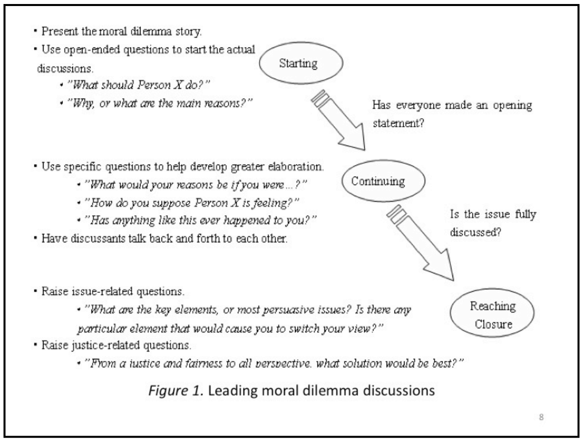
6



Moral dilemma discussion

Dilemma discussion in classroom (Berkowitz, 1984)

- story-telling + questions + discussion
- treating pupils respectively & supportive atmosphere
- more effective than values clarification (Niles, 1985)
- successful projects in U.S. & Israel (Slavin, 2003)



- Present the moral dilemma story.
- Use open-ended questions to start the actual discussions.
  - *"What should Person X do?"*
  - *"Why, or what are the main reasons?"*
- Use specific questions to help develop greater elaboration.
  - *"What would your reasons be if you were…?"*
  - *"How do you suppose Person X is feeling?"*
  - *"Has anything like this ever happened to you?"*
- Have discussants talk back and forth to each other.
- Raise issue-related questions.
  - *"What are the key elements, or most persuasive issues? Is there any particular element that would cause you to switch your view?"*
- Raise justice-related questions.
  - *"From a justice and fairness to all perspective, what solution would be best?"*

Starting
Has everyone made an opening statement?
Continuing
Is the issue fully discussed?
Reaching Closure

Figure 1. Leading moral dilemma discussions

Applying in China

In what ways would it help?

- autonomous learning
  right/wrong judgement => moral reasoning
- easy to accept
- pertinent to real-life issues
  bridge principles and reality

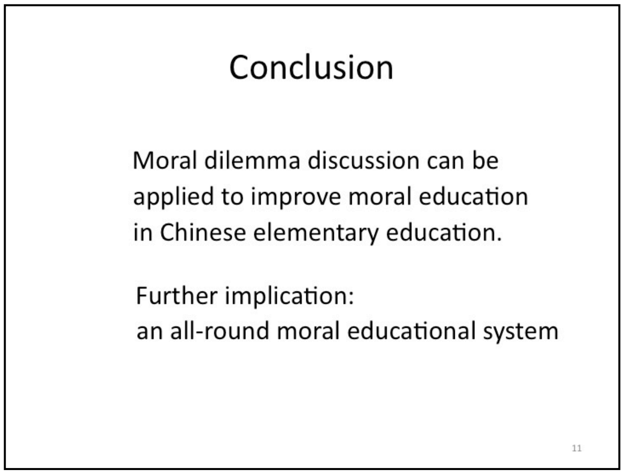
Applying in China

What are probable challenges?

- universality in moral development
- cultural difference
- role of teacher

What are practical suggestions?

- classroom instruction
- peer interaction



Conclusion

Moral dilemma discussion can be
applied to improve moral education
in Chinese elementary education.

Further implication:
an all-round moral educational system

Thanks for your attention!

Acknowledgments

I am grateful to Ed Gordon for the invitation to reflect upon how educational assessment can best serve teaching and learning. From the outset of my career at Teachers College, Columbia University, Ed encouraged me to take on formidable challenges and was generous in providing financial and, more importantly, intellectual support. I am fortunate to have had such a wise mentor throughout the years.

I am also grateful to Eric Larsen, a colleague at Teachers College, Columbia University, for helping me put together this document. He, like Ed, has been a steady influence as I have attempted to make sense of various approaches to assessing students in this country as well as abroad. He has been especially helpful in encouraging me to draw on linguistics and its allied discipline discourse analysis in deconstructing what goes on in the peculiar genre of multiple-choice testing that has such power over our lives.

I would like to thank E. Wyatt Gordon and Maralin Roffino for their help in shaping this document. They have shown patience in communicating with me during the final stages of its preparation while I was in Japan.

References

Adames, J. (1987). *A study of the pre-reading process of selected English as a second language college students* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Aronowitz, R. (1984). Reading tests as texts. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 43–62). Norwood, NJ: Ablex.

Baker, E. (2008). Learning and assessment in an accountability context. In K. E. Ryan & L. A. Shepard (Eds.), *The future of educational accountability* (pp. 277–291). New York: Routledge.

Baker, E. (2009). The influence of learning research on the design and use of assessment. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert design of optimal learning environments* (pp. 333–355). New York: Cambridge University Press.

Baldwin, J. (1998). Stranger in a village. *Collected essays* (T. Morrison, Ed.). New York: Library of America.

Bennett, R. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5–12.

Bennett, R. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, 9(5), 1–23.

Bennett, R. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2), 70–91.

Bennett, R. (*2011*). Automated scoring of constructed-response literacy and mathematics items. Available at http://www.acarseries.org/papers

Bennett, R., & Bejar, I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Bhasin, J. (1990). *Main-idea tasks in reading comprehension tests and the responses of bilingual poor comprehenders* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: School of Education, King's College.

Bloom, B. S. (1984). *Taxonomy of educational objectives*. New York: Longman.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, *12*(1). Available at http://epaa.asu.edu/epaa/v12n1

Chu, H. (1993). *Assessing Chinese kindergarten children in New York City* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Clarke-Medura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research in Technology and Education, 42*(3), 309–328.

Coyle, M. (1992). *The New Jersey high school proficiency test in writing: A pragmatic face on an autonomous model* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106,* 1047–1085.

Darling-Hammond, L. (2010). New-generation assessment of common core standards: Moving toward implementation. Available at http://www.acarseries.org/papers

Dewey, J. (1998). *The essential Dewey* (Vols. 1–2, L. A. Hickman & T. Alexander, Eds.). Bloomington: Indiana University Press.

Ehren, M. C. M., & Hatch, T. (forthcoming). Responses of schools to accountability systems using multiple measures: The case of New York City elementary schools.

Engel, S. (2011). Measuring our success. *Teachers College Record* (ID Number 16318). Available at http://www.tcrecord.org

Gee, J. (1992). What is reading? Literacies, discourses, and domination. *Journal of Urban and Cultural Studies*, *2,* 65–77.

Gordon, E. W. (1970). Toward a qualitative approach to assessment. *The Report of the commission on tests*. New York: College Entrance Examination Board.

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 11–32.

Gordon, E. W. (1999*). Education and social justice: A view from the back of the bus.* New York: Teachers College Press.

Hakuta, K. Improving education for all children: Meeting the needs of language minority children. In D. Clark (Ed.), *Education and American youth*. Washington, DC: Aspen Institute.

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hill, C. (1992). *Testing and assessment: An ecological approach* [Inaugural lecture for the Arthur I. Gates Chair in Language and Education]. New York: Teachers College, Columbia University.

Hill, C. (1995). Testing and assessment: An applied linguistics perspective. *Educational Assessment, 2,* 179–212.

Hill, C. (1996). *Exemplar essay project: Theory of knowledge*. International Baccalaureate.

Hill, C. (1998). *English in China: Educating for a global future*. Excerpt available at http://162.105.138.207/webcourse/advanced_english/English%20in%20China.htm

Hill, C. (2000, March 18). Practicing without learning. *The New York Times,* p. A15.

Hill, C. (2001a, December 27). Pitfalls of annual testing. *The Christian Science Monitor,* p. 9.

Hill, C. (2001b). Short-answer questions in College English testing. In *Research on Teaching College English in China* (pp. 172–184). Beijing, China: Beijing University Press.

Hill, C. (2001c). *Linguistic and cultural diversity: A growing challenge to American education.* New York, NY: The College Board.

Hill, C. (2003). Integrating digital tools into a culturally diverse curriculum: An assessment model for the Pacesetter Program. *Teachers College Record, 105,* 278–296.

Hill, C. (2004). Failing to meet the standards: English Language Arts Test for New York State. *Teachers College Record, 106,* 1086–1123.

Hill, C. (2012). Educational research: The challenge of using an academic discipline (reprint of the Lawrence A. Cremin Lecture delivered in 2010). *Teachers College Record, 114*, 1–42.

Hill, C., Black, J., & McClintock, R. (1994) Assessing student understanding and learning in constructivist study environments. In M. R. Simonson, N. Maushak, & K. Abu-Omar (Eds.), *16th annual proceedings of selected presentations at the 1994 national convention of the Association for Educational Communications and Technology.* Washington, DC: AECT.

Hill, C., & Larsen, E. (1992). *Assessment in secondary education: A review of emerging practices*. Berkeley: University of California, National Center for Research in Vocational Education.

Hill, C., & Larsen, E. (2000). *Children and reading tests.* Stamford, CT: Ablex.

Hill, C., & Parry, K. (1988). *Reading assessment: Autonomous and pragmatic models of literacy* (LC Report 88-2). New York: Teachers College, Columbia University, Literacy Center.

Hill, C., & Parry, K. (1989). Autonomous and pragmatic models of literacy: Reading assessment in adult education. *Linguistics and Education, 1,* 233–289.

Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy in reading assessment. *TESOL Quarterly, 26,* 433–461.

Hill, C., & Parry, K. (1994). *From testing to assessment: English as an international language.* Harlow, UK: Longman.

Ingulsrud, J. (1988). *Testing in Japan: A discourse analysis of reading comprehension test items* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Jaschik, S. (March 26, 2005). Fooling the College Board. *Inside Higher Education*.

Johnson, A. (2009). *Objectifying measures: The dominance of high-stakes testing and the politics of schooling.* Philadelphia, PA; Temple University Press.

Jonçich, G. (1968). *The sane pragmatist: A biography of Edward L. Thorndike*. Wesleyan Press: Middletown, CT.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stetcher, B. M. (2000). What do test scores in Texas tell us? Santa Monica, CA: RAND.

Koretz, D., & Barron, S. I. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS) (MR-1014-EDU). Santa Monica, CA: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests. In R. L. Linn (Chair), *The effects of high stakes testing. Symposium conducted at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education.* Chicago, IL.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Martinez, M. (2009). *Learning and cognition: The design of the mind.* Boston, MA: Allyn & Bacon.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge/Falmer.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439–483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.) (2008). *Assessment, equity, and opportunity to learn*. New York: Cambridge University Press.

National Academy of Education. (2009). Standards, assessments, and accountability. In L.

Shepard, J. Hannaway, & E. Baker (Eds), *Education policy white paper.* Washington, DC: Author.

National Research Council. (1999). How people learn: Bridging research and practice. In M. S. Donovan, J. D. Bransford, & J. W. Pellegrino (Eds.), *Commission on behavioral and social sciences and education.* Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaser, R. (Eds). Board on Testing and Assessment, Center for Education. Division of Social Sciences and Education. Washington, DC: National Academy Press.

New York State Testing Program. (2002). *Scoring guide and scorer practice set* (English Language Arts, Grade 4). Albany, NY: New York State Education Department.

Nix, D., & Schwartz, M. (1979). Toward a phenomenology of reading comprehension. In R. Freedle (Ed.), *New directions in discourse processing* (pp. 183–196). Norwood, NJ: Ablex.

Parry, K. (1986). *Readers in context: A study of northern Nigerian students and school certificate texts* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2002). Challenging the validity of automated essay scoring, *Computers in Human Behavior, 18,* 103–134.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Rodriguez, R. (1998). *The browning of America*. Available at http://www.pbs.org/newshour/essays/february98/rodriguez_2–18.html

Russ, J., Ehren, M.C.M., & Lesaux, N. (forthcoming). *Strategies teachers use to coach students to do well on the ELA state test: The case of New York City and Boston public elementary schools.*

Russell, M., & Haney, W. (2000, March 28). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives*, *8*(19). Available at http://epaa.asu.edu/epaa/v8n19.html

Russell, M., & Abrams, L. (2004). Instructional uses of computers for writing: The effect of state testing programs. *Teachers College Record, 106,* 1332–1357.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing* (CSE Technical Report 482). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Schwartz, D., & Arena, D. (2009). *Choice-based assessments for the digital age*. Stanford, CA: Stanford University, School of Education.

Sebrechts, M., Bennett, R., & Rock, D. (1991). Agreement between expert-system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76,* 856–862.

Shepard, L. (2000). The role of assessments in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shepard, L. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *72*, 232–238.

Sims-West, N. (1996). *An investigation of gender difference on the Scholastic Aptitude Test of Verbal Ability* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Thorndike, E. (1915). An improved scale for measuring ability in reading. *Teachers College Record, 16,* 31–53, 445–467.

Trabasso, T. (1981). On the making of inferences during reading and their assessment. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 56–75). Newark, DE: International Reading Association.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, H., & Hill, C. (2011). A paradigm shift for English language teaching in Asia: From imposition to accommodation. *The Journal of Asia, TEFL*, *8,* 231-258.

Winerip, M. (May 5, 2005). SAT Test Rewards Length and Ignores Errors. *The New York Times.*

Yuan, Y. (1997). *Reader response to the Taiwan Joint College Entrance Examination: English reading section* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Zhang, W. (2003). *Doing English digital* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

张薇 (2006). 英语数字素养的研究型评价模式,《外语教学与研究》, 第2期,115–121页 [Zhang, W. (2006). Assessing English digital literacy: The project approach. *Foreign Language Teaching and Research*, *2,* 115–121.]

Figure 11

Application Task

The poet writes about a contrast between his present life and that of his grandfather. What are some differences between the ways in which you and an older person you know look at things? Select one of these differences and give reasons to explain it. How does this difference affect your communication with this person?

I have an administrator who I have communicated with much during my time in high school named Bruce. Throughout high school, I have made it my goal to excel in certain areas of curriculum so as to stand out from the rest of my peers. Frustrated with my practice, Bruce, a much older man once in the armed forces, asked me, "Why are you trying to make things so much harder on yourself than they have to be?"

That struck me as odd. Throughout my entire juvenile life people have been telling me not to fall behind because no matter what I do there's always someone waiting to take my position and hold me back. So I replied, "There's so many people out there smarter and more qualified. I have to work hard now so I can make a better living later."

Bruce looked at me and sighed. Not a sigh of depression but one of sudden realization. "Look," he said, "you're going to find that you spend your entire life as a young man in search of money. You're going to work hard and maybe get it. But by the time you've earned enough to stop working so damned hard, you're going to be too old to enjoy it the way you wanted to."

I laughed at him when he told me this, years ago to date. After all, it made no sense to make money the primary incentive for staying and doing well in school and then turning around and deflating the importance of the correlation between money and happiness as Bruce had done.

During those years, I've continued to excel; hard work and constant struggle have defined my life long before and after my conversation with Bruce. Even now, I understand the message he was trying to convey, but I haven't really changed my course of action.

I don't and probably won't believe Bruce. That is, until and unless logic and experience prove him right. Although it seems feasible, it's not something I care to believe in. After all, I am a teenager, a young man in the prime of my life. I believe acting on Bruce's belief could ruin me; becoming so jaded, so weary of the struggles of life before I encounter even half of them seems foolish.

It is the word of an older generation against a younger generation, a person with the knowledge of experience against a person with the hope that naivety brings to the inexperienced. However, he was right about one thing, I will reach my goal. It's simply a question of, "When?"

Feedback

Content. We like the way in which your piece is structured as a real exchange between you and Bruce. This approach enables you to develop a lot of rich content. That said, we find your presentation of the contrast between “enjoying life while you can” and “working hard to ensure future success” sometimes hard to follow. For example, you suggest that Bruce contradicts himself: “It made no sense to make money the primary incentive for staying and doing well in school and then turning around and deflating the importance of the correlation between money and happiness.” Yet in the preceding text it is you, not Bruce, who suggests that future success is the main reason for working hard in school.

Your presentation of the contrast is either-or, which is easy for the writer, but less straightforward for readers, who may have thoughts that don’t fit neatly into the writer’s scheme. For example, you don’t allow for the possibility that hard work can be a deeply satisfying, even enjoyable, part of life. Also, we don’t see how giving yourself more downtime might make you “so jaded, so weary of the struggles of life.”

Clarity. You use paragraphing effectively—each of your seven paragraphs is a good length and is internally coherent. We especially liked the way in which you began a new paragraph by focusing on a reaction to what has just been said: “That struck me as odd,” “Bruce looked at me and sighed,” and “I laughed at him when he told me this.”

Of course, you could have used paragraphing to respond to the task as it was structured: first, a presentation of various differences, next a focus on a single difference and its source, and finally a characterization of the ways in which this difference affects communication. It might have been good if you had found some way to acknowledge the structure of the task while approaching the material in your own way.

Critical/Creative Thinking. Your personal voice is heightened by the way you use quotation. We like the fact that you quoted yourself as well as Bruce. We especially like your technique of commenting on the quoted speech: for example, you end the first paragraph with speech attributed to Bruce and then start a new paragraph in which you first describe your internal reaction and then what you said. This combination of internal and external reaction helps to convey the complexity of your response to what Bruce said.

Summary. You managed to write a great deal in just over 37 minutes (336 words on the interpretation task and 432 words on the application task). And the process log provides evidence that you did some revision. Still it might have been helpful if you had used at least some of the remaining time to read your work carefully and do even further revision. We look forward to reading what you will write on the next assessment—your voice is strong and engaging on the page.

Sample Process Log

We also used the computer's capacity to keep track of time to help us monitor how students went about responding to our tasks. Hence, along with students' written responses, we collected data on the order in which they moved through the various sections of an assessment activity and how long they spent on a given section. Table 1 shows the amount of time that the same student spent on the various tasks built around Leroy Quintana's poem.

TABLE 1

Section	Time (min:sec)	Cumulative (min:sec)
Resource	37:01	37:01
Orientation	0:04	0:04
Planning Task	0:07	0:11
Interpretation Task	2:18	2:29
Planning Task	0:17	2:46
Orientation	0:01	2:47
Planning Task	1:14	4:01
Interpretation Task	10:16	14:17
Application Task	0:04	14:21
Interpretation Task	2:30	16:51
Application Task	18:44	35:35
Return to Work/Final Exit	0:03	35:38
Application Task	1:17	36:55
Return to Work/Final Exit	0:06	37:01
Application Task	0:09	37:10
Return to Work/Final Exit	0:05	37:15
Sent file to server		37:15

The top row shows that this student kept the Resource window (the text of the poem) open during the entire session. The remaining rows show his sequence in moving from one task to the other and how much time he spent before switching to another window.

Table 2 shows our analysis of the raw data in Table 1, using familiar terms for crucial phases in the writing process—*previewing*, *drafting*, and *revising*.

TABLE 2

Section	Function	Time	Total	Words	Words/min
Resource		37:01	37:01		
Orientation		0:04 0:03	0:07		
Planning	previewing previewing revising	0:07 0:17 1:14	1:42		
Interpretation	previewing drafting revising	2:18 10:16 2:30	15:04	336	22.3
Application	previewing drafting revising revising	0:04 18:44 1:17 0:09	20:14	432	21.3

As indicated by the last column labeled **Words/min**, this student maintained a comparable rate of production while responding to two quite different tasks: the Interpretation task and the Application task. In the case of certain students, the information in this column revealed that even though they spent considerable time on a particular task, they were not able to produce much writing, which allowed us to provide more tailored feedback.

We also discovered the words/minute ratio to be useful when we were evaluating the viability of a particular task during field-testing. If this ratio was consistently low for students, it provided evidence that the task itself was blocking them from showing what they can do and thus needed revision.

Digital Project Model: Assessing Students in Chinese Universities

The digital project model was developed at Teachers College, Columbia University, by Zhang Wei and was subsequently implemented in a popular course known as *Doing English Digital* at Beijing University. The educational goals of this model were first articulated in the William P. Fenn Lectures, *English in China: Educating for a Global Future*, which I presented at Beijing University and twelve other universities throughout the People's Republic of China in 1998. These lectures proposed that the traditional College English course, which is required for all students not majoring in English, should become more digitally oriented by requiring students to use the Internet to conduct research in their major field of study, such as biochemistry, geology, economics, or history.

In carrying out such research, students develop an array of skills in using digital tools:

searching for relevant information (search tools such as Google)

evaluating the information (online lists of reliable resources in an academic discipline)

organizing the information (digital tools such as Inspiration)

making an oral presentation (digital tools such as PowerPoint)

making a written presentation (digital tools integrating text and graphics)[9]

This research experience can play an important role in preparing Chinese students to participate in a global future, since it helps them integrate their knowledge of English as the lingua franca of the Internet with digital skills that are crucial in transnational communication (Hill, 1998).[10]

Structure of *Doing English Digital*

Fundamental to the success of this digital project model is that it is curriculum-embedded and instructionally-oriented. At its core are eleven modules that provide various kinds of scaffolding to support students as they complete their research projects. Each module consists of five components: tasks, guidelines, examples, tools, and readings. These components have been designed to anticipate various problems that students encounter in conducting their research projects. The guidelines provide step-by-step help for completing the tasks in each module. The examples mainly come from previous student work or links to other websites. The tools offer different kinds of procedural and conceptual scaffolding, such as search tools (e.g., keyword searches, topical indexes, search engines), organizing tools (e.g., Inspiration), software for graphic organization, publishing tools (e.g., Adobe), software for Webpage development, and assessment tools such as checklists and evaluation rubrics.

[9] In communicating with mass audiences throughout China, I deliberately used the traditional distinction between oral and written communication. In digital communication, however, this distinction is often blurred, since the oral can be present in a written document (e.g., a video link that the reader can activate) and the written can be present in an oral presentation (e.g., PowerPoint slides that present written text the audience can read). Such hybrid forms of communication have become normative, especially for those who are growing up with hand-held devices that they use in their daily lives.

[10] It is commonly assumed that such communication takes place with a native speaker of English, but it often takes place with another individual who is not a native speaker, especially in Asian countries where native speakers do not abound and English is increasingly used as a lingua franca for transnational communication in the domains of politics, business, and education (Wang & Hill, 2011). It is important to bear in mind that not just in Asia, but throughout the world there are now considerably more non-native speakers of English than native speakers.

Doing English Digital has now been offered seven times at Beijing University and so Zhang Wei makes available on the course website previous research projects, which are categorized according to major topics (see Figure 12). These sample projects are useful in helping students to select a topic in their major field and then review the existing research.

Physics	Psychology	Economics
Education	Law	Literature
Philosophy	Politics	Advertising
Cultural Studies	Women's Studies	Literacy Studies

Figure 12

As students attempt to select a topic, they post potential topics online and provide feedback to each other. Once they have selected a topic and reviewed the relevant research online, they move on to the next stage, in which they post potential ways of organizing their research online and, once again, provide feedback to each other. As they begin to develop a written presentation, they submit an initial draft to Zhang Wei, who provides feedback, but also uses this draft to help authenticate that the final draft is fundamentally the work of the individual student.

During the final stage of the course, students make oral and written presentations that are evaluated according to the rubric that Zhang Wei has adapted from the digital testing model (see Figure 13):

Content	focused controlling ideas	sufficient evidence	credible materials
Clarity	coherent patterns of organization	consistent control of language	effective use of graphics
Critical/Creative Thinking	thoughtful treatment of the topic	fair-minded evaluation of varied perspectives	active interaction with sources

Figure 13

Zhang Wei has added a third subcriterion under each of the major criteria that focuses on digital aspects of the student project. Under *Content*, *credible materials* is used to evaluate whether students have used Internet resources that are well vetted and thus reliable (they are required to include live links in the digital version of their projects so that evaluators can easily go online and check up on the resources they have used). Under *Clarity*, *effective use of graphics* is used to evaluate whether tables and figures are well integrated with the text (students—and not just in China—often laboriously repeat the information in tables and figures in the text itself rather than moving on to provide strategic commentary on what has been graphically presented). Under *Critical/Creative Thinking*, *active interaction with sources* is used to evaluate whether students have reshaped the online material so that it is effectively integrated into their project.

In applying the rubric to student projects, evaluators use three levels of scoring: *excellent*, *good*, *passing* (along with +/– for each level). Two evaluators respond to each project and a third is used when the first two do not agree. In order to insure a stable use of the rubric, Zhang Wei followed a procedure that I developed when working as a consultant for the International Baccalaureate Program. In the first stage, sample student projects are scored holistically by evaluators who are experienced in using the rubric. In the second stage, the rubric is used to conduct detailed textual analyses of projects that exemplify different levels of scoring in order to create an exemplar packet (see Hill, 1998, for a more detailed description of this process).

Zhang Wei then uses the exemplar packet to train evaluators. In her dissertation research (Wei, 2003), the consistency in scoring was especially high for the written presentations ($r = .82$, $p < .01$). Not surprisingly, it was somewhat lower for the oral presentations ($r = .74$, $p < .01$), given the inherent complexity of an oral presentation in which speech must be integrated with visual information on PowerPoint slides.[11]

Zhang Wei also distributes the exemplar packet to students during the early stages of Doing English Digital to help them internalize the rubric. One of the major benefits of this digital project model is that it teaches students to internalize basic criteria that they can use to evaluate their own writing not only in this course but in other courses as well—and, indeed, in their later careers beyond the university.

[11] Zhang Wei and I are currently working on a more detailed rubric for evaluating PowerPoint presentations. This rubric includes criteria for evaluating (1) the construction of slides—for example, whether the use of language is appropriate (i.e., key words and phrases rather than entire propositions), and (2) the use of slides—for example, whether the presenter's oral communication, including gestures, is effectively integrated with the visual information available in the slides.

Sample Research Project

To show how Doing English Digital works, I will examine a research project conducted by Yue Yu, a student majoring in psychology at Beijing University. This project takes on an important topic in education—how best to develop moral thinking in children.[12]

His written presentation and the PowerPoint slides for his oral presentation can be found in the Appendix.[13] The written project is relatively short (2,031 words including references) and the number of slides limited—there are only 12, including an introductory title page and a final slide in which he thanks his fellow students for their attention. Zhang Wei deliberately restricts the length of both presentations, since the students she teaches have virtually no experience in extended speaking or writing in English.

Yue Yu, like most of his fellow students, has not spent time in a country where English is spoken and has had relatively limited exposure to the language in his formal education. In China, most students study English for about five hours per week in secondary school, where teachers have traditionally lectured about grammar as a means of controlling large classes. It is thus not surprising that various kinds of infelicities can be found in their initial efforts to use English in spoken and written communication.

Given this context, Yue Yu's use of English to express the sophisticated thinking evidenced in his project is quite remarkable. He received a score of *excellent* on both the oral and written presentations. Let us briefly consider how Zhang Wei (2010) applied the rubric in evaluating his presentations, bearing in mind that her purpose was to develop sample materials for training evaluators. She moves systematically from one criteria to the next, each time focusing on features of the student essay in relation to the subcriteria.[14]

[12] One of the attractive features of Doing English Digital is that students take on topics that are relevant to the larger society. When I last observed the course at Beijing University, a student was exploring how best to approach the topic of comparing Chinese and American university students in their use of social media.

[13] Zhang Wei requires students to include an evaluation of the reliability of their sources. I have included one example of such an evaluation in the Appendix, which can be found after the references.

[14] Her use of the rubric contrasts with the one reported in the digital testing model, where the focus was on providing feedback to the student.

Evaluating the Written Presentation

Content

Focused controlling ideas

The writer[15] begins by observing Chinese elementary education is built around reciting slogans. After reviewing various alternative proposals, he introduces the possibility of using moral dilemma stories as an alternative. He then introduces a theoretical framework as well as practical classroom methods for using these stories. He ends by pointing out various benefits as well as potential difficulties in using these stories.

Sufficient evidence

The writer provides relevant detail to support his ideas: for example, the first 12 lines provide rich documentation of the current methods of teaching moral education.

Credible materials

The writer has used professional websites that provide reliable information.

Clarity

Coherent patterns of organization

The writer has effectively used headings, paragraph markers, and connective phrasing to signal the structure of his essay. Consider, for example, lines 41–64 that discuss the potential benefits of using moral dilemma stories. This section begins with a heading that consists of a rhetorical question: *In what way would moral dilemma discussion benefit Chinese moral instruction*? (This is the first of three rhetorical questions used as parallel subheadings.) In this section, the writer discusses three benefits, each signaled by connective material initiating a new paragraph:

> Line 44: "One of the major benefits…"
> Line 53: "Apart from its benefit of autonomous learning…"
> Line 58: "Moreover, moral dilemma discussion is…"

Consistent control of language

The writer, despite the occasional awkward locution or grammatical infelicity, maintains a firm control of language at both the macro-level (as indicated by the above examples) and the micro-

[15] In discussing the application of the rubric, Zhang Wei uses the generic phrase "the writer" rather than a personal name, since anonymity is preserved throughout the evaluation process.

level (as indicated by a consistent use of vocabulary items such as “autonomous” and “moral reasoning” that maintain an appropriate register for the topic.

Effective use of graphics

The writer constructs a table that succinctly presents Kohlberg’s developmental stages of moral reasoning: the first two are presented in a column labeled ‘Preconventional,’ the second two in a column labeled ‘Conventional,’ and the final two in a column labeled ‘Postconventional.’ He also constructs a flow chart that illustrates how a moral dilemma story can be effectively presented in an elementary classroom. In the case of these graphics, the text that follows does not laboriously recycle the information presented, but rather moves on to provide strategic commentary.

Critical/Creative Thinking

Thoughtful treatment of the topic

The writer takes on an important topic—the moral education of children—and begins by pointing out that the approach in Chinese elementary schools has not been sufficiently thoughtful. He immediately engages his fellow students by providing vivid examples of the slogans that they were forced to repeat as children in primary school.

Fair-minded evaluation of varied perspectives

The writer reviews various approaches to moral education by both Chinese and Western scholars, but ends up recommending the one that he thinks would be most appropriate for children. After outlining benefits that this approach could bring, he is careful to address problems that are likely to arise if it is implemented in Chinese elementary schools. In addressing these problems, he points out the importance of adjusting the approach so that it is more congruent with Chinese cultural norms (i.e., children expect their teachers to provide authoritative opinions).

Active interaction with sources

The writer is willing to think critically about the Western model and adjust it in the light of what Chinese children expect from a teacher. This pragmatic spirit is present throughout the essay: the writer provides little direct quotation but rather rethinks the source material so that it fits the particular topic under consideration.

Evaluating the Oral Presentation

It is difficult to apply the rubric to an oral presentation based on PowerPoint slides. As we are all painfully aware, such slides are often deficient in both design and use. It is not uncommon that a presenter simply reads lengthy propositions crowded onto a slide while the audience squirms impatiently.

Yue Yu managed to avoid both kinds of problems. If you turn to the Appendix, you will see that his slides are parsimoniously constructed. Consider, for example, slide 2 on page 45. He uses the single word 'Outline' as a title and then provides short phrases—or merely a single word in the case of 'Introduction' and 'Conclusion'—to describe the four sections of his presentation. He uses short questions to further break down the third section, which is the heart of his presentation (and he avoids the further use of bullet points for these subsections).

This spirit of parsimony is also evidenced in the slides that Yue Yu constructed to guide his presentation of the third section. Consider, for example, slide 9 on page 46, which deals with the potential benefits of using moral dilemma stories. Under the bulleted heading 'In what ways would it help?' he lists three short phrases to guide his presentation:

autonomous learning
easy to accept
pertinent to real-life issues

When speaking, he used these phrases as mnemonic devices to cue both himself and his audience as he moved through the discussion of potential benefits (a strategic use of a pointer reinforced the power of these visual cues). Their mere presence on the slide was a signal to the audience that he had carefully planned his presentation and was prepared to speak extemporaneously. Given this greater freedom, he was able to maintain eye contact with his audience instead of looking down at a text.

Concluding Reflections

I have examined in some detail a digital testing model designed for American high school students and a digital project model designed for Chinese university students. I would now like to highlight certain features of these models that show particular promise for the future. I would then like to propose that these two kinds of models are best viewed as complementary and hence should be integrated into a more comprehensive model, which will be described as a *digital* *assessment model*, that can be used not only to support classroom teaching and learning but also certify high school students and select them for further educational opportunities.

Digital Testing Model

The digital testing model is not limited to traditional print literacy, but rather provides students the resources—print, sound, image, and animation—that they are accustomed to working with in a digital age. Students are provided a range of tools that allow them to work efficiently with these resources: for example, they can copy and paste material from film as well as text, or they can conduct a search for crucial material and rapidly assemble it in a strategic database. Hence this model reflects greater authenticity, since it allows students to engage in the kind of work that they ordinarily do when using a computer.

This model also has the virtue of presenting students with a set of integrated tasks. The planning task provides grounding in the resources that students draw on in responding to constructivist tasks: first an interpretation task in which they respond critically to the resources and then an application task in which they place the resources in a broader context. In responding to these two tasks, students work with digital tools as well: a notepad for planning what they will write, a live word counter for monitoring how much they are writing, and a spell checker for correcting typos and misspellings.

As students revise what they have written, they have access to familiar tools: for example, cutting and pasting allows for material to be reordered easily. As one student pointed out, when she comes up with a good idea, she simply writes it out and keeps it “at the front of what I am writing” so that she can draw on it when an appropriate context emerges. She also observed that if her fingers are not on a keyboard, she is not able “to do any real thinking and get any words flowing onto the page.” These words force us to consider whether assessment is fair when it requires students to respond in handwriting, thus depriving them of the tools they are accustomed to using. Of course the question of fairness is confounded by the fact that within our multicultural society students vary considerably in the degree to which they have access to computers. As we move more deeply into the digital age, this issue will become more prominent.

Finally, I would like to call attention to the process log, which allows us to analyze how students spend their time as they work with the resources and tasks. For example, we can determine whether they initially preview all three tasks, whether they use digital tools efficiently to assemble a database, and whether they spend sufficient time drafting and then revising their responses**.** Thus, the process log allows us to highlight time management along with *Content*, *Clarity*, and *Critical/Creative Thinking* in the feedback that we provide. The challenge we face in developing a digital testing model is to preserve broad values while providing students insights into how well they manage digital resources and tools.

Digital Project Model

The digital project model also leads to student work that is characterized by greater authenticity. In a digital age, using the Internet to find information about a particular topic is an essential activity. The course Doing English Digital is set up to teach students a comprehensive set of skills that they can use to find information online and then communicate it effectively.

This greater authenticity is reinforced by the social interaction that students engage in as they develop their research projects. During the early stages—identifying the topic, finding appropriate resources, developing a coherent plan for the presentation—they interact with each other and their teacher not only face-to-face but also through the course website. Once they begin to write what they plan to present, they continue to interact with the teacher, who provides feedback on early drafts.

Rubric Design

I would like to call attention to the rubric used in Doing English Digital to evaluate student work and provide feedback. It is adapted, as previously noted, from the one used in the digital testing model, which, in turn, was adapted from a rubric built for the International Baccalaureate. This rubric has the virtue of focusing on important values in writing while avoiding excessive detail. In using simple terms to identify three broad areas—*Content*, *Clarity*, *Critical/Creative Thinking*—it sets up a framework that is easy for evaluators to use and for students to internalize.

As the standards movement has developed in this country, rubrics have become increasingly complicated writing has come to play an important role in testing at both the state level (e.g., the New York State Tests used for certification) and the national level (e.g., the SAT used for selection). Unfortunately, this movement has spawned rubrics that reflect widely circulated standards that have value in the larger educational enterprise but are inappropriate for evaluating the kind of writing that can be done in a testing situation.

State Level. Inappropriately inflated rubrics are especially noticeable in tests designed for children at the state level. Figure 14 presents the rubric used to evaluate fourth graders' writing on the English Language Arts Test in New York State (2002). The first column lists general qualities, while the second provides descriptions of how these qualities are manifested in responses that receive the highest score (level 4).

Quality	Responses at Level 4
Meaning: The extent to which the response exhibits understanding and interpretation of the task and text(s)	*Taken as a whole:* • fulfill all or most requirements of the tasks • address the theme or key elements of the text • show an insightful interpretation of the text • make connections beyond the text
Development: The extent to which ideas are elaborated, using specific and relevant evidence from the text(s)	*Taken as a whole:* • develop ideas fully with thorough elaboration • make effective use of relevant and accurate examples from the text
Organization: The extent to which the response exhibits direction, shape, and coherence	*The extended response:* • establishes and maintains a clear focus • shows a logical sequence of ideas through the use of appropriate transitions or other devices
Language Use: The extent to which the response reveals an awareness of audience and purpose through effective use of words, sentence structure, and sentence variety	*The extended response:* • is fluent and easy to read, with vivid language and a sense of engagement and voice • is stylistically sophisticated, using varied sentence structure and challenging vocabulary

Figure 14

In a review of the New York State Test for fourth graders, I called attention to the mismatch between the criteria found in the rubric and the writing that children are able to do in the particular conditions that the test affords.

> Consider, for example, such descriptions of language use as "is fluent and easy to read, with vivid language and a sense of engagement or voice" and "is stylistically sophisticated, using varied sentence structure and challenging vocabulary." In state education departments throughout the country, phrases like these have been recycled in rubrics used to evaluate what children write on language arts tests. It is disconcerting that standards associated with the highly edited work of seasoned adult writers, working on familiar material over months or even years, is being applied to what children, working under the pressure of a high-stakes test, manage to get on the page when they have about 15 minutes to respond to three tasks about a story that they have just heard for the first time. (Hill, 2004, 1099–1101).[16]

[16] Since retirement, I have been mentoring a child in Harlem through a program for children who have a parent in prison. In helping him prepare for this test, I discovered that he ended up limiting his written responses because the large-size printing that he likes to use when writing for official purposes would extend well beyond the text

National Level. Since the SAT began to evaluate student writing in 2005, it has received a good deal of criticism for the approach it is using. The writing section includes not only a written essay but also multiple-choice tasks that require students to identify errors, complete sentences, and improve paragraphs. The total score is heavily weighted toward the multiple-choice component (75%).

The scoring of the essays is based on a rubric and carried out by two trained readers. If their scores differ by more than one point on a 6-point scale, a senior reader is called in to assign the final score. Since the essays are quite short (only 25 minutes is allowed to read the prompt and write a response), they can be rapidly scored (the average time used to score an essay is 3 minutes).

This brief writing and rapid scoring has led to fundamental questions about the value of including this kind of writing task on the SAT. Les Pearlman, who directs undergraduate writing at MIT, found that the length of the essay strongly correlates with the assigned score ($r < .9$): the shortest essays (about 100 words) tend to receive the lowest scores and the longest essays (about 400 words) the highest scores (Winerip, 2005).[17]

Pearlman also questions the official policy of ignoring factual errors when evaluating the essays. He argues that a key feature of undergraduate education at MIT is instilling in students a respect for the accurate use of factual information. From Pearlman's perspective, such respect is fundamental to the development of scientific thinking.[18]

boxes provided for answers. In explaining why his responses were so short, he said that he was afraid he would "lose points if his writing goes outside the box." When doing homework with him, I had discovered that his teacher subtracted points whenever his responses did not fit into the text boxes provided in his workbook.

I should further note that even if his printing had been small, the text boxes on the test were generally too small to accommodate the information called for by the tasks. If this child were to take the test on a computer—he is quite comfortable on a computer because of his love of videogames—the text box would automatically expand to accommodate whatever he writes.

[17] As far as I can ascertain, Pearlman did not carry out a multifactor analysis. I suspect that features such as effective sequencing of arguments would correlate positively with essay length. After all, one needs a certain amount of textual space in order to develop effective argumentation.

[18] Under Pearlman's guidance, an undergraduate at MIT took the SAT and wrote an essay about Franklin Delano Roosevelt and the Depression that was deliberately filled with factual errors. His essay was, however, the desired

Digital Assessment Model

Given the problems that attend evaluation of student writing in a testing situation, I recommend the development of a comprehensive model that includes a project component as well as a testing component. The term *digital assessment model* can be used to refer to this more balanced approach, which maintains the positive benefits associated with traditional testing while allowing for a more responsible appraisal of student writing. As illustrated by Doing English Digital, when students are allowed a more extended time frame and provided scaffolding that supports the writing process, they are able to produce writing that can be evaluated fairly with rigorous standards.

An extended time frame does allow for the possibility of a student receiving help from others, which the traditional approach to testing is designed to prohibit. Doing English Digital is designed so that students receive responses to their writing not only from the instructor but also from other students in the course. From the vantage point of Zhang Wei, these responses are fundamental to what goes on in writing projects in the real world, and hence assessment should take account of the degree to which an individual student can make effective use of feedback. At the same time, Zhang Wei relies on the various drafts that individual students produce as a means of verifying that the final draft is essentially their own work. Given that these drafts are digitally stored, she is able to examine them closely to detect both the changes that signal an effective response to feedback and the continuities that signal the authentic voice of an individual writer.[19]

Once an extended time frame is introduced, the cost of evaluating student writing increases dramatically. As already observed, the average time for a reader to evaluate the relatively short

length and contained vocabulary items such as “plethora” that are used in essays that receive a high score. His essay received a score of 10: the maximum number of points is 12, since each rater can assign up to 6 points (Jaschik, 2007). I should note that in the rubric developed for the International Baccalaureate and adapted for both the digital testing model and the digital project model, *accuracy* is included as a subcriterion under *Content*.

[19] Digital technologies will increasingly be able to authenticate the work of individual students by analyzing samples of their writing for stylistic features. But even if such technologies are perfected, collecting valid samples for individual students might turn out to be too difficult and hence the judgment of thoughtful readers will, no doubt, still be needed.

written essay on the SAT is three minutes. The sample essay by Yue Yu is 2,031 words, which is about five times longer than the lengthier essays written for the SAT.

In the future, automated scoring will play an increasingly prominent role in holding down cost. As Bennett (2011) has observed, automated scoring is built around surface features, ranging from spelling and punctuation to discourse markers of cohesion, which are markedly different from the features represented in the rubrics that we have been discussing. As the field of artificial intelligence continues to develop, we can anticipate increasingly reliable ways of using surface features as indices to deeper levels of structure. It is important to bear in mind that in our ordinary acts of reading we have access to deeper levels of structure through a judicious sampling of surface features.[20]

As Bennett further points out, it is difficult to know just how trained evaluators make use of a rubric when evaluating student writing. Anecdotal evidence suggests that they are able to use a complex array of sampling techniques to arrive at holistic judgments that are not the result of a mechanical application of the rubric (although these judgments can still be reasonably consistent with its values). Bennett makes a tantalizing suggestion that I plan to test out in a forthcoming research project: use two independent systems of automated scoring and bring in a human rater only when they disagree.[21]

[20] Automated scoring of student writing is called for in the consortium *Partnership for Readiness in College and Career*s (PARCC), one of the two major consortia that are being funded by the U.S. Department of Education to develop assessment systems for the Common Core Standards. According to knowledgeable sources, the complexity of developing reliable automated scoring systems has been considerably underestimated, and it will be difficult to meet the deadline set for the academic year 2014-15.

[21] I will collaborate with Wang Haixiao, who chairs the Department of Applied Foreign Languages at Nanjing University, on a research project in which we will use both China-based automated scoring and Western-based automated scoring to evaluate research projects that were scored by the rubric in the Doing English Digital course at Beijing University. We hypothesize that the two methods of automated scoring will produce scores more similar than those based on the rubric, given that they both are oriented toward a presumably comparable set of surface features. The leading vendor of automated scoring in China, like many vendors in this country, has not made public the system it uses. As Bennett points out, the development of automated scoring is handicapped by the widespread lack of transparency.

Digital Archives

A comprehensive digital assessment model could be built around an archival system used to carry out the basic functions of certification and selection. Given the capacity of digital technologies to efficiently store and retrieve information, constructing a unified system at the national level is technically feasible at the present time. Before one can be put in place, however, formidable political obstacles, most notably those having to do with the strong tradition of states' rights, will have to be overcome. The movement to adopt Common Core Standards is clearly a step toward developing a more unified system, and the fact that 43 states have already approved these standards provides grounds for cautious optimism.

How might digital archives for individual students be constructed so as to strengthen the relations between assessment and classroom teaching and learning? Let us consider, in turn, the testing component and the project component.

Testing Component. This component would consist of carefully designed classroom activities carried out under the strict conditions associated with testing. In order to insure comparability, these activities could be based on certain strands in a common curriculum, perhaps those in American history that have to do with developing civic responsibilities (see the Common Core Standards for material that would be widely accepted across the political spectrum and thus not be opposed on partisan grounds).

These classroom activities would take place on a regular basis throughout the academic year and all the student work would be digitally archived. For the purposes of accountability, there would be externally appointed evaluators, working with classroom teachers, who would use methods of random sampling to evaluate student responses to a limited number of classroom activities in selected areas of the curriculum.

Since these classroom activities would take place on a regular basis, students would be less likely to think of them as tests, especially since teachers would use process logs to provide helpful feedback on matters such as time management and the writing process. In effect, everyday activities of the classroom would come to function as tests only as they are selected by a process of random sampling.

Given the demands of accountability that accompany any assessment system, this testing component would necessarily carry substantial weight. The fact that this component would be digitally administered could play an important role in accountability: for example, real-time records would be available in the digital archives and could be used to document that the strict time limits associated with testing are properly observed.

Project Component. Here, too, random sampling methods would be used to select samples of student work from the digital archives that would be evaluated by the team of external evaluators and classroom teachers. Since the number of projects that an individual student can carry out is limited, the team would evaluate only one or two projects for each student in a given subject matter (e.g., English Language Arts). The system might be designed to allow students to select one project that they would submit to the evaluation team. They would submit not only the project, but also a statement that explains why they have selected it as representing their best work.

In any evaluation of student projects, it is imperative that teachers be included so they can deepen their experience in using the rubric and reinforce its standards in their daily interactions with students. The ultimate goal of any assessment system should be to insure that teachers continuously circulate high standards in the classroom so that students bring them to bear on all the work that they do, not only in school but also in the larger society.[22]

Certification. There are different ways in which a state education agency could use information based on digital archives in granting students a high school diploma. For purposes of efficiency, it could use only a numerical score based on the testing component if it is sufficiently high. Hence, information from the project component would be introduced only when the score from the testing component is marginal and needs to be supplemented.

Another approach would be to use information from both components simultaneously, with the possibility that the state agency might provide greater weight to one of the components (presumably the testing component). Ultimately, decisions about the use of archival data would depend upon policy decisions at the state level. Given the resistance to centralized authority in this country, it is important that as much autonomy as possible be maintained at the state level.

[22] As Linda Darling-Hammond (2004, 2010) has observed, as assessment activities become integrated into the everyday classroom, they come to play an important role in the professional development of teachers. See Bennett (2010) for discussion of the ways in which the CBAL project that is underway at ETS contributes to teacher development.

Selection. Admissions offices in American colleges and universities would develop individual policies about how to use information from digital archives. Since these offices are already committed to using samples of student writing, they would welcome the opportunity to receive randomly selected student writing evaluated within the project component. Such writing could be supplemented by a sample that individual students select from digital archives and submit along with a statement of why they value this particular writing. As for the quantitative component of an admissions dossier, the score generated by the testing component of the digital assessment model could be used in place of an SAT or ACT score.

A final thought—a comprehensive digital assessment model, if properly designed and administered, could lead to a diminished use—or even a gradual withering away—of externally mandated tests based on the multiple-choice format. Such a change could lead to greater integrity in the American classroom: It would free students from the debilitating anxieties they often experience in preparing for these tests and teachers from the burden of devoting an inordinate amount of class time to test-prep activities.

As the larger world explores the ways digital technologies can transform educational assessment, countries such as France and China with a strong central government are in a position to act boldly and use digital technologies to create systems that more effectively integrate assessment with teaching and learning. Despite all the political obstacles to developing an integrated assessment system in this country, the Gordon Commission could fulfill its mandate on the future of assessment by articulating a vision in which all the information to be used in evaluating students is generated in carefully designed classroom activities that are stored in digital archives. This country is now entertaining proposals to build digital archives for personal health information, so why not formulate a proposal for a comprehensive assessment model that would use digital archives in certifying high school students and selecting them for further educational opportunities?

Appendix

Applying Moral Dilemma Discussion in Chinese Elementary Classroom

Yue Yu
Yuanpei Pilot Program and Department of Psychology
Peking University

"Love the motherland, the people, labor, and science and take good care of public property," this is the famous "Five Love" slogan used in Chinese elementary school. In a typical moral education class, pupils would echo slogans like this again and again, sometimes with the words go in from one ear and out of another. It is natural to ask the question when viewing such scenarios: is this kind of instruction effective?

In fact, researchers have raised the same question. Current Chinese moral education in elementary school mainly teaches social moral concepts as the "Five Love", and behavior norm such as "respect teachers" and "follow the disciplines" (Xia, et al., 2005). Patriotism and collectivism are highly emphasized while little concern has been put on pupil's character and moral reasoning. As for teaching method, the dominant classroom instruction is exhortation, which is sometimes referred as "bag of virtues" or "values clarification" (Zhang, 2002; Xia, et al., 2005). This kind of instruction has shaped pupils who can only recite slogans without knowing how to apply them, and school is disconnected with family and community education (Yang, 1997). As Zhang (2002) pointed out, Chinese moral education is facing quandaries.

Former researchers have given various suggestions to improve Chinese moral education. These include changing virtual-situation education into real-situation education (Fu, 2005), paying more attention to inherent moral culture (Liao, 2000), corresponding the instruction with pupils' moral reality (Zhang, 2002), and combining school education with family and community education (Yang, 1997). However, these concepts seem too theoretical to be taken into classroom practice. In this paper, I would introduce a distinctive teaching approach used in western countries—moral dilemma discussion, and discussed its possible appliance in Chinese elementary classroom.

Moral Dilemma Discussion

Moral dilemma discussion approach, or New-Socratic approach, is a teaching technique derived from Kohlberg's theory of moral development (Kohlberg, 1981). According to his framework, the life-long moral development can be divided into six stages (see table 1), each representing a unique type of thinking defined by how we process moral-ethical and value questions. Higher stages of moral development take account of broader perspective, represent more complex and abstract thought, contain more personal empathy, and provide more principle-based solutions to social problems.

Table 1: *Kohlberg's Stages of Moral Reasoning*

I. Preconventional Level	II. Conventional Level	III. Postconventional Level
Stage 1: Punishment and Obedience Orientation. Physical consequences of action determine its goodness or badness.	**Stage 3: "Good Boy-Good Girl" Orientation.** Good behavior is whatever pleases or helps others and is approved of by them. One earns approval by being "nice."	**Stage 5: Social Contract Orientation.** What is right is defined in terms of general individual rights and in terms of standards that have been agreed on by the whole society.
Stage 2: Instrumental Relativist Orientation. What is right is whatever satisfies one's own needs and occasionally the needs of others. Elements of fairness and reciprocity are present, but they are mostly interpreted in a "you scratch my back, I'll scratch yours" fashion. Individual adopts rules and will sometimes subordinate own needs to those of the group.	**Stage 4: "Law and Order" Orientation.** Right is doing one's duty, showing respect for authority and maintaining the given social order for its own sake. People define own value in terms of ethical principles they have chosen to follow.	**Stage 6: Universal Ethical Principle Orientation.** What is right is defined by decision of conscience according to self-chosen ethical principles. These principles are abstract and ethical, not specific moral prescriptions.

Note. Adapted from "*Educational Psychology: Theory and Practice (7th ed.)*" by R. E. Slavin, 2003, Boston: Pearson Education. p. 55.

His framework emerged from interviews with his research subjects using moral dilemma stories, which describe troublesome moral situations.

Kohlberg's story-telling approach has been used in the moral classroom to increase the level of the pupil's moral maturity (Kohlberg, 1981; Berkowitz, 1984). Refined by Berkowitz (1984), such discussion typically contains three successive phases: starting, continuing, and reaching closure, as illustrated in Figure 1. During discussion teacher should figure out the current moral status of individual pupil and treat them respectively, and they should also hold supportive view to pupils' discussion (Berkowitz, 1984). The application of dilemma-discussion approach in the instructional practices of moral education proves to be consistently effective for children's moral development (Thoma, 1984). For example, in an experiment servicing institutionalized delinquent and predelinquent students, moral discussion group showed significantly higher impact on moral reasoning compared to values clarification group and control group (Niles, 1985). In fact, there appeared many successful projects of dilemma-discussion approach all over the U.S. since 1980 (Slavin, 2003).

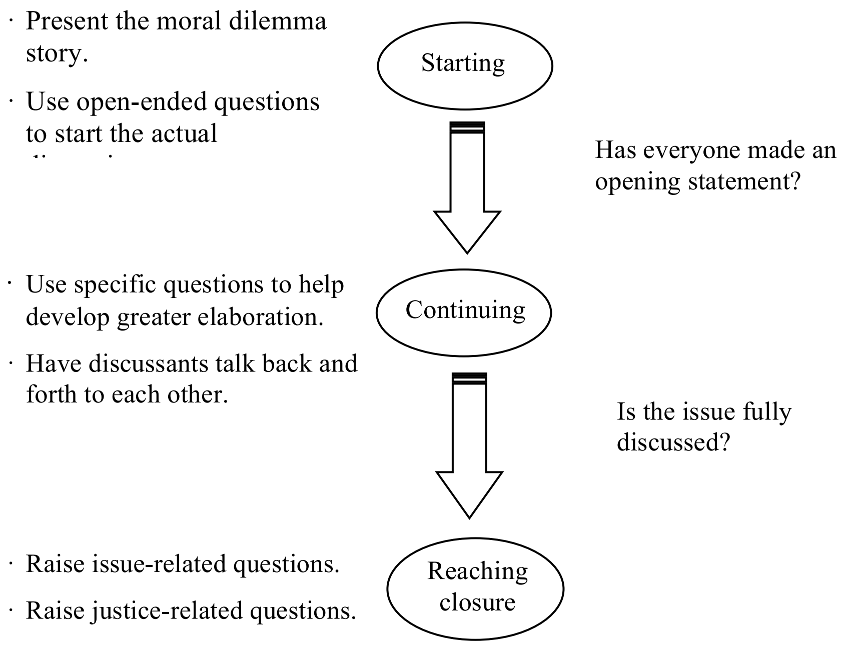

· Present the moral dilemma story.
· Use open-ended questions to start the actual
Starting
Has everyone made an opening statement?
· Use specific questions to help develop greater elaboration.
· Have discussants talk back and forth to each other.
Continuing
Is the issue fully discussed?
· Raise issue-related questions.
· Raise justice-related questions.
Reaching closure

Figure 1. Leading moral dilemma discussions

Applying Moral Dilemma Discussion in Chinese Elementary Classroom

In what way would moral dilemma discussion benefit Chinese moral instruction?

Compared with Chinese traditional moral instructions, dilemma-discussion approach has remarkable benefits in its autonomous learning, easy acceptance and pertinence to real-life issues.

One of the major benefits of moral dilemma discussion is that it helps foster moral reasoning and promote autonomous learning of pupils. Kohlberg (1981) has claimed that virtue judgment is just putting a label on certain people or behavior, while moral decision always relies on other thinking process. In fact, the base of real-life moral actions is not right-wrong judgments, but the way to analyze and reason these situations. Experiments showed that students in lower moral reasoning stages cheat in tests four times as much as their classmates in higher stages, while their judgment on propriety of cheating remains the same (Kohlberg, 1981). Moral dilemma discussion focuses on improving children's reasoning skills on moral issues, rather than simply identifying something as right or wrong. Since moral reasoning in Chinese students is less critical and logical (Xie, 2001), this kind of training is especially valuable.

Apart from its benefit of autonomous learning, moral dilemma discussion approach is also easier for pupils to accept. Dilemma-discussion approach is based on the identification of children's current moral developmental stage, and aimed at pushing them forward within their capacity. In contrast, traditional moral education in China often talks about abstract virtue criteria, which are sometimes far beyond the pupil's level of comprehension. And only through understanding can moral judgments be digested and applied to action.

Moreover, moral dilemma discussion is highly pertinent to issues in real life. Materials of moral dilemma can be selected from headline stories from newspapers, everyday incidents, popular moral issues, and incidents from movies or readings. The advantage is two-fold: on the one hand, real issues can arouse much interest, thus increasing the motivation and involvement of pupils; on the other hand, the judgments of real moral situations can be applied to real life behavior more easily. Since students taught by traditional Chinese moral instructions often show dissociation between their moral judgments and moral action (Lu, 2002; Fu, 2005), real issues can bridge principles and reality.

What are probable challenges in the application?

When borrowing a teaching approach from one country to another, cultural differences are of primary concern for its feasibility. As for moral development, literature has revealed a cross-culture character of Kohlberg's stages, which lends support to the usage of moral dilemma in different countries. Snarey (1985) has reported two studies, conducted in Kenyan and Turkish respectively, where moral dilemma material was used with properly transposed contents. The result showed a similar sequence of moral development from the less complex to the more complex levels, and the main cross-cultural difference showed up was between urban and tribal societies rather than western versus nonwestern societies. A later survey in both Taiwan and U.S. also found the important dimensions identical in the response to the dilemmas, regardless of the subjects' language and religion differences (O'Neill, 1998). Hence it is fairly reasonable to assume that Chinese children may go through a similar development in moral maturity.

Although the universality of moral development is guaranteed, other problems might occur in teaching practice due to cultural difference. The role of teacher, for example, is quite different between China and western countries. According to dilemma-discussion approach, teachers should withhold their own views, and let students reason by themselves (Berkowitz, 1984). Whether this point is suitable for Chinese moral education is open for discussions. Since moral education in China is often combined with ideological and legal education (Zhang, 2002), it bears responsibility of setting up social convention and political standpoint for pupils. Moral dilemma discussion itself cannot achieve all of the aims, exhortation and illustration of objective moral standard is still needed for children's civic development.

What are practical suggestions in applying moral dilemma discussion?

Classroom discussion is the core of dilemma-discussion approach. With concern to the reality of China, I suggest that moral dilemma discussions replace some (not all) of the ideological and moral classes in higher grades of elementary school. This grade period is chosen because children move from preconventional morality stage into conventional stage between age 9 and age 12 (Kohlberg, 1981; refer to table 1), so it is critical time for their moral development. During a discussion class, the material of moral dilemma can be searched from many ways as mentioned, and an atmosphere should be established that every pupil can raise his or her opinion no matter it is right or wrong. In the closure part, slightly different from Kohlberg's view, I think the teacher should give an instructive answer to the dilemma, which represents the stage of moral growth that is a little higher than the pupils'. This kind of combination between dilemma discussion and values clarification can better fit the multi-aimed moral class in China.

Besides classroom instruction, another scenario for moral dilemma discussion can be peer interaction. According to William Damon (1988), children at elementary-school age would discuss concepts of fairness and justice when confronted with difficult group decisions, which would enable them “to become a full citizen in society, with all of a citizen’s leadership prerogatives and fellowship responsibilities” (p. 86). Hence, moral dilemma discussion in peer group should be encouraged though establishing a supportive and democratic campus atmosphere.

Conclusion

What is the most effective way to cultivate virtuous citizens? This question is widely concerned and long debated in China. In this paper I suggest applying a western education approach—moral dilemma discussion—to moral education system in China after reasonable modifications. The dilemma-discussion approach emphasizes autonomous learning compared to traditional approach, and is easier to accept and more pertinent to real-life issues. Moral dilemma discussion has been proved to be effective in fostering children’s moral reasoning, and is widely used in western countries. With evidence of cross-cultural character of moral development, it’s reasonable to anticipate that dilemma-discussion approach would improve moral education in Chinese elementary school. Together with traditional value clarification, this new approach would broaden and vivify moral instruction in elementary classroom, thus fostering pupils’ internal moral development in a more effective way. A further step can be combining moral dilemma discussion in school with pupils’ behavior in family and community, so as to establish an all-around moral educational system to produce righteous citizens.

References

Berkowitz, M. (1984). Process analysis and the future of moral education. Paper presented at the annual meeting of American Educational Research Association, New Orleans, April 23.

Damon, W. (1988). *The moral child.* New York: Free Press.

Fu, W. (2005). Moral conflict of authenticity and students' moral development. *Educational Research, 12(3),* 13-16. Retrieved March 2007, from *CNKI* database.

Kohlberg, L. (1981). *Essays on moral development, Volume I & II.* New York: Harper and Row.

Niles, W. J. (1985). Effects of a moral development discussion droup on delinquent and predelinquent boys. *Journal of Counseling Psychology, 33(1),* 45-51. Retrieved May 2007, from *Sciencedirect* database.

O'Neill, P. & Petrinovich, L. (1998). A preliminary cross-cultural study of moral intuitions. *Evolution and Human Behavior, 19,* 349-367. Retrieved March 2007, from *Sciencedirect* database.

Slavin, R. E. (2003). *Educational Psychology: Theory and Practice (7th ed.).* Boston: Pearson Education.

Snarey, J. (1985). Cross-cultural universality of social-moral development. *Psychological Bulletin, 97(2),* 202-232.

Thoma, S. (1984). Do moral education programs facilitate moral judgment? A meta-analysis of studies using the Defining Issues Test. *Moral Education Forum, 9(4),* 20-25.

Bibliographical information & Source evaluation

Article 1

Search strategy: searched on "ScienceDirect" using "Moral dilemma" & "education" as key words.

Bibliographical information:

Author name: O, Neill Patricia; Petrinovich, Lewis

Date of publication: 1998

Title of the article: A Preliminary Cross-Cultural Study of Moral Intuitions

Name of the journal: *Evolution and Human Behavior*

Name of the publisher/database: Elsevier

Date of retrieval: 07.3.29

URL: http://www.sciencedirect.com/science/article/B6T6H-3VB35PS-1/2/807ae734f67d388cf2bb4c334db0116e

Source evaluation:

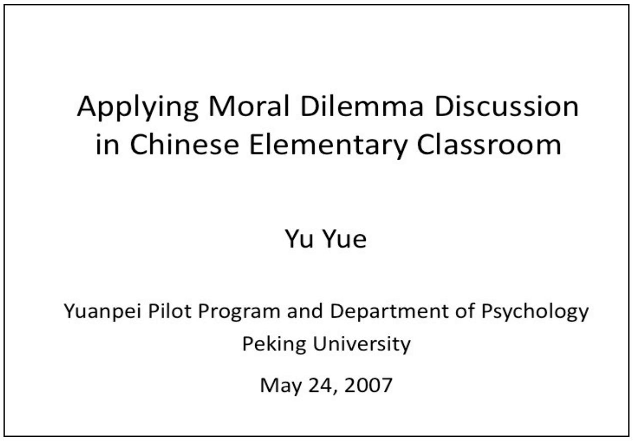
Relevance: The article discusses the universality of moral beliefs and is thus relevant to my topic.

Authority: As I've never heard of the journal *Evolution and Human Behavior* nor the authors, it is not easy to justify the authority of this article.

Accuracy: The article is supported by abundant theoretical and empirical evidence and is thus reliable.

Currency: The article is a little behind the times since it's published in 1998.

PowerPoint Slides Used in Yue Yu's Oral Presentation

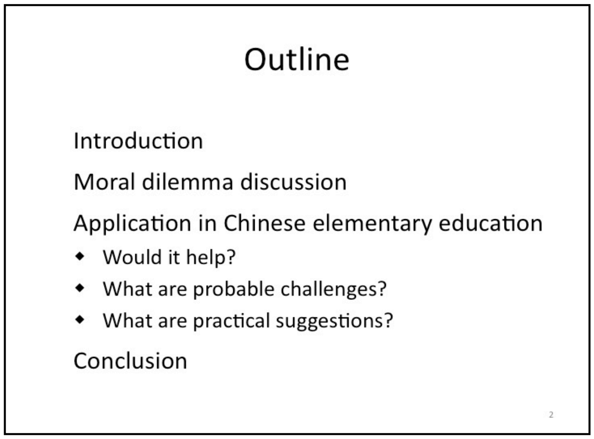


Applying Moral Dilemma Discussion in Chinese Elementary Classroom

Yu Yue

Yuanpei Pilot Program and Department of Psychology
Peking University

May 24, 2007



Outline

Introduction

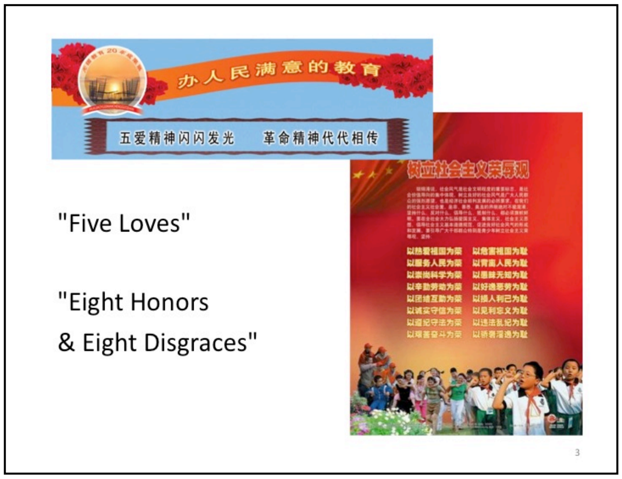
Moral dilemma discussion

Application in Chinese elementary education

- Would it help?
- What are probable challenges?
- What are practical suggestions?

Conclusion

2

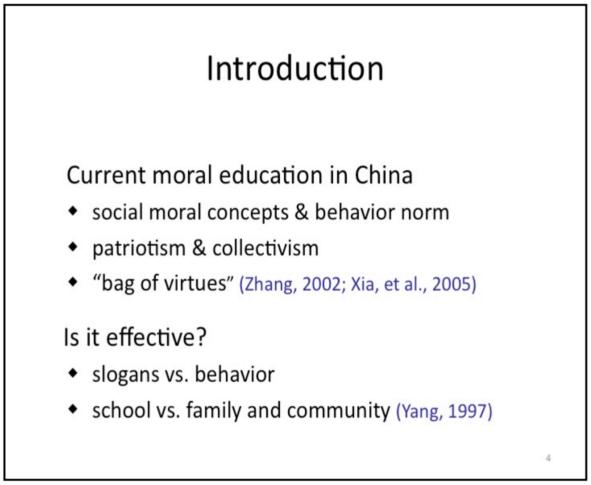


办人民满意的教育
五爱精神闪闪发光 革命精神代代相传
树立社会主义荣辱观
以热爱祖国为荣 以危害祖国为耻
以服务人民为荣 以背离人民为耻
以崇尚科学为荣 以愚昧无知为耻
以辛勤劳动为荣 以好逸恶劳为耻
以团结互助为荣 以损人利己为耻
以诚实守信为荣 以见利忘义为耻
以遵纪守法为荣 以违法乱纪为耻
以艰苦奋斗为荣 以骄奢淫逸为耻

"Five Loves"

"Eight Honors
& Eight Disgraces"

3



Introduction

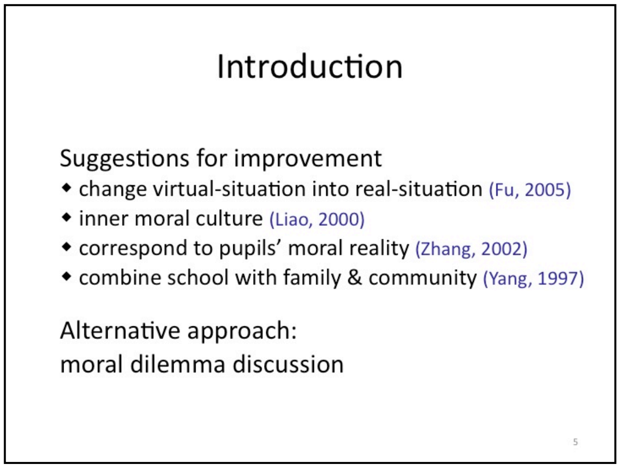
Current moral education in China

- social moral concepts & behavior norm
- patriotism & collectivism
- "bag of virtues" (Zhang, 2002; Xia, et al., 2005)

Is it effective?

- slogans vs. behavior
- school vs. family and community (Yang, 1997)

4



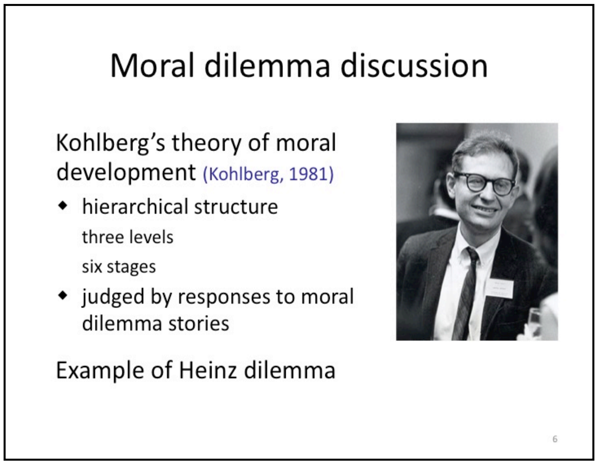
Introduction

Suggestions for improvement

- change virtual-situation into real-situation (Fu, 2005)
- inner moral culture (Liao, 2000)
- correspond to pupils' moral reality (Zhang, 2002)
- combine school with family & community (Yang, 1997)

Alternative approach:
moral dilemma discussion

5



Moral dilemma discussion

Kohlberg's theory of moral development (Kohlberg, 1981)

- hierarchical structure
  - three levels
  - six stages
- judged by responses to moral dilemma stories

Example of Heinz dilemma

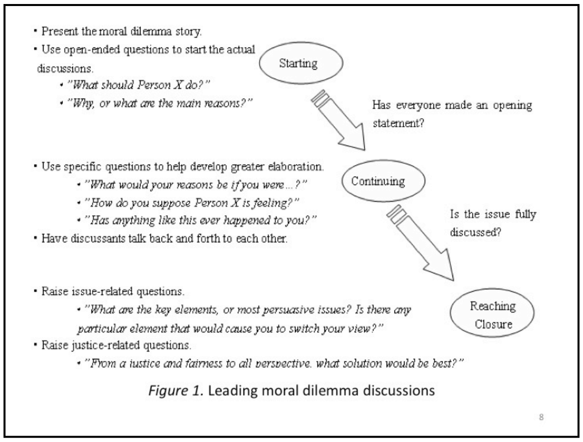
6

Moral dilemma discussion

Dilemma discussion in classroom (Berkowitz, 1984)

- story-telling + questions + discussion
- treating pupils respectively & supportive atmosphere
- more effective than values clarification (Niles, 1985)
- successful projects in U.S. & Israel (Slavin, 2003)

7


Starting
Has everyone made an opening statement?
Continuing
Is the issue fully discussed?
Reaching Closure

- Present the moral dilemma story.
- Use open-ended questions to start the actual discussions.
  - *"What should Person X do?"*
  - *"Why, or what are the main reasons?"*
- Use specific questions to help develop greater elaboration.
  - *"What would your reasons be if you were…?"*
  - *"How do you suppose Person X is feeling?"*
  - *"Has anything like this ever happened to you?"*
- Have discussants talk back and forth to each other.
- Raise issue-related questions.
  - *"What are the key elements, or most persuasive issues? Is there any particular element that would cause you to switch your view?"*
- Raise justice-related questions.
  - *"From a justice and fairness to all perspective, what solution would be best?"*

Figure 1. Leading moral dilemma discussions

8

Applying in China

In what ways would it help?

- autonomous learning
  right/wrong judgement => moral reasoning
- easy to accept
- pertinent to real-life issues
  bridge principles and reality

9

Applying in China

What are probable challenges?

- universality in moral development
- cultural difference
- role of teacher

What are practical suggestions?

- classroom instruction
- peer interaction

10

Conclusion

Moral dilemma discussion can be applied to improve moral education in Chinese elementary education.

Further implication:
an all-round moral educational system

11

Thanks for your attention!

12

Acknowledgments

I am grateful to Ed Gordon for the invitation to reflect upon how educational assessment can best serve teaching and learning. From the outset of my career at Teachers College, Columbia University, Ed encouraged me to take on formidable challenges and was generous in providing financial and, more importantly, intellectual support. I am fortunate to have had such a wise mentor throughout the years.

I am also grateful to Eric Larsen, a colleague at Teachers College, Columbia University, for helping me put together this document. He, like Ed, has been a steady influence as I have attempted to make sense of various approaches to assessing students in this country as well as abroad. He has been especially helpful in encouraging me to draw on linguistics and its allied discipline discourse analysis in deconstructing what goes on in the peculiar genre of multiple-choice testing that has such power over our lives.

I would like to thank E. Wyatt Gordon and Maralin Roffino for their help in shaping this document. They have shown patience in communicating with me during the final stages of its preparation while I was in Japan.

References

Adames, J. (1987). *A study of the pre-reading process of selected English as a second language college students* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Aronowitz, R. (1984). Reading tests as texts. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 43–62). Norwood, NJ: Ablex.

Baker, E. (2008). Learning and assessment in an accountability context. In K. E. Ryan & L. A. Shepard (Eds.), *The future of educational accountability* (pp. 277–291). New York: Routledge.

Baker, E. (2009). The influence of learning research on the design and use of assessment. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert design of optimal learning environments* (pp. 333–355). New York: Cambridge University Press.

Baldwin, J. (1998). Stranger in a village. *Collected essays* (T. Morrison, Ed.). New York: Library of America.

Bennett, R. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5–12.

Bennett, R. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, 9(5), 1–23.

Bennett, R. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2), 70–91.

Bennett, R. (*2011*). Automated scoring of constructed-response literacy and mathematics items. Available at http://www.acarseries.org/papers

Bennett, R., & Bejar, I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Bhasin, J. (1990). *Main-idea tasks in reading comprehension tests and the responses of bilingual poor comprehenders* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: School of Education, King's College.

Bloom, B. S. (1984). *Taxonomy of educational objectives*. New York: Longman.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, *12*(1). Available at http://epaa.asu.edu/epaa/v12n1

Chu, H. (1993). *Assessing Chinese kindergarten children in New York City* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Clarke-Medura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research in Technology and Education, 42*(3), 309–328.

Coyle, M. (1992). *The New Jersey high school proficiency test in writing: A pragmatic face on an autonomous model* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106,* 1047–1085.

Darling-Hammond, L. (2010). New-generation assessment of common core standards: Moving toward implementation. Available at http://www.acarseries.org/papers

Dewey, J. (1998). *The essential Dewey* (Vols. 1–2, L. A. Hickman & T. Alexander, Eds.). Bloomington: Indiana University Press.

Ehren, M. C. M., & Hatch, T. (forthcoming). Responses of schools to accountability systems using multiple measures: The case of New York City elementary schools.

Engel, S. (2011). Measuring our success. *Teachers College Record* (ID Number 16318). Available at http://www.tcrecord.org

Gee, J. (1992). What is reading? Literacies, discourses, and domination. *Journal of Urban and Cultural Studies*, *2,* 65–77.

Gordon, E. W. (1970). Toward a qualitative approach to assessment. *The Report of the commission on tests*. New York: College Entrance Examination Board.

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 11–32.

Gordon, E. W. (1999*). Education and social justice: A view from the back of the bus.* New York: Teachers College Press.

Hakuta, K. Improving education for all children: Meeting the needs of language minority children. In D. Clark (Ed.), *Education and American youth*. Washington, DC: Aspen Institute.

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hill, C. (1992). *Testing and assessment: An ecological approach* [Inaugural lecture for the Arthur I. Gates Chair in Language and Education]. New York: Teachers College, Columbia University.

Hill, C. (1995). Testing and assessment: An applied linguistics perspective. *Educational Assessment, 2,* 179–212.

Hill, C. (1996). *Exemplar essay project: Theory of knowledge*. International Baccalaureate.

Hill, C. (1998). *English in China: Educating for a global future*. Excerpt available at http://162.105.138.207/webcourse/advanced_english/English%20in%20China.htm

Hill, C. (2000, March 18). Practicing without learning. *The New York Times,* p. A15.

Hill, C. (2001a, December 27). Pitfalls of annual testing. *The Christian Science Monitor,* p. 9.

Hill, C. (2001b). Short-answer questions in College English testing. In *Research on Teaching College English in China* (pp. 172–184). Beijing, China: Beijing University Press.

Hill, C. (2001c). *Linguistic and cultural diversity: A growing challenge to American education.* New York, NY: The College Board.

Hill, C. (2003). Integrating digital tools into a culturally diverse curriculum: An assessment model for the Pacesetter Program. *Teachers College Record, 105,* 278–296.

Hill, C. (2004). Failing to meet the standards: English Language Arts Test for New York State. *Teachers College Record, 106,* 1086–1123.

Hill, C. (2012). Educational research: The challenge of using an academic discipline (reprint of the Lawrence A. Cremin Lecture delivered in 2010). *Teachers College Record, 114*, 1–42.

Hill, C., Black, J., & McClintock, R. (1994) Assessing student understanding and learning in constructivist study environments. In M. R. Simonson, N. Maushak, & K. Abu-Omar (Eds.), *16th annual proceedings of selected presentations at the 1994 national convention of the Association for Educational Communications and Technology.* Washington, DC: AECT.

Hill, C., & Larsen, E. (1992). *Assessment in secondary education: A review of emerging practices*. Berkeley: University of California, National Center for Research in Vocational Education.

Hill, C., & Larsen, E. (2000). *Children and reading tests.* Stamford, CT: Ablex.

Hill, C., & Parry, K. (1988). *Reading assessment: Autonomous and pragmatic models of literacy* (LC Report 88-2). New York: Teachers College, Columbia University, Literacy Center.

Hill, C., & Parry, K. (1989). Autonomous and pragmatic models of literacy: Reading assessment in adult education. *Linguistics and Education, 1,* 233–289.

Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy in reading assessment. *TESOL Quarterly, 26,* 433–461.

Hill, C., & Parry, K. (1994). *From testing to assessment: English as an international language.* Harlow, UK: Longman.

Ingulsrud, J. (1988). *Testing in Japan: A discourse analysis of reading comprehension test items* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Jaschik, S. (March 26, 2005). Fooling the College Board. *Inside Higher Education*.

Johnson, A. (2009). *Objectifying measures: The dominance of high-stakes testing and the politics of schooling.* Philadelphia, PA; Temple University Press.

Jonçich, G. (1968). *The sane pragmatist: A biography of Edward L. Thorndike*. Wesleyan Press: Middletown, CT.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stetcher, B. M. (2000). What do test scores in Texas tell us? Santa Monica, CA: RAND.

Koretz, D., & Barron, S. I. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS) (MR-1014-EDU). Santa Monica, CA: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests. In R. L. Linn (Chair), *The effects of high stakes testing. Symposium conducted at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education.* Chicago, IL.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Martinez, M. (2009). *Learning and cognition: The design of the mind.* Boston, MA: Allyn & Bacon.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge/Falmer.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439–483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.) (2008). *Assessment, equity, and opportunity to learn*. New York: Cambridge University Press.

National Academy of Education. (2009). Standards, assessments, and accountability. In L.

Shepard, J. Hannaway, & E. Baker (Eds), *Education policy white paper.* Washington, DC: Author.

National Research Council. (1999). How people learn: Bridging research and practice. In M. S. Donovan, J. D. Bransford, & J. W. Pellegrino (Eds.), *Commission on behavioral and social sciences and education.* Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaser, R. (Eds). Board on Testing and Assessment, Center for Education. Division of Social Sciences and Education. Washington, DC: National Academy Press.

New York State Testing Program. (2002). *Scoring guide and scorer practice set* (English Language Arts, Grade 4). Albany, NY: New York State Education Department.

Nix, D., & Schwartz, M. (1979). Toward a phenomenology of reading comprehension. In R. Freedle (Ed.), *New directions in discourse processing* (pp. 183–196). Norwood, NJ: Ablex.

Parry, K. (1986). *Readers in context: A study of northern Nigerian students and school certificate texts* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2002). Challenging the validity of automated essay scoring, *Computers in Human Behavior, 18,* 103–134.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Rodriguez, R. (1998). *The browning of America*. Available at http://www.pbs.org/newshour /essays/february98/rodriguez_2–18.html

Russ, J., Ehren, M.C.M., & Lesaux, N. (forthcoming). *Strategies teachers use to coach students to do well on the ELA state test: The case of New York City and Boston public elementary schools.*

Russell, M., & Haney, W. (2000, March 28). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives*, *8*(19). Available at http://epaa.asu.edu /epaa/v8n19.html

Russell, M., & Abrams, L. (2004). Instructional uses of computers for writing: The effect of state testing programs. *Teachers College Record, 106,* 1332–1357.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing* (CSE Technical Report 482). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Schwartz, D., & Arena, D. (2009). *Choice-based assessments for the digital age*. Stanford, CA: Stanford University, School of Education.

Sebrechts, M., Bennett, R., & Rock, D. (1991). Agreement between expert-system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76,* 856–862.

Shepard, L. (2000). The role of assessments in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shepard, L. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *72*, 232–238.

Sims-West, N. (1996). *An investigation of gender difference on the Scholastic Aptitude Test of Verbal Ability* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Thorndike, E. (1915). An improved scale for measuring ability in reading. *Teachers College Record, 16,* 31–53, 445–467.

Trabasso, T. (1981). On the making of inferences during reading and their assessment. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 56–75). Newark, DE: International Reading Association.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, H., & Hill, C. (2011). A paradigm shift for English language teaching in Asia: From imposition to accommodation. *The Journal of Asia, TEFL*, *8,* 231-258.

Winerip, M. (May 5, 2005). SAT Test Rewards Length and Ignores Errors. *The New York Times.*

Yuan, Y. (1997). *Reader response to the Taiwan Joint College Entrance Examination: English reading section* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Zhang, W. (2003). *Doing English digital* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

张薇 (2006). 英语数字素养的研究型评价模式,《外语教学与研究》, 第2期,115–121页 [Zhang, W. (2006). Assessing English digital literacy: The project approach. *Foreign Language Teaching and Research*, *2,* 115–121.]