Educational Studies in Japan: International Yearbook
No.4, December, 2009, pp.79-89

Academic Achievement Survey and Educational Assessment Research[1]

TANAKA, Koji*
translated by VAN STEENPAAL, Niels**

The recent "Nationwide academic achievement and study situation survey" was clearly influenced by the idea of "authentic assessment", an educational assessment perspective focused on "quality" and "engagement". However, when "performance assessment", the assessment method corresponding to this focus, is adopted in academic achievement surveys, it runs the risk of turning into a rigid hollow structure. In this paper I will reflect on the ideal application of performance assessment in academic achievement surveys, and will investigate the concepts of "consequential validity", "equity", and "moderation" in regard to their potential to further develop the discourse.

1 Stating the Problem

We can safely say that all the post-war reforms in school education have centered on the issue of academic achievement in Japan. Two representative illustrations of this focus are the 1950s debate revolving around the issue of "declining academic achievement", and the discussion on "academic achievement on examinations" around 1975. In both of these cases the debate was triggered and fueled by the results of academic achievement surveys. The Kubo Shun'ichi survey (1951) and the Public Institute for Educational Research Survey (1975–1976) achieved great fame in this way and have exerted influence on the revisions of the curriculum guidelines ever since. Likewise, the latest revision (March 2008) was inspired by the academic achievement debate around the turn of 21th century. It was furthermore the most fundamental reform in decades in that the focus was changed from "New Academic Achievement" to "Solid Academic Achievement (comprehensive learning abilities)".

Although the current debates at the beginning were solely concerned with declining academic achievement of college students[1], the latter academic achievement debate, influenced by academic achievement surveys and especially 2003PISA, has taken a different course. PISA-type "literacy" assessments are being incorporated into the educational policies of countries around the world as the universal standard for the new global economy.[2] In Japan, this trend is reflected in

* Kyoto University
e-mail: l50095@sakura.kudpc.kyoto-u.ac.jp
**PhD student, Kyoto University
e-mail: vansteenpaal.niels@at7.ecs.kyoto-u.ac.jp

that MEXT, concerned about weakening "reading literacy", in December 2005 proposed the "Program for the improvement of reading literacy" and the "Teaching resources for the improvement of reading literacy". Furthermore, the fact that in the latest revision of the curriculum guideline the issue of "linguistic achievement" was emphasized, and that the "Nationwide academic achievement and study situation survey", carried out as a census survey since April 2007, distinguishes between A-type problems (concerning basic "knowledge") and B-type problems (concerning the "application" of knowledge in everyday situations),[3] are also clear signs of influence by PISA.

Whereas in the past the academic achievement surveys served only to trigger and fuel the debate on the subject, the current situation is that they guide the course of educational reforms. As subjecting the "academic achievement survey" to internal analysis as unraveling its politics[4] has therefore become a matter of high priority. In this paper I will analyze the academic achievement survey from the perspective of educational assessment theory as internal analysis[5] and will clarify the essence of assessment that is currently exerting strong influence on PISA, namely "authentic assessment" or "performance assessment", that is occupied with "quality" and "engagement".[6] I will furthermore examine how the assessment is critically refined as it is adopted in academic achievement surveys in USA. Through the above analysis and examination I aim to identify the problematic issues concerning the "Nationwide academic achievement and study situation survey" and to point out the tasks that lay ahead.

2 Authentic Assessment

(a) the context of its conception

In contrast with TIMSS that measures the extent to which a specific school curriculum is mastered, PISA-type "literacy" places emphasis on the proficiency of knowledge and skills that children need in order to live their daily lives. "Mathematical literacy" for example is described in a functional manner as "concerned with the capacities of students to analyze, reason, and communicate ideas effectively as they pose, solve, and interpret mathematical problems in a variety of situations."[7] Measuring this kind of PISA-type "literacy" therefore requires the creation of concrete problems that are situated in "authentic" contexts. It is in this characteristic that we can clearly recognize the influence of the idea of "authentic assessment".

The term "authentic assessment" made its first appearance within the context of educational assessment in the latter half of the 1980s in America through the work of Wiggins.[8] Inspired by the famous report "Nation at Risk" (1983), in which the importance of raising academic achievement was emphasized, this period witnessed the large-scale introduction of "standardized tests". These "high-stakes" tests were carried out by state governments in order to inspect the educational results of schools and school districts, and to respond to demands for accountability. However, before too long resistance arose against the top-down enforcement of these tests, and questions were asked as to whether educational results could actually be measured trough standardized tests.

"Standardized tests" more often than not present children with artificial and fragmentary (one-shot) questions that demand rote memorization of knowledge. Furthermore, they create a "test atmosphere" that, by exhibiting various kinds of ritualized aspects, is far removed from the normal class atmosphere. These characteristics caused skepticism about the effectiveness of standardized

tests in assessing a child's academic achievement. A high grade on a standardized test might reflect a specific kind of ability that is of value within a school, but does it actually guarantee that the necessary skills to live and work in society are mastered? And, might not the uniformity of standardized tests serve to amplify racial and social differences? It is against this background of critique against standardized tests that the idea of authentic assessment made its first appearance.[9]

(b) criticism towards standardized testing

Despite his critical stance towards standardized tests, Wiggins maintains that even though there are limits to the "test" as an assessment method, we should not discard the method in its entirety.[10] When we take the specific criticism towards "standardized tests" to also apply to "tests" in general, we not only ignore the fact that standardized tests are specifically "group-based" testing, but we also avoid the important task of critically reconstructing the meaning embedded in "tests" as such.

Wiggins therefore advocates a distinction between "establishing a standard" in order to clarify the common level, and "standardization" in order to support "standardized testing". In the past, "establishing a standard" was basically equated with "standardization" and meant setting up an assessment standard (norm) based on the average performance level of a group. However, the "bell curve", used in the analytic process of this kind of standardization, presupposes the existence of well performing children and ill performing children. As soon as teachers become accustomed to this kind of "relative assessment" they are likely to avoid the difficult, but fruitful task of establishing criteria that aim to clarify the common academic achievement level .

Wiggins divides "tests" into "norm-referenced tests" and "criterion-referenced tests" (based on "objective-referenced assessment") and maintains that his emphasis on "authentic assessment" should be understood as a further development of "objective-referenced assessment". "Further development" here most likely meaning, orientated towards the development of an assessment method that overcomes the limits of tests. However, as I shall explain in further detail later, when the assessments methods implied by "authentic assessment" are adopted in large-scale academic surveys, there is a risk that they will, contrary to their original intention, transform into devices of oppression. There is an increasing awareness of the necessity for the construction of a theory that will prevent this kind of transformation.

(c) the meaning of "authentic" : focusing on "quality"

Wiggins has described "authentic assessment" as a method that "replicates or simulates the contexts in which adults are 'tested' in the workplace, in civic life, and in personal life."[11] Shaklee has referred to it as the assessment of children within a process that involves them with "realistic tasks."[12]

The emphasis here on terms like "life" and "realistic tasks" obviously implies criticism towards the uniformity and artificiality of the standardized tests. It is supposed that the ability of a child is formed through the involvement with "authentic" tasks, and that it is therefore exactly this process that should be the object of assessment. Teaching and assessment are thus not thought of as separate, but rather as two continuous aspects of the same process.

Saying that the assessment task has to be "realistic" within the context of the child's life means that the task has to be a familiar and unavoidable part of that child's life. However, we should note that "realistic" within the context of authentic assessment implies more than just familiarity and necessity. Wiggins maintains that "authenticity" in assessment corresponds to the quali-

tatively higher-level objectives of "synthesis" and "application" that have been described by Bloom in his "Taxonomy of Educational Objectives".[13] "Synthesis problems" for example are explained by Bloom as problems that have not been treated during class but are attempted by children by using various materials in a manner like an open-book examination.

Wiggins thus relates "authenticity" in assessment with the qualitatively higher-level difficulty of the assessment problems posed to children. Problems that reflect the familiar daily lives of children inspire them to attempt these problems, but are also particularly difficult in that these daily lives are worlds in which the information necessary to solve problems is either abundant or scarce, thus demanding the ability to refine and synthesize knowledge. Awareness of this paradox between "familiarity" and "difficulty" contained within "authenticity" is of crucial importance to a correct understanding of the idea of "authentic assessment".

3 Performance Assessment

(a) restructuring validity and reliability

We have established that the essence of the idea of "authentic assessment" lies in its focus on higher-level "quality". However, when we want to turn this idea of "authentic assessment" into a concrete assessment method, we have to address the methodological principals of "validity" and "reliability". Whereas validity is a concept that describes the extent to which the assessment object has been measured accurately, reliability describes the extent to which the assessment object has been measured consistently. Although closely related, these concepts have developed in an oppositional manner; where one was emphasized, the other was neglected. For example, while tests made by teachers themselves were only concerned with validity at the expense of reliability, high-stakes tests that demanded accountability focused exclusively on reliability while sacrificing validity. In the case of large-scale academic achievement tests the demand was for reliability. And this demand was accordingly satisfied by making almost exclusive use of "objective tests".

Overcoming this conflict between validity and reliability requires the development of an assessment method of high validity, together with the establishment of an assessment standard that is able to secure reliability. This has more than all become the task at hand for "authentic assessment" understood as a further development of "objective-referenced assessment". The fact that the recent "Nationwide academic achievement and study situation survey" established B-type problems signifies a commitment to this new task. I will now examine how the concepts of "validity" and "reliability" came to be reconstructed and understood in an integrative manner.

The concept of "validity" as an assessment method principle has traditionally been divided into "construct validity", "content validity", and "criterion-related validity", with the latter further dividing into "concurrent validity" and "predictive validity". I will first offer a brief explanation of these concepts.[14]

"construct validity"—describes the extent to which the assessment method is adequately measuring the theorized constructive concept that is taken as the object of assessment. It is therefore necessary to accurately define one's constructive concept beforehand.

"content validity"—describes the extent to which the assessment method is accurately representing or abstracting the object of assessment based on the constructive concept. It is therefore necessary to identify the important items within the domain in question.

"concurrent validity"—describes the extent to which the assessment method is successful in measuring the constructive concept when compared to another method measuring the same concept. As a premise, this other method must however possess high validity.

"predictive validity"—describes the extent to which the assessment method is able to accurately predict future results. As a premise, one has to however assume that the constructive concept remains unchanged.

From the perspective of authentic assessment, "construct validity" is the most important kind of validity because it examines the evidential grounds for the "educational objectives" being referenced to, and describes to what extent these objectives are reflected in the assessment method. We can think of "content validity" as the concrete function of this "construct validity". Furthermore, the criteria of the "criterion-related validity" as "concurrent validity" and "predictive validity" has traditionally been discussed on the basis of the "objectivity" of "relative assessment" (for example intelligence tests). From the perspective of "objective-referenced assessment" and "authentic assessment", it is therefore necessary to examine the significance and meaning of "concurrent validity" and "predictive validity" when their criteria become educational objectives.

Gipps has further refined "construct validity" into the concept of "curriculum fidelity", an idea that demands that the assessment method covers the entire spectrum of the curriculum and that it is matched with its particular domain and level.[15] In cases where, despite the fact that qualitatively higher-level educational objectives (for example, expressive ability or problem solving ability) are established, an objective test corresponding to lower-level educational objectives is used as an assessment method, two problems arise. Firstly, it will remain unclear whether the higher-level objectives have been reached or not. And secondly, children will adapt to the lower-level tests. Gipps suggestion of "curriculum fidelity" emphasizes this problem and urges the development of an assessment method that corresponds to qualitatively higher-level educational objectives.

"Reliability" is a concept describing the extent to which the precision of the assessment results are stable and coherent, regardless of where, when and by whom the measurement was carried out. In the past, we see that the "measurement movement" that arose as criticism against "absolute assessment", focused on the sole pursuit of "reliability" through the use of statistical methods, and as a result lost all "validity". How should we then understand the concept of "reliability" within the context of "authentic assessment"?

"Reliability" has traditionally been divided into the "reliability of the assessment method" and the "reliability of grading".[16] The first reliability describes the extent to which the assessment method is stable. Methods are, for example, the "test-retest method" in which the same test group is measured twice with the same assessment method, and the "parallel testing method" in which two measurements are carried out within the same population by two assessment methods comparable in form and difficulty. The second type of reliability describes the extent to which the grades are consistent. Examples are the "inter-rater reliability" that describes the consistency of grades by different raters, and the "intra-rater reliability" that describes the consistency of grades for one child when assessed multiple times by the same rater.

As has been pointed out, an assessment method that pursues "reliability" without regard for "validity" will result in an "objective test" or "standardized tests". That is why Gipps, in tandem with "curriculum fidelity" as a new concept of validity, has suggested "comparability" as a new concept of reliability.[17] Comparability describes the extent to which different raters share the same

understanding of the assessment standard, and are assessing the performance of the assessment object impartially according to the same rubric. In order to develop an assessment method that can correspond to "quality", "authentic assessment" theory is reconstructing the concepts of validity and reliability to try and understand them in an integrative manner. It is against this background that the idea of "performance assessment" was suggested.

(b) assessing "performance"

Performance within the context of "performance assessment" refers to the external presentation of an internal state of mind through for example gesture, behavior, painting, or language. "Performance assessment" tries to understand the rich aspects of learning as expressed through the involvement with "authentic tasks". By assessing children through tasks like free essays, live demonstrations, and presentation, "performance assessment" aims to comprehend the qualitatively higher-level aspects of academic achievement like intelligence, insight, and expression.[18]

In a popular dictionary of American education "performance assessment" is described as a method measuring "how well students apply knowledge to the real world."[19] This kind of description is a common one and we also find it reflected in the B-type problems(ex. too-much information problem, free essay) of the recent "Nationwide academic achievement and study situation survey". We should note however that free essays that can be carried out by nothing more than "pen and paper", even though they might require "application" and "integration", are by Wiggins considered as a "prompt". Wiggins uses "performance assessment" in the restricted sense of a method that requires the child to perform a certain role.[20] This means that depending on the type and accuracy of the assessment method adopted, there will be a difference in the quality of understanding of the performance. The performance task emphasized by Wiggins can therefore only be realized through synthesis with actual teaching practice, and is clearly not suited for large-scale simultaneous academic achievement tests.

In order to secure the reliability of "performance assessment" we have to develop rubrics. A rubric consists of "scales", descriptors that signify assessment standards, and concrete samples (also referred to as anchors).[21] The potential for transmission and verification of rubrics is heightened by linking each scale with a representative sample (for example in the case of "data collection", papers written by the child). Because the scales are accompanied by pre-decided descriptors and samples, they can be considered "ordinal scales" or "interval scales" rather than "nominal scales".[22] It is exactly because of this "reliability" that rubrics have been proposed as a method to guarantee the objectivity of assessment standards.

It is important however that the rubric is also made available to children in a readily understandable form. Although there might be some resistance to making the assessment index openly available, it serves a clear and significant purpose. Firstly, in the difference between "open" and "secret" standards lies the difference between "objective-referenced assessment" and "absolute assessment". Making the rubric openly available furthermore enables the critical exploration and correction needed when problems arise during or after assessment.

Even more importantly, an "open" rubric can serve children as a guideline in its study activities and self assessment. A grading by use of rubrics is nothing more than an assessment of children at that particular time, and does not signify any kind of ultimate judgment. When for example a child receives a grading of 2 for a science experiment, it is important that the teacher and child share the same understanding of how to improve on the learning process in order to receive a grading of 3. It is this important process of shared understanding that rubrics are expected to facilitate.

Although for the recent "Nationwide academic achievement and study situation survey" the equivalent of a rubric was presented in the form of the "Solution types" within the "Explanation materials" and "Outline of the survey results", there is still plenty enough room for improvement when it comes to the establishment of rubrics. It is fair to say that the "confusion" reported in newspapers, brought to light some fundamental problems concerning the establishment of rubrics and the proficiency of raters.[23]

4 Performance Assessment in Academic Achievement Surveys

(a) high-stakes academic achievement surveys

In America in the early 1990s the "standards movement" arose. The movement was highly influenced by the idea of "authentic assessment" and managed to convey their message to PISA, eventually leading to the adoption of performance assessment in state level academic achievement tests. Then, as a result of the establishment of the No Child Left Behind Act in 2001, the demand for accountability rose, causing the spread of high-stakes state level tests that were supposed to respond to this demand. The principle of competition inherent in these tests led to the reallocation of teachers and sometimes even pushed schools to the brink of closure.[24]

Even though it was admitted that, as method of educational assessment, "performance assessment" was superior over "standardized tests", the above-mentioned situation gave rise to criticism against its adoption in high-stakes academic achievement surveys.[25] And although the "standards movement" got caught up in the high-stakes state level tests and was now focusing on exam training to improve test scores, there were also those who defended the original position of the "standards movement" as rooted in the protection of the rights of the socially vulnerable.[26]

Despite differences in approach, all criticism sprung from the shared concern that, when adopted in high-stakes academic achievement surveys, "performance assessment" would end up as a tool for indoctrination and discrimination no different than "standardized tests", constricting the functioning of education and schools. It was feared that adopting rubrics would create distance from the child's actual activities.[27] It is interesting to note that identical concerns were voiced regarding the B-type problems of the recent "Nationwide academic achievement and study situation survey".

That "authentic assessment" has always shared these concerns can be clearly witnessed from the fact that its position demanding "engagement" in educational assessment also includes grass-roots criticism against the "standards movement" and the No Child left Behind Act that, by prioritizing business profit, both promote a new kind of discrimination.[28] It is against this background of "engagement", and guided by the methodological principles for academic achievement as implied by Gipps, that I would like to ruminate on the potential of "performance assessment".

(b) methodological principles for academic achievement surveys

Besides reconstructing the concepts of validity and reliability, Gipps also suggested some concepts that can be considered as methodological principles for academic achievement surveys. One of these is the concept of "consequential validity", which describes the kind of consequences that result from the practice of a certain assessment method.[29] When for example a higher-level assessment method with high "curriculum fidelity" is put into practice, there is a risk that teachers will come to emphasize exam training at the expense of reflection upon the quality of lesson study

and the formation of higher-level academic achievement. The concept of "consequential validity" serves to indicate this kind of risk.

Overcoming this problem however will require a reform of educational practice that fully realizes the limits of "performance assessment" adopted in "pen and paper"-style academic achievement surveys, and that will train children in daily life performances. This will demand a teaching strategy that is aware of the structure of performance, and that is based on the multifarious requirements of classroom and child.[30]

Then there is the principle of "equity", which intervenes in the process of test creation and grading, and tries to provide equal conditions for the examinees by taking into account various cultural biases like sex, nationality, race, ethnicity and class.[31] Depending on the materials used in the test problems of international academic achievement surveys, some cultural regions are clearly at a disadvantage. For example, to a child living in a country without a train network, test problems that include materials on trains pose obvious difficulties. In these cases, the aim should be to use materials relating to the public transportation network popular in that child's own country. We see that equity is based on the same emphasis on realistic context we also find in "authentic assessment", and that it can help us avoid the pitfalls of academic achievement surveys.

Lastly, there remains to be discussed the principle of "moderation", which runs through not only the issue of academic achievement but through educational assessment in its entirety.[32] Although the principle was originally proposed as a method to heighten "comparability", it is also related to the emphasis on the expertise of teachers and the establishment of democracy within the context of educational assessment. Moderation can be divided into the method of unifying the assessment process, and the method of unifying the assessment results. Often used in "performance assessment", the principle of group moderation is a method that unifies the assessment results in order to create a rubric, and it is valued as an effective method to heighten the proficiency of the teacher in educational assessment.[33] Because it creates a rubric in a bottom-up fashion, building up assessment criteria in partnership as it were, it is also a strategy that helps to avoid fixation of assessment criteria. In a recent work, Wiggins suggested "curriculum management", a strategy for constructing schools that would realize "authentic assessment", and would increase the expertise of teachers[34]. Because this implies construction of moderation on school level, we should understand "curriculum management" as yet another strategy to defend against the fixation of educational assessment.

As we have seen, "authentic assessment" consists of strategies and principles that, grounded in the awareness of "engagement", developed in opposition to the large-scale academic achievement surveys. These strategies and principles can become effective methodological principles in the analysis of not only the "Nationwide academic achievement and study situation survey", but also the academic achievement surveys on school and district level. We especially have to pay attention to how the principle of moderation, which reconstructs the politics of academic achievement surveys from within, is going to be utilized. While learning from American and European experience, also Japan is now venturing into the territory of research and practice of academic achievement surveys.

Note

1. This article was originally published in Japanese at *The Japanese Journal of Educational Research,* Vol.75,No.2,2008.

References

1 The current academic debate was triggered by a piece of sensational journalism published in the *Shūkan Asahi* (Asahi Weekly) titled "Tōdai, Kyōdaisei mo 'Gakuryoku Hōkai'" (Crumbling Academic Achievement even for Students of Tokyo and Kyoto University). See, Tanaka, Kōji. "Konnichi no gakuryoku mondai de towarete iru koto." Tanaka Kōji (ed.). *Gakuryoku to hyōka no 'ima' o yomitoku*. Tokyo: Nihon hyōjun, 2004.

2 See, Harada, Nobuyuki (ed.). *Tashika na gakuryoku to yutaka na gakuryoku: kakkoku kyōiku kaikaku no jittai to gakuryokumoderu*. Kyoto: Minerubua shobō, 2007; Ōmomo, Totsuyuki. Uesugi, Takamichi. Inokuchi, Junzō. Ueda, Takeo (eds.). *Kyōiku kaikaku no kokusai hikaku*. Kyoto: Minerubua shobō, 2007.

3 For analyses of the assessment tasks of several recent surveys see, Tanaka, Kōji (ed.). *Atarashii gakuryokutesuto o yomitoku: PISA, TIMSS, Zenkoku gakuryoku gakushū jōkyō chōsa,Kyouiku katei jishi jokyou chosa*. Tokyo: Nihon Hyōjun, 2008.

4 See, Nakajima, Tetsuhiko. "Zenkoku gakuryokutesuto mittsu no mondaiten: mokuteki, kengen, kojinjōhō." *Kyōiku*, May 2007.

5 This paper is based on the following understanding of academic achievement surveys as a method of educational assessment.
"An academic achievement survey is a survey conducted among students ranging from elementary school to higher education (recently also including university) with the purpose of assessing the state of academic achievement in an organized, systematic and scientific manner. The results of this survey are then used to reflect upon the social (regional, economical and cultural conditions) and educational (conditions and regulations concerning school and classroom) environments that are considered to have influence on this state of academic achievement.
The accuracy and validity of this kind of survey however depends on several factors for example, the academic achievement perspective on which the survey problems are based, the state of research on educational assessment and curricula, the period during which the survey is conducted, the range of the survey object, and the way in which the results are processed. Matters like the nature of the conducting institute, its members, and the expertise of the survey conductors and raters also influence the success of the survey.
In the case of international surveys, the need for survey items with high comparability furthermore requires consideration of the linguistic, cultural, and educational conditions of each country."
In this article I will reflect upon the history of educational assessment research and will offer an internal analysis of the large-scale academic achievement census surveys that are being carried out on national and district level. The first effort at this kind of internal analysis is probably, Mitsui, Daisuke. Murakoshi, Kunio. "Sengo no gakuryoku chōsa: sono hōhōron no kentō." (1,2,3.) *Kokumin kyōiku*, nr. 27, 28, 31, 1976~77.

6 Terms representative of this new trend in educational assessment are for example, "authentic assessment", "performance assessment", and "direct assessment". In Japan this trend was first introduced in, Sawada, Minoru. "Amerika gasshūkoku ni okeru kyōiku hōkō kaikaku no saizensen." Matsuura, Yoshimitsu. Nishikawa, Nobuhiro (eds.). *Kyōiku no paradaimu tenkan*. Tokyo: Fukumura shuppan, 1997.
In this article, I use "authentic assessment" as a general term to refer to the views of Gipps, C and Wiggins, G, and understand "performance assessment" as the representative assessment method corresponding to these views.

7 OECD (ed.). *The Pisa 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: OECD Publications, 2003, p.24.

8 Janesick, Valerie J. *The Assessment Debate*. Santa Barbara, CA: ABC-Clio, 2001, p.6. For research relating to the works of Wiggins see, Endō, Takahiro. "G. Uiginzu no 'kanpa' gakushū." *Kyōiku hōhōgaku kenkyū* 2004:30; Nishioka, Kanae. "Uiginzu to Makutai ni yoru 'gyakumuki sekkei' ron no igi to kadai." *Karikyuramu kenkyū*, 2005:14.

9 See, Burke, Kay (ed.). *Authentic Assessment: A Collection*. Thousand Oaks, CA: Corwin Press, 1992.

10 Wiggins, G. "A True Test: Toward More Authentic and Equitable Assessment." *Phi Delta Kappan*, May 1989. See also, Takata, Kōji. "Ōsentikku asesumento toha dono yō na hyōka hōhō ka." *Rika no kyōiku*, December 2001.

11 Wiggins, G. *Educative Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass Publishers, 1998,p.24.

12 Shaklee, Beverly D(others). *Designing and Using Portfolios*. Boston: Allyn and Bacon, 1997, p.6.

13 Wiggins, G. op. cit. pp. 24–25. Wiggins gives "understanding" as the essential definition of "ability" and further divides it into the six facets of "explanation", "interpretation", "application", "perspective", "empathy", and "self knowledge". "Perspective" here means to observe a situation from multiple perspectives, and "empathy" refers to reflecting on a situation from the perspective of someone else. Wiggins G. and McTighe, J. *Understanding by Design.* (2nd Edition). Alexandria, VA: ASCD, 2005, chap. 4.

14 See, Hashimoto, Shigeharu. *Shinkyōiku hyōkahō sōsetsu*. Tokyo: Kaneko shobō, 1976, vol.1, chap. 6.

15 Gipps, Caroline V. *Beyond testing: towards a theory of educational assessment*. London: Falmer Press, 1994, p.172.

16 Hashimoto, Shigeharu. op. cit. vol.1, chap. 6; Gipps, Caroline V. op. cit. pp.67–69.

17 Gipps, Caroline V. op. cit. p.171.

18 An introduction to performance assessment in Japanese is for example, Yoshida, Shinichirō. *Tesuto dake deha hakarenai! Hito o nobasu 'hyōka' toha*. Tokyo: NHK shuppan, 2006.

19 Collins, John W. and O'Brien, Nancy Patricia (eds.). *The Greenwood Dictionary of Education*. Westport, CT: Greenwood Press, 2003, p. 263.

20 Wiggins, G. and McTighe, J. op. cit. p. 153. Wiggins uses the following example of a performance task scenario (McTighe, J and Wiggins, G. *Understanding by Design:Professional Development Workbook.*ASCD,2004,p.171).
Goal:
• Your goal is to help a group of foreign visitors understand the key historic, geographic, and economic features of our region.
Role:
• You are an intern at the Regional Office of Tourism
Audience:
• The audience is a group of nine visitors (who speak English)
Situation:
• You have been asked to develop a plan, including a budget, for a four-day tour of the region. Plan your tour so that the visitors are shown sites that best illustrate the key historic, geographic, and economic features of our region.
Product Performance and Purpose:
• You need to prepare a written tour itinerary and a budget for the trip. You should include an explanation of why each site was selected and how it will help the visitors understand the key historic, geographic, and economic features of our region. Include a map tracing the route for the tour.
Standards and Criteria for Success:
• Your proposed tour plan (including itinerary, budget and route map) needs to include
- The key historic, geographic, and economic features of the region.
- A clear rationale for the selected sites.
- Accurate and complete budget figures.

21 Wiggins uses the following example of an analytic-trait rubric for a fifth grade science experiment (data collection).
4. Data were collected and recorded in an orderly manner that accurately reflects the results of the experiment.
3. Data were recorded in a manner that probably represents the results of the experiment.
2. Data were recorded in a disorganized manner or only with teacher assistance.
1. Data were recorded in an incomplete, haphazard manner or only after considerable teacher assistance.
Wiggins, G. op. cit. p. 167.

22 Within "scales" we can distinguish, "nominal scales" (in which numbers are arbitrarily decided labels lacking mathematical properties), "ordinal scales" (in which numbers indicate relative position), "interval scales" (an ordinal scale possessing equal intervals. Ex. the Celcius scale), and the "ratio scale" (an interval scale possessing an absolute zero point. Ex. the Kelvin scale). The "measurement movement" treated "nominal scale" measurements as if they were based on a "ratio scale", and analyzed them in a statistical manner. See, Adachi, Jirō. "Kyōiku hyōka ni kakawaru jakkan no gainen no kentō." *Kyōikugaku kenkyū* 43:2, June 1976.

23 "Gakuryoku Chōsa Saiten de Konran." *Asahi Shinbun*, June 15th 2007.

24 See, Tsuneyoshi, Ryōko. "Kōkyōiku ni okeru hai-sutēkusu na tesuto: purinsu-jōjizugun no rikonsutichūshon." *Kyōikugaku kenkyū* 67:4, March 2000; Tsuneyoshi, Ryōko. "Hai-sutēkusu na gakuryokutesuto o mochiita mikuroreberu no kyōiku kanri: amerika, merīrandoshū gakuryokutesuto teihenkō no jirei kara." *Kyōiku mokuhyō-hyōka gakkai kiyō* 17, 2007; Akaboshi, Shinsaku. *Amerika kyōiku no shosō*. Tokyo: Gakubunsha, 2007.

25 See, Eisner, E. W. "The Uses and Limits of Performance Assessment." *Phi Delta Kappan*, May 1999. And, in the same issue, Haertel, E. H. "Performance Assessment and Education Reform."

26 Thompson, S. "The Authentic Standards Movement and Its Evil Twin." *Phi Delta Kappan*, January 2001.

27 Mabry, L. "Writing to the Rubric." *Phi Delta Kappan*, May 1999.

28 Janesick has reported on this issue in the case of Texas and Florida, and introduced the "FAIRTEST" of the grass roots movement. Janesick, Valerie. *Authentic Assessment*. New York, NY: Peter Lang, 2006. A point of emphasis within "authentic assessment" is that those who are directly affected by the assessment results, the so-called stakeholders, should have a right to "engage" in the act of assessment. For more on the "Stakeholder Approach" within educational assessment see, Weiss, C.H. "The Stakeholder Approach to Evaluation: Origins and Promise." in House, Ernest R. (ed.). *New Directions in Educational Evaluation*. London: Falmer Press, 1986.

29 Gipps, Caroline V. op. cit. p.63. This proposal was suggested within the context of the intensified inter-school competition in England that resulted from the policy to publicly announce the results of the National Test.

30 See, Tomlinson, C. A. and McTighe, J. *Integrating Differentiated Instruction and Understanding by Design*. Alexandria, VA: ASCD, 2006.

31 Gipps, Caroline V. op. cit. pp.148–151.

32 Gipps, Caroline V. ibid. pp.137–139; Randnor, H. and Shaw, K. "Developing a Collaborative Approach to Moderation." in Torrance, Harry (ed.). *Evaluating Authentic Assessment: Problems and Possibilities in New Approaches to Assessment*. Buckingham: Open University Press, 1995.

33 The procedure of group moderation is suggested by Wiggins as follows.

1. The scale on which the performance/work is to be graded is confirmed.
2. The performance/work is graded by at least three teachers.
3. A guidance plan is created on the basis of the performance/work that received the same grade by all teachers. Wiggins, G. op. cit. chap. 7.

34 See, Wiggins, G. and McTighe, J. *Schooling by Design: Mission, Action and Achievement*. Alexandria, VA: ASCD, 2007. For the development of the idea of "curriculum management" in Japan see, Nishioka, Kanae (ed.). *'Gyakumuki sekkei' de tashika na gakuryoku o hoshō suru*. Tokyo: Meiji Tosho, 2008.